Measurement of Serious Emotional Disturbance in Population-based Studies of US Children: Considerations and Recommendations

Katholiki Georgiades, PhD, Associate Professor, Department of Psychiatry and Behavioural Neurosciences and David R. (Dan) Offord Chair in Child Studies, McMaster University, Offord Centre for Child Studies and McMaster Children's Hospital

Peter Szatmari, MD, Chief of Child and Youth Mental Health Collaborative, The Hospital for Sick Children and Centre for Addictions and Mental Health, Professor and Head of the Division of Child and Youth Mental Health, University of Toronto, Patsy and Jamie Anderson Chair in Child and Youth Mental Health

This paper was commissioned by the National Academies of Sciences, Engineering, and Medicine Standing Committee on Integrating New Behavioral Health Measures into the Substance Abuse and Mental Health Services Administration's Data Collection Programs. Opinions and statements included in the paper are solely those of the individual authors, and are not necessarily adopted, endorsed or verified as accurate by the National Academies of Sciences, Engineering, and Medicine. Support for the Standing Committee was provided by a contract between the National Academy of Sciences and the U.S. Department of Health and Human Services.

08/07/2016

BACKGROUND

This paper was written to inform the work of the National Academies of Sciences, Engineering and Medicine Standing Committee on Integrating New Behavioral Health Measures into the Substance Abuse and Mental Health Services Administration's (SAMHSA) Data Collections Programs. The committee was asked to assist SAMHSA with its responsibilities of expanding behavioral health data collections in several areas, including the measurement of Serious Emotional Disturbance (SED) in children. The most recent definition states that Serious Emotional Disturbance (SED) in children from birth to 18 years, is defined as the presence of current or past-year diagnosable mental, behavioral, or emotional disorder of sufficient duration and severity to meet *DSM-III-R* (or equivalent) diagnostic criteria. In addition, the disorder must result in functional impairment which substantially interferes with or limits the child's role or functioning in family, school or community activities (Federal Register, 1993). Children identified with an SED are eligible to receive services funded by Federal community mental health services block grants; therefore, state-level prevalence estimates are essential to quantify service needs by individual states to determine block grant funding levels. According to the federal definition, SED does not include either substance use disorders or developmental disorders that are captured by other national data and funding sources.

This paper provides an overview of important practical, methodological and conceptual issues that warrant careful consideration when selecting an instrument to establish 12-month prevalence of SED in the US child population. We conclude with recommendations for the conduct of measurement evaluation studies that could provide a comprehensive evidence base to inform SAMHSA's approach to achieving its overall objective.

Practical Considerations

Declining response rates, increased costs, and respondent burden

Steadily declining response rates in household surveys over the past three decades (National Research Council, 2013) threaten to undermine the potential usefulness of such surveys for informing public health policy and service needs. Non-response is particularly high among at-risk populations (i.e., single-parent households, families with young children, ethnic/racial minorities and areas characterized with high crime rates), introducing the potential for bias in estimates of disorder arising from these surveys. The increase in non-response is also associated with increased costs of survey implementation, due to repeated attempts to contact sampled individuals and missed appointments.

As a result of these concerns, respondent burden is an important consideration when selecting an instrument to use for measuring SED in national, epidemiologic studies. Practical considerations, including time, costs and ease of administration, need to be carefully considered, in conjunction with the psychometric properties (reliability and validity) of potential measures under consideration.

It is generally assumed that structured and semi-structured diagnostic interviews are the 'gold standard' for measuring mental disorders in general population samples of children (we use the term "children" to include both children and adolescents). In comparison to questionnaires, many assume that structured diagnostic interviews provide more reliable and valid classifications because they were explicitly developed to measure *DSM* defined mental disorders and provide opportunities for the standardization of data collection. The perceived psychometric benefits of the interview, however, need to be carefully evaluated and weighed against the considerable burden on respondents (i.e., time), constraints for administration, and costs associated with training and implementation.

Questionnaires, on the other hand, can be brief, inexpensive to implement, pose little burden on respondents, can be administered in a variety of settings to multiple informants using various modes of administration (i.e. on line). In light of the declining and differential response rates and the resource implications (in terms of time and financial) associated with administering diagnostic interviews, the viability of use in epidemiological surveys is called into question. The practical advantages of questionnaires provide a compelling case to carefully examine and contrast the psychometric properties (reliability and validity) of these two approaches: mental health questionnaires versus structured diagnostic interviews. If there are incremental psychometric advantages associated with the interview, are they large enough to offset the potential methodological limitations (burden, financial costs, non-response) that may lead to distortions in the yielded data?

Methodological and Conceptual Considerations

Lack of a 'gold standard' criterion for ascertaining the presence/absence of mental disorder

A fundamental challenge in the assessment of childhood mental disorder is the absence of a 'gold standard' determination of the presence/absence of disorder. As such, there is no direct way of measuring the validity of diagnoses (Kraemer, 2014). Instead, establishing validity means challenging it using a variety of external criteria. The more challenges a diagnosis can withstand, the more likely the diagnosis is valid (Kraemer, 2014). However, in the absence of reliability, the diagnosis is unlikely to withstand many challenges because unreliable measurement attenuates all correlations. For that reason, a critical first step is to have a reliable instrument to evaluate the evidence supporting or refuting its validity.

Reliability of Diagnostic Interviews

This report is not meant to provide an exhaustive review of the psychometric properties of diagnostic interviews and mental health questionnaires but rather to draw attention to the importance of carefully considering the psychometric properties of each measurement approach. Some have argued that test-retest reliability – the correlation of two or more independent diagnoses per individual within a time interval during which the individual is unlikely to change diagnostic status- is the most important for clinical decision making (Kraemer, 2014). For a binary measure (presence/absence of diagnosis), test-retest reliability is assessed using the intraclass kappa coefficient, a chance corrected measure of agreement (Cohen, 1960; Kraemer, Periyakoil, & Noda, 2002).

A recent review identified 17 studies which provide test-retest reliability estimates for six diagnostic interviews –the CAPA, PAPA, K-SADS-PL, DISC, DICA and MINI-KID - administered to parents only, parents and adolescents combined, or adolescents only to classify the presence/absence of psychiatric disorder (Boyle et al., under review). Nine of the 17 studies were done in clinic samples; two, in mixed clinic and community samples; and 6 used community based samples. The review revealed substantial variability in test-retest reliability estimates across disorders within the same diagnostic instrument and between diagnostic instruments. For example, evaluation of the PAPA (Egger et al., 2006) revealed test-test reliability estimates ranging from 0.39 for GAD to 0.74 for ADHD. Average test-retest reliability estimates (kappa) across disorders for each instrument were 0.60 for the PAPA, 0.75 for the CAPA, 0.47 for the K-SADS, between 0.48-0.68 for the DICA, 0.75 for the MINI-KID and between 0.31-0.76 for the DISC (Boyle et al., under review). In general, average reliability estimates were higher in parent or parent and adolescent reports combined (0.57) compared to adolescent only (0.45); and lower in community samples (0.48) compared to clinic or mixed clinic/community (0.58). The specific sources of reliability however, cannot be determined because binary diagnoses are based not only on the endorsement of symptoms, but also on indicators of their clinical and diagnostic significance, such as the number and constellation of symptoms, age at onset, duration, recurrence, severity and/or impairment.

This review clearly demonstrates that reliability estimates of diagnostic interviews vary as a function of the sample (clinic versus community), respondent (parent and/or adolescent) and disorder type. The lower intraclass kappa estimates for community samples for whom prevalence of psychiatric disorder is much lower than clinical samples and for rarer psychiatric disorders are not surprising, in light of the base rate problem with kappa (Kraemer, 2014). In homogeneous populations, where the prevalence of disorder is consistently very low or very high, even minimal error can significantly reduce reliability estimates. Applying standards for evaluating reliability estimates derived from medical diagnoses (Kraemer, 2014), most of the kappas derived from diagnostic interviews administered to parents and/or adolescents fall between 0.4-0.60, which are characterized as “good” but not “very good”. An important consideration, however, for fulfilling SAMHSA’s mandate to establish 12-month prevalence estimates of SED in US children aged 0-18 years, is the base rate problem mentioned above. The lower prevalence of SED in community samples and in younger ages may have significant impacts on the reliability of binary outcome measures.

Reliability of Mental Health Questionnaires

Most questionnaires that assess a broad spectrum of emotional and behavioral symptoms that define the most common disorders in childhood produce dimensional ratings of disorder rather than binary outcomes. As a result, test-retest reliability estimates for dimensional scales are reported using intraclass correlations (ICCs) or pearson correlations. While intraclass kappas and intraclass correlations are directly comparable (Fleiss & Cohen, 1973), pearson correlations are not.

A recent review was conducted of child and adolescent mental health outcome measures suitable for informing clinical practice at the individual and service levels, including use for local and national benchmarking (Deighton et al., 2014). The review identified 11 measures, ten of which were self-report questionnaires that had the potential for use in practice and policy and available psychometric evidence. The identified measures were: 1) Achenbach System of Empirically Based Assessment (ASEBA), 2) Beck Youth Inventories (BYI), 3) Behavior Assessment System for Children (BASC), 4) Behavioral and Emotional Rating Scale (BERS), 5) Child Health Questionnaire (CHQ), 6) Child Symptom Inventories (CSI), 7) Health of the National Outcome Scale for Children and Adolescents (HoNOSCA), 8) Kidscreen, 9) Pediatric Symptom Checklist (PSC), 10) Strengths and Difficulties Questionnaire (SDQ), 11) Youth Outcome Questionnaire (YOQ). The implementation features (including respondent, age range, response scales, length and financial costs) and the psychometric properties (validity, internal consistency and reliability) of each measure were reviewed (Deighton et al., 2014).

For most of the mental health questionnaires reviewed (Deighton et al., 2014), test-retest reliability estimates (as reported by an ICC or pearson *r*) were higher than those reported for binary diagnoses derived for structured diagnostic interviews. For example, test-retest reliability correlations for the ASEBA CBCL, TRF and YSR ranged from 0.82-0.94 (8 day interval), 0.60-0.95 (16 day interval), and 0.68-0.91 (8 day interval), respectively. While the ASEBA instruments may be considered too time consuming for the purposes of SAMHSAs mandate, an instrument worth considering is the Brief Problem Checklist (Chorpita et al., 2010). Item response theory (IRT) and factor analytic methods were applied to the CBCL and YSR, to develop and evaluate a 12-item child and caregiver version of a Brief Problem Checklist (BPC) telephone interview. Evaluation of this instrument provides support for its psychometric properties. Test-retest correlations for child reported internalizing and externalizing symptoms were 0.76 and 0.72 (8 day time interval), respectively. For the caregiver version, comparable estimates were 0.76 (internalizing) and 0.78 (externalizing). It should be noted that this instrument has been evaluated as an orally administered interview over the phone and not self-complete.

Other instruments selected in the review by Deighton et al. (2014) include the Pediatric Symptom Checklist (PSC). Test-rest estimates for the PSC for parent report were 0.86 (1 week interval) and 0.45 for youth report (Deighton et al., 2014). The test-rest time interval for the youth report, however, was 4 weeks, which may be considered too long an interval. When assessing test-retest reliability, the interval between independent assessments must be long enough to ensure blindness among the ratings but short enough to ensure that disorder status has not changed.

Reviews of the psychometric properties of the SDQ for 4-12 year olds (Stone et al., 2010) and 3-5 year olds (Kersten et al., 2016) have also been conducted. Weighted mean test-retest correlations across 6 studies for the total difficulties score of the SDQ 4-12 year old version, were 0.76 for parent report and 0.84 for teacher report (Stone et al., 2010). The test-retest time interval, for some of the included studies (Goodman, 1999; 2001), however, was too long, ranging from 4-6 months. Estimates derived from these studies are not likely to be valid reliability estimates, but rather, estimates of consistency over time. It is important to note, however, that proper evaluation of test-retest reliability, within a shorter time frame, would most likely result in even higher estimates.

Applying standards for evaluating test-retest reliability estimates based on intraclass correlation coefficients for dimensional measures, 0.6-0.8 is considered ‘good’ and 0.4-0.6 is considered ‘adequate’ (Kraemer et al., 2012). Many of the mental health questionnaires identified above, such as the ASEBA CBCL, PBC and SDQ fall well within the ‘good’ range. The higher test-retest reliability estimates reported for dimensional ratings of diagnoses, compared to binary diagnoses, are to be fully expected given the increase in variance associated with dimensional scales. In fact two meta-analyses, of 58 studies and 59,575 participants quantified the relative reliabilities and validities achieved by binary versus dimensional measures of psychopathology and examined conditions under which the relative performance of the two forms of assessment might differ (Markon, Chmielsewski, & Miller, 2011). Overall, the estimated reliabilities indicate that use of dimensional measures of psychopathology increases reliability by 15%, compared to binary measures. The comparable estimate for validity is an increase of 37% associated with the use of dimensional measures. Disorder type and sample (clinical versus community) did not moderate these phenomena. Given typical observed effect sizes, the increase in validity has substantial implications for reductions in sample size requirements to achieve standard power levels. The increase in variance associated with dimensional ratings also results in measures that are more sensitive to change over time, and the brevity of such instruments, could allow for repeated assessments over a short period of time that could be combined to provide more reliable estimates. Given the psychometric and practical benefits associated with brief, dimensional measures of child and adolescent mental disorder, the diagnostic accuracy of such measures have been routinely evaluated (Sheldrick et al., 2015).

Diagnostic Accuracy Studies

Most studies designed to evaluate the diagnostic accuracy of questionnaires assume that structured interviews are the ‘gold standard’ for classifying mental disorders in children. As a result, the questionnaire is evaluated against structured diagnostic interviews. Although often considered the ‘gold standard’, structured interviews lack perfect reliability (as noted above) and this will have a direct impact on evaluation studies examining the diagnostic accuracy of questionnaires against interviews (Sheldrick et al., 2015). The imperfect reliability of diagnostic interviews and the impact this has on diagnostic accuracy studies of questionnaires is not widely recognized. Low reliability among reference tests (i.e., diagnostic interview) can attenuate the magnitude of the correlation with a questionnaire (Sheldrick et al., 2015). It can also inflate the association, if the error variance across the two instruments is correlated, which is often the case when a single informant is used for both instruments (Sheldrick et al., 2015).

In a comparative study of three diagnostic interviews for children and adolescents, Angold et al. (2012) report widely different prevalence estimates for 1 or more diagnoses, based on the same sample, ranging from 17.9% using the DAWBA to 46.0% using the DISC and 32.2% using the CAPA. The level of agreement (based on kappa) between interview pairs for specific disorders was very low, ranging from 0.13 (anxiety) to 0.57 (ADHD); and between 0.25-0.38 for any diagnoses. The highest accuracy in which one structured interview detected the results of another was a sensitivity and specificity of approximately 85% for either estimate (Angold et al., 2012). Considering this is the highest level of accuracy for which one diagnostic interview can detect the results of another, it seems quite reasonable to expect the accuracy of brief questionnaires to fall within a similar, and most probably, slightly lower range given differences in the mode of administration.

In a recent meta-analysis of the prevalence of child and adolescent mental disorders, Polanczyk et al. (2015) found that diagnostic interview was a significant moderator of prevalence variability, whereas diagnostic algorithm (i.e., DSM version or ICD) was not. Using the K-SADS as the comparison estimate, studies that used CIDI generated estimates 25.12% (CI 95% 6.59–43.64) higher, and studies that adopted Isle of Wight, DICA, and clinical interview generated estimates 22.91% (CI 95% 2.08–43.74), 13.17% (CI 95% 0.96–25.38), and 31.58% (CI 95% 10.98–52.18) lower. The wide variation in prevalence estimates produced by different structured interviews - all claiming to measure mental disorders in children - raises concerns about their validity. It also raises concerns about using structured interviews as the basis for establishing the accuracy of dimensional measures of child and youth mental disorder.

Another major concern associated with structured diagnostic interviews is the potential loss of important and meaningful information. For example, most structured diagnostic interviews provide only binary responses (presence/absence) at the symptom and disorder level. Compounding this loss of information, is the fact that many structured diagnostic interviews now have brief screening questions that determine whether or not a participant will proceed to a full assessment of DSM criteria linked to specific disorders. If a participant screens out of a particular module, we are left with virtually no symptom information and DSM criteria for that specific disorder and potentially other disorders if the screening questions are linked to multiple modules or disorders. This is a very important consideration, especially in the early stages of measurement development and evaluation, when SAMHSA is determining the optimal approach for classifying children and adolescents with SED.

Analyses of the clinical reappraisal study of the NCS-A that compared clinical follow up of the lay-administered Composite International Diagnostic Interview-Adolescent version (CIDI) with the K-SADS (i.e., gold standard) found much greater concordance using symptoms and severity ratings from the CIDI rather than categorical diagnosis (Kessler et al, 2009). Subsequent validation studies of specific disorders from the NCS-A, also incorporated specific symptoms and qualifiers rather than the categorical diagnosis for ADHD (Green et al, 2010), panic and phobic disorders (Green et al, 2011), and distress disorders (Green et al, 2012). This further illustrates the limitation of binary diagnoses and the importance of dimensional measures in achieving optimal reliability and validity for many of these disorders. These findings also suggest that the value of diagnostic interviews could be improved by deriving additional information, including number of symptoms, and dimensional ratings of severity and impairment.

The challenge with dimensional rating scales, however, is establishing a credible threshold or cut-off that can determine, with a certain level of accuracy, the prevalence of SED in the population of US children 0-18 years. Dimensional ratings can be converted to binary classifications and provide more reliable variance and finer discrimination among individuals to establish threshold options to screen for disorder, compared to binary measures. Imposing such thresholds may lead to a loss of valuable information, but the dimensional nature of questionnaires allows for an empirical evaluation of the psychometric adequacy (reliability and validity) of varying thresholds.

Different approaches can be used to establish and evaluate those thresholds. The most common is to compare varying thresholds applied to the questionnaire against a ‘gold standard’, which is most often a structured or semi-structured diagnostic interview (Sheldreick et al., 2015). Thresholds are then established that result in the greatest diagnostic accuracy (i.e., greatest sensitivity and specificity), relative to the ‘gold standard’. An alternative approach may be to draw on already existing data, such as meta-analytic studies (Polanczyk et al., 2015), that provide worldwide prevalence estimates of psychiatric disorders in children and adolescents in the general population and establish thresholds on the questionnaire that will result in similar estimates.

Another alternative is to apply empirical methods to establish thresholds. For example, Factor Mixture Modeling (FMM; Lubke & Tueller, 2010), can be applied to both symptom and impairment ratings, to help determine thresholds that can 'probabilistically' classify children with SED (Sheldrick et al., 2015). The concurrent and predictive validity of these classifications can be evaluated using independent, external criteria across independent samples. Given additional considerations for the assessment of childhood mental disorders, such as respondent and age effects, FMM can be applied separately to specific respondents and age groups, to determine the extent to which there is convergence in thresholds. To address the increased likelihood of misclassification of children who score at or near the selected thresholds, Sheldrick et al. (2015) recommend interpreting the classifications probabilistically, rather than definitively. There will be greater uncertainty for children scoring at or near the threshold, compared to those that score much higher or lower. In setting thresholds, the costs and benefits of identification and treatment must be carefully considered (Sheldrick et al., 2015). Resources must be sufficiently available for children who score at or above the threshold to receive treatment or further assessment.

Additional Considerations

Informant Effects. Perhaps one of the most consistent findings in the assessment of children's mental health is the low-to-moderate correspondence across informants (Achenbach et al., 1987; De Los Reyes et al., 2015; Kraemer et al., 2003). A recent meta-analysis of 341 studies published between 1989 and 2014 examined cross-informant correspondence of children's internalizing and externalizing mental health concerns, as reported by parents, teachers and youth. Across all studies, the overall cross-informant correlation was 0.28 (95% CI [0.22, 0.33]; p<0.001). Informant correspondence was larger for externalizing concerns r=0.30 (95% CI [0.30, 0.31]; p<0.001), compared to internalizing concerns, r=0.25 (95% CI [0.24, 0.25]; p<0.001). The magnitude of correspondence also varied as a function of (a) informant pair - informants who observed children in the same contexts (i.e., mother-father), exhibited greater levels of correspondence compared to all other informant pairs (i.e., parent-teacher, teacher-child, parent-child); and (b) measurement method - cross-informant correspondence was larger for reports using dimensional scales (e.g., emotional-behavioral questionnaires), relative to categorical and binary measures (e.g., diagnostic interviews) (De Los Reyes et al., 2015). The latter finding is to be expected given greater estimates of reliability and validity associated with dimensional scales, compared to categorical (Markon et al., 2011).

Low-to-moderate levels of cross-informant agreement likely arises as a result of the independent and conjoint influences of (a) child's actual characteristics (e.g., symptoms, impairment, competencies), (b) the context in which the child is observed, (c) the informant's perspective (or biases), and (d) measurement error (Kraemer et al., 2003). One of the main assumptions of a multi-informant approach is that each informant provides a unique and valid perspective of the child's mental health concerns. Consequently, optimal selection of informants should avoid high cross-informant correspondence (or collinearity) and attempt to correct deficiencies in one informant's perspective, by incorporating a second informant whose perspective is independent (Kraemer et al., 2003). This lack of correspondence between informants may facilitate a more valid measurement approach by providing the opportunity to empirically derive measures of the child's underlying mental health concern that is distinct from, or free of, other sources of variation attributable to context and perspective (Kraemer et al., 2003). It should be noted however, that these other sources of variation are of potential interest, particularly, if childrens' expression of mental health concerns vary across contexts (De Los Reyes et al., 2015). The reliability and validity of each of these measures (i.e., mental health concern, context and perspective) can be evaluated and validated in independent samples (Kraemer et al., 2003).

Despite hundreds of research studies examining cross-informant agreement, it remains unclear whether a multi-informant approach adds incremental validity to the assessment of child mental disorder and which informant(s) to include (De Los Reyes et al., 2015). Very few studies have examined the incremental validity of multiple informants' reports (i.e., the extent to which each informant's report contributes unique variance in the prediction of criterion outcomes, over and above another informant's report) and those that do suffer from methodological constraints (De Low Reyes et al., 2015). The few studies that exist provide some important findings for consideration, including: (1) Incremental validity varies as a function of measurement method. Comparisons of the incremental validity of dimensional rating scales versus structured diagnostic interviews for the assessment of childhood ADHD suggest that dimensional rating scales provide greater incremental validity. The use of *DSM-IV* structured diagnostic interview did not contribute unique variance over and above dimensional rating scales (Pelham et al., 2005); (2) A multi-informant assessment approach of internalizing, externalizing and total broad band mental health problems that combines parent, teacher and child reports, yields larger correlations with an independent criterion variable (i.e., referral status), relative to a single informant approach (De Los Reyes et al., 2015); and (3) There is evidence of incremental validity of parent and teacher reports in the prediction of ADHD diagnoses (Pelham et al., 2005; Power et al., 1998; Smith et al., 2000) and child self-reports of anxiety, over and above parent reports, in the prediction of anxiety disorder diagnoses (Villabø et al., 2012).

Determining whether a multi-informant assessment approach should be adopted to address SAMHSA's mandate of estimating 12-month prevalence of childhood SED will require careful consideration of the following factors: (1) the ease of identification and recruitment of multiple informants who can provide reliable ratings of the child's mental health concerns, and who observe the child in different contexts to maximize the opportunities to validly detect contextual variations in the expression of child mental health concerns; (2) the age of the child will determine the availability of independent informants; (3) the availability of instruments completed by multiple informants that contain identical items, content, response options and scaling, to allow for comparisons across informants and the empirical integration of multi-informant data in order to produce more reliable and valid assessments of child mental disorder; and (4) cost-benefit analyses that weighs the incremental value associated with multi-informant reports against the added burden, costs and implementation challenges.

Secondary analyses of existing general population surveys of children that contain identical measures administered to multiple informants may help determine the incremental validity associated with a multi-informant approach and inform the selection of informant(s) by age and mental health domain. In such analyses, the criterion validity measures should be constructed blind or independent of all of the informant's reports in order to reduce contamination (De Los Reyes et al., 2015). These analyses may also help establish benchmarks for determining how much incremental validity is warranted to justify the inclusion of multiple informant data in general population studies, and to weigh this against the potential costs and burden.

Based on the existing evidence and considerations mentioned above, we recommend the following: First, across the full age spectrum (i.e., 2-18 years), we recommend obtaining parent/primary caregiver report. This will provide a single informant that is consistent across all ages and that will allow for comparisons. Second, for the younger children (i.e., 2-12 years), we recommend a second informant, who can provide an independent assessment and who has the opportunity to observe the child outside of the home (i.e., teacher or early child care provider). Third, for adolescents (i.e., 10-18 years), we recommend collecting youth self-reports. For a portion of the sample, specifically the 10-12 year olds, the recommendation is to collect data from 3 informants (parents, teacher and youth). This type of design provides the opportunity to examine measurement invariance of the selected instruments across all three informants in the 10-12 year old age group. If there is evidence of invariance, it may provide justification for empirically integrating the multi-informant data (i.e., parent + teacher and parent + youth) to produce a measure of SED across the full age spectrum. Subsequent, evaluation of the reliability and validity of the multi-informant measure and validation in independent, comparable samples will be warranted. This type of design is most feasible if brief dimensional measures are selected of common childhood mental health problems that are identical across informants and age groups. The item pool can be supplemented for each age group, if needed, but a common core set of items must remain constant across ages and informants to allow for comparisons, integration and evaluation.

Age. If maintaining identical instruments across informants and age groups is perceived as an essential methodological requirement, there are only a few instruments available that can address, albeit not perfectly, this requirement. This is a critical consideration, otherwise age and instrument will be confounded, introducing the possibility of methodologically induced differences in the prevalence of SED by age group. As mentioned above, if brief dimensional measures are selected, then the possibility of enhancing the core item pool with additional items that are developmentally sensitive for each age grouping becomes more feasible.

The two instruments that come closest to meeting this requirement are: 1) The ASEBA measures (CBCL, YSR, TRF), which are available across multiple informants for the broadest age group of children ranging from 1½ -18 years of age, and through to adulthood, and 2) The SDQ also covers a broad age range of children to be administered to multiple informants, with psychometric evidence available for children 2-17 years of age. Recognizing that the length of the ASEBA instruments (i.e., CBCL, YSR, TRF), estimated to take about 15 minutes to complete, and costs may preclude its selection for meeting SAMHSAs mandate, there may be opportunities to use empirical methods to reduce the content without comprising the psychometric properties of the instruments. For example, the 12-item Brief Problem Checklist (Chorpita et al., 2010), is a telephone interview administered to parents and youth, that was developed by applying item response theory and factor analyses to the YSR (Abechebach & Rescorla, 2001) and the CBCL (Abechebach & Rescorla, 2001) to a clinical sample of 2,332 children aged 8-12 years of age. The instrument yields a Total Mental Health Problems Scale, along with two sub-scales, Internalizing and Externalizing Scales, which correspond to the most commonly assessed and treated childhood mental health concerns. It was designed to routinely monitor clinical outcomes over time, and may prove to be a promising instrument for consideration and further evaluation in general population studies. The strengths of this instrument include (a) the symptom coverage is broad enough to capture the most common childhood mental health concerns and that are often co-morbid with less common childhood mental disorders; (b) the brevity of its use, estimated to take 1 minute to administer, allows for the opportunity to consider repeated administrations over a short time period (1-2 weeks) to allow for a more reliable and valid assessment of mental health concerns across informants; and (c) strong psychometric evidence to support its use in clinical settings for routine monitoring of clinical progress during the course of treatment. Additional measurement evaluation work is warranted however since it has been primarily evaluated in a clinical setting with youth aged 8-12 years and their primary caregiver, and administered as a telephone interview.

The ASEBA team has also developed the Brief Problem Monitor (BPM; Achenbach, McConaughy, Ivanova & Rescorla, 2011), that is primarily based on the Brief Problem Checklist (Chorpita et al., 2010), with some adaptations, including the inclusion of an Attention Problems Sub-Scale, a broader age group (6-18 years), multiple informant versions (parent, teacher and youth), and the potential for various modes of administration (paper-pencil, computer, online). The psychometric evidence for the BPM is rather limited at this point and appears to be based on the full administration of the CBCL, YSR and TRF and subsequent extraction and evaluation of the subset of items that comprise the corresponding BPM (Achenbach et al., 2011; Piper et al., 2014). To our knowledge, the BPM has not been administered and evaluated as a standalone instrument.

Culture and Language. Given the racial and linguistic diversity of the US population, it is also important to consider the extent to which the selected instruments have been translated, administered and evaluated cross-nationally. Again, the two instruments that are most widely used cross-nationally, and/or have multi-cultural norms available are the ASEBA Measures and the SDQ.

Mode of Administration. Few studies exist that have examined the impact of mode of questionnaire administration (i.e., face-to-face, telephone, self-complete, paper-pencil, computer assisted etc) on the quality and accuracy of data obtained on child mental health concerns. The few reviews that have been conducted suggest (a) higher item response for face-to-face interviews compared to paper-pencil and telephone surveys; (b) greater social desirability biases in face-to-face and telephone surveys, compared to self-administration, but few differences between face-to-face and telephone surveys; and (c) few differences in self-administration of paper-pencil versus electronic questionnaires (Bowling, 2005; Fouladi et al., 2002). There is general consensus that sensitive questions should be asked using self-administration methods, such as audio-computer self-administration and computer assisted self-completion. Self-administration methods do contribute to higher levels of reporting sensitive data and less social desirability bias, compared to interview methods (Bowling, 2005; Brener et al., 2003; Johnson et al., 2001; Silber & Rosenthal, 1986; Tourangeau & Smith, 1996; Turner, Ku, Rogers et al., 1998; Weeks, 1992). In terms of acceptability, youth (12-25 years) prefer self-administered instruments compared to interview based formats (Bradford & Rickwood, 2012) and the limited available evidence that contrasts paper-pencil versus web-administration among children and adolescents suggest very few differences (Eaton et al., 2010; van de Looij-Jansen & De Wilde, 2008). Studies examining respondents' preferences more generally report that individuals prefer face-to-face interviews to telephone interviews, and electronic self-completion to paper self-completion questionnaires (Bowling, 2005).

In selecting a mode of administration, there are many factors that need to be taken into account including feasibility, efficiency, and cost-effectiveness of each mode and for each informant, along with the cognitive burden imposed on respondents by the mode, control over the quality of the data collected and the process of data collection and capture. Presumably, asking about mental health concerns of one's child, or about oneself is considered sensitive in nature and may be more conducive to electronic, self-administration modes, either through in home computer assisted self-interview techniques or web-based/internet based collection methods (with or without audio assistance). If the assessment of SED is going to be combined with existing/ongoing household surveys then a computer assisted self-administered interview in the home may suffice, with the capacity to include audio assistance to address literacy issues in youth and adult caregivers. Alternatively, an internet based/ mobile device approach may also be worth consideration if sampling methods are available to enlist a representative sample of children, youth and their families using mobile devices or other electronic mechanisms (e-mail). Several features of web-questionnaires make them appealing including the potential to reach geographically dispersed populations rather efficiently, interactive data capture with rapid data checking and adjustments, the potential for instant feedback in the form of summary statistics of individual responses, and the potential to randomize survey questions to address methodological research questions (Ekman & Litton, 2007; Fricker & Schonlau, 2002; Jackob & Zerbak, 2006; Wyatt, 2000). One of the challenges however, may be relatively low response rates (~40%) and non-response bias (Cook et al., 2000), which are common concerns across all modes of administration and general population survey designs.

Broad based mental health measure. SED in childhood is defined as the presence of current or past-year diagnosable mental, behavioral, or emotional disorder of sufficient duration and severity to meet *DSM-III-R* (or equivalent) diagnostic criteria. Based on this definition, an empirically validated dimensional measure of emotional and behavioral symptoms that defines the most common disorders in childhood, and are often co-morbid with less prevalent disorders, could potentially address SAMHSA's mandate. This recommendation is made on the basic understanding that SAMHSA is not interested in producing disorder specific prevalence estimates, but rather an overall estimate of SED.

By taking a more global approach to the assessment of SED (rather than specific diagnoses), it will alleviate the reporting of extensive co-morbidity that characterizes many previous national household surveys. Structured diagnostic interviews generate as many as 25 categorical diagnoses. Overlap in diagnostic criteria across disorders can contribute to double- or triple-counting the same symptom criteria across different diagnostic modules. As such, individual children are represented in multiple categories leading to an inflation in prevalence rates in any one disorder. For example in the only nationally representative sample of U.S. adolescents, the National Co-morbidity Adolescent Survey (NCS-A), 42% of youth met criteria for more than one lifetime disorder (Merikangas et al, 2010). Broad mental health questionnaires attempt to minimize similar symptom descriptors and cross-loadings of items to multiple disorder categories. Factor analyses and other sophisticated statistical approaches investigating the underlying structure of childhood and adolescent psychopathology have shown that there are between 2-3 underlying factors that tap common manifestations of these conditions (Achenbach et al., 1987; Blanco et al., 2013; Kessler et al., 2009), and lend support to a more global assessment of childhood psychopathology.

The importance of measuring rarer disorders and symptoms, such as mania, hypomania, psychotic symptoms, suicidality, eating disorders, was discussed at the expert panel meetings in September and November of 2014 (SAMSA, Center for Behavioral Health Statistics and Quality, 2014a; SAMSA, Center for Behavioral Health Statistics and Quality, 2014b). These are disorders that require specialized services and each case requires many resources. The high co-morbidity of these conditions with more common mental disorders and the symptom overlap in diagnostic criteria across these disorders are still likely to result in the identification of these children and adolescents, using broad based mental health questionnaires. Symptom screeners for these conditions that are included in most of the diagnostic interviews could also be added to ensure that youth with these conditions are covered in surveys of SED.

DSM-5. The introduction of the *DSM-5* childhood disorders has not led to major changes in the characterization of most of the disorders that would be considered under the rubric of SED. One exception is the extension of the age of onset of ADHD from 7 to 12 years that has led to an increase in the prevalence rate of ADHD from 7.38% (DSM-IV) to 10.84% (DSM-5). The later onset group was comparable in terms of severity and patterns of co-morbidity to the earlier onset group, but were more likely to be from lower income and ethnic minority families (Voort et al., 2014). Aside from ADHD, the aggregate anxiety, mood, and behavior disorder criteria do not differ substantially between the two diagnostic systems.

Auxiliary information and conditions. Whenever possible, auxiliary sources of information should be used to evaluate the validity of the SED classification. Indices of teacher-reported difficulties, parent and youth's global perception of mental health problems and need for professional help, contacts with health systems and community agencies and ongoing pharmacological and psychological treatment for mental or behavioral problems should be included wherever possible. Other considerations might include linkages with prescription and health services administrative databases and with education, social service and justice system databases. By taking a global approach to SED (not disorder specific), the identification of independent, construct validity measures may be easier, given the limited number of risk factors that are disorder specific (Shanahan et al., 2008).

There is abundant evidence that substance use disorders, developmental conditions, chronic physical health conditions and the familial and social context contribute as much to impairment as individual mental disorders. Therefore, even though they are excluded from the SED definition, they should be assessed in order to identify the full spectrum of conditions leading to impairment in children and adolescents. Children with multiple morbidity comprise the majority of children and youth receiving mental health services in the general population. More research on the attributable risk of medical comorbidity, and familial/social factors to general impairment and disability beyond individual disorders is needed in order to provide a more comprehensive picture of the correlates of SED in the general population.

Summary & Recommendations

Based on our review of the available evidence and the constraints which SAMHSA is working under to expand the collection of behavioral health data to include the measurement of SED in children, we recommend the use of a brief, dimensional measure of emotional and behavioural symptoms that can be self-administered to multiple informants and repeated over a very short time interval (1-2 weeks) to disaggregate measurement error. The instrument must contain a core set of identical items, content and scaling across informants and age groups, but can be supplemented with additional items that may be more developmentally sensitive. In terms of informants, we recommend that one parent completes the assessment for the full age spectrum (2-18 year olds), and that the 2nd informant varies depending on age. Among the 2-12 year olds, we recommend obtaining teacher or early child care provider assessments and among 10-18 years collecting youth self-reports. It is important to note, with this study design, 10-12 year olds, will have assessments completed by 3 informants which would provide the opportunity to examine measurement invariance across informants in the age group. If there is evidence of invariance, it may provide justification for empirically integrating the multi-informant data (i.e., parent + teacher and parent + youth) to produce a measure of SED across the full age spectrum. Empirical integration and evaluation of multi-informant and repeated data within a 1-2 week interval will be required and will likely produce more reliable and valid assessments of SED in children. Given the sensitive nature of the information SAMHSA is seeking to collect, we recommend computer assisted self-interview with audio assistance in the home. This can be done through the internet, a mobile device assessment or during in-home data collection for other surveys, such as the NSDUH.

Thresholds will need to be established for classifying children with SED probabilistically. We recommend applying empirical methods, such as Factor Mixture Modeling Techniques, that combine the assessments of emotional and behavioral symptoms with the ratings of impairment to classify children. The concurrent and predictive validity of these classifications can be evaluated using independent, external criteria across independent samples. Given additional considerations for the assessment of childhood mental disorders, such as respondent and age effects, FMM can be applied separately to specific respondents and age groups, to determine the extent to which there is convergence in thresholds. Existing data may be currently available to begin to develop and evaluate these methods of classification.

The recommendations made above are heavily determined by three primary considerations: (1) the psychometric evidence supporting the use of brief, dimensional measures of emotional and behavioral symptoms; (2) the presumptive practical and financial benefits associated with administering brief dimensional measures, including cost, length, burden on respondent, training and versatility (i.e., can be administered in a variety of settings to multiple informants using various modes of administration); and (3) the importance of disaggregating sources of variation in ratings of childhood emotional and behavioral symptoms (i.e., informant, context, error, etc), in an attempt to get more valid and reliable ratings of childhood disorder (see Table).

The available instruments that come closest to meeting our recommendations are the ASEBA instruments (CBCL, YSR, TRF, and BPM) and the SDQ.

We would like to conclude by qualifying that our recommendations are specific for meeting an explicit measurement objective –the measurement of childhood SED at the population level – and do not apply to clinical contexts where structured and semi-structured interviews serve broader objectives of diagnosis and treatment planning. SAMHSAs restricted measurement objective is very specific and the goal is to maximize reliability and validity of the assessment data.

Table. *Practical and Methodological Considerations for the Selection of an Instrument to Measure Childhood SED in the US Population*

	Structured, Diagnostic Interview	**Broad Based Dimensional Mental Health Questionnaire**
PRACTICAL CONSIDERATIONS		
Administration Time (respondent burden)	~ 1-1.5 hours	~5-20 minutes
Cost to Use	Dependent on instrument but the administration time and training requirements adds substantially to cost	Dependent on instrument
Need permission to adapt	Dependent on instrument; majority are copyright	Dependent on instrument; majority are copyright
Ease of Administration		
Mode	Single mode is most common: face-to-face	Multiple modes available (telephone, paper-pencil, computer)
Setting	Restricted primarily to in home	Multiple settings
Training	Requires significant training	No training needed
METHODOLOGICAL CONSIDERATIONS		
Psychometric Evidence		
Reliability	Test-retest range of intra-class kappa: 0.4-0.6	Test-retest range of intra-class correlation: 0.7-0.9 (Markon et al. (2011) estimate dimensional measures increase reliability by 15% compared to binary measures)
Validity		Markon et al. (2011) estimate dimensional measures increase validity by 37% compared to binary measures
Established Cut-offs/Algorithms	Yes	Dependent on instrument but will need validation if impairment ratings will be combined to classify children with SED
DSM Criteria	Yes	Dependent on instrument; but far fewer of them have DSM oriented scales
Multi-informant	More restricted	Yes
Age Coverage	More restricted	Broad
Identical Item/Content Coverage Across Ages & Informants		Dependent on instrument; but more viable

	Structured, Diagnostic Interview	**Broad Based Dimensional Mental Health Questionnaire**
Comprehensive Coverage of Diagnoses & Clinical Characteristics	Yes –includes and combines symptoms, impairment, age of onset, recurrence	No
Cross-National Use & Evaluation	More limited	Dependent on instrument but several (ASEBA and SDQ) widely used cross-nationally and available in many languages
Repeated Assessments	Much more difficult given respondent burden	Easily administered on repeated occasions
Suitable for Big Data	No	Yes
Sensitivity to Change	Less sensitive	Yes
Dimensional Ratings	No	Yes

References

Angold, A., Erkanli, A., Copeland, W., Goodman, R., Fisher, P. W., & Costello, E. J. (2012). Psychiatric diagnostic interviews for children and adolescents: a comparative study. *Journal of the American Academy of Child & Adolescent Psychiatry, 51*(5), 506-517.

Achenbach, T. M., McConaughy, S. H., Ivanova, M. Y., & Rescorla, L. A. (2011). Manual for the ASEBA Brief Problem Monitor (BPM). *Research Center for Children, Youth, and Families. University of Vermont. Retrieved from http://www. aseba.org*

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

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: implications of cross-informant correlations for situational specificity. *Psychological Bulletin, 101*(2), 213-232.

Blanco, C., Krueger, R. F., Hasin, D. S., Liu, S. M., Wang, S., Kerridge, B. T., & Olfson, M. (2013). Mapping common psychiatric disorders: structure and predictive validity in the national epidemiologic survey on alcohol and related conditions. *JAMA Psychiatry*, *70*(2), 199-207.

Bowling, A. (2005). Mode of questionnaire administration can have serious effects on data quality. *Journal of Public Health, 27*(3), 281-291.

Boyle, M., Duncan, L., Georgiades, K., Bennet, K. J., Gonzalez, A., Van Leishout, R. J., Szatmari, P., MacMillan, H. L., Kata, A., Ferro, M. A., Lipman, E. L., & Janus, M. (2016). Classifying child and adolescent psychiatric disorder by problem checklists and standardized interviews. (Under review).

Bradford, S., & Rickwood, D. (2011). Psychosocial assessments for young people: a systematic review examining acceptability, disclosure and engagement, and predictive utility. *Adolescent health, medicine and therapeutics*, *3*, 111-125.

Brener, N. D., Billy, J. O., & Grady, W. R. (2003). Assessment of factors affecting the validity of self-reported health-risk behavior among adolescents: evidence from the scientific literature. *Journal of Adolescent Health*, *33*(6), 436-457.

Chorpita, B. F., Reise, S., Weisz, J. R., Grubbs, K., Becker, K. D., & Krull, J. L. (2010). Evaluation of the Brief Problem Checklist: child and caregiver interviews to measure clinical progress. *Journal of Consulting and Clinical Psychology*, *78*(4), 526-536.

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

Cook, C., Heath, F., & Thompson, R. L. (2000). A meta-analysis of response rates in web-or internet-based surveys. *Educational and Psychological Measurement*, *60*(6), 821-836.

De Los Reyes, A., Augenstein, T. M., Wang, M., Thomas, S. A., Drabick, D. A., Burgers, D. E., & Rabinowitz, J. (2015). The validity of the multi-informant approach to assessing child and adolescent mental health. *Psychological Bulletin, 141*(4), 858-900.

Deighton, J., Croudace, T., Fonagy, P., Brown, J., Patalay, P., & Wolpert, M. (2014). Measuring mental health and wellbeing outcomes for children and adolescents to inform practice and policy: a review of child self-report measures. *Child and Adolescent Psychiatry and Mental Health*, *8*(1), 1-14.

Eaton, D. K., Brener, N. D., Kann, L., Denniston, M. M., McManus, T., Kyle, T. M., & Ross, J. G. (2010). Comparison of paper-and-pencil versus Web administration of the Youth Risk Behavior Survey (YRBS): risk behavior prevalence estimates. *Evaluation Review*, *34*(2), 137-153.

Egger, H. L., Erkanli, A., Kee;er, G., Potts, E., Walter, B. K., & Angold, A. (2006). Test-retest reliability of the Preschool Age Psychiatric Assessment (PAPA). *Journal of the American Academy of Child and Adolescent Psychiatry*, *45*, 538-549.

Ekman, A., & Litton, J. E. (2007). New times, new needs; e-epidemiology. *European Journal of Epidemiology*, *22*(5), 285-292.

Federal Register (1993). Substance Abuse and Mental Health Services Administration, Center for Mental Health Services. Definition of children with a serious emotional disturbance. Federal Register, 58(96), 29425.

Fleiss, J. L., & Cohen, J. (1973). Equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, *33*, 613-619.

Fouladi, R. T., Mccarthy, C. J., & Moller, N. (2002). Paper-and-pencil or online? Evaluating mode effects on measures of emotional functioning and attachment. *Assessment*, *9*(2), 204-215.

Fricker, R. D., & Schonlau, M. (2002). Advantages and disadvantages of Internet research surveys: Evidence from the literature. *Field Methods*, *14*(4), 347-367.

Goodman, R. (1999). The extended version of the Strengths and Difficulties Questionnaire as a guide to child psychiatric caseness and consequent burden. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 40(5)*, 791-799.

Goodman, R. (2001). Psychometric properties of the strengths and difficulties questionnaire. *Journal of the American Academy of Child & Adolescent Psychiatry*, *40*(11), 1337-1345.

Green, J. G., Avenevoli, S., Gruber, M. J., Kessler, R. C., Lakoma, M. D., Merikangas, K. R., & Zaslavsky, A. M. (2012). Validation of diagnoses of distress disorders in the US National Comorbidity Survey Replication Adolescent Supplement (NCS-A). *International Journal of Methods in Psychiatric Research*, *21*(1), 41-51.

Green, J. G., Avenevoli, S., Finkelman, M., Gruber, M. J., Kessler, R. C., Merikangas, K. R., ... & Zaslavsky, A. M. (2011). Validation of the diagnoses of panic disorder and phobic disorders in the US National Comorbidity Survey Replication Adolescent (NCS-A) supplement. *International Journal of Methods in Psychiatric Research, 20*(2), 105-115.

Green, J. G., Avenevoli, S., Finkelman, M., Gruber, M. J., Kessler, R. C., Merikangas, K. R., ... & Zaslavsky, A. M. (2010). Attention deficit hyperactivity disorder: concordance of the adolescent version of the Composite International Diagnostic Interview Version 3.0 (CIDI) with the K-SADS in the US National Comorbidity Survey Replication Adolescent (NCS-A) supplement. *International Journal of Methods in Psychiatric Research*, *19*(1), 34-49.

Jackob, N., & Zerback, T. (2006). Improving quality by lowering nonresponse: A guideline for online surveys. *Cadenabbia, Italy: Quality Criteria in Survey Research VI*.

Johnson, T. P., & Mott, J. A. (2001). The reliability of self-reported age of onset of tobacco, alcohol and illicit drug use. *Addiction*, *96*(8), 1187-1198.

Kersten, P., Czuba, K., McPherson, K., Dudley, M., Elder, H., Tauroa, R., & Vandal, A. (2016). A systematic review of evidence for the psychometric properties of the Strengths and Difficulties Questionnaire. *International Journal of Behavioral Development*, *40*, 64-75.

Kessler, R. C., Avenevoli, S., Green, J., Gruber, M. J., Guyer, M., He, Y. & Merikangas, K. R. (2009). National comorbidity survey replication adolescent supplement (NCS-A): III. Concordance of DSM-IV/CIDI diagnoses with clinical reassessments. *Journal of the American Academy of Child & Adolescent Psychiatry*, *48*(4), 386-399.

Kraemer, H. C. (2014). The reliability of clinical diagnoses: state of the art. *Annual Review of Clinical Psychology*, *10*, 111-130.

Kraemer, H. C., Kupfer, D. J., Clarke, D. E., Narrow, W. E., & Regier, D. A. (2012). DSM-5: how reliable is reliable enough? *American Journal of Psychiatry*, *169* (1), 13-15.

Kraemer, H. C., Measelle, J. R., Ablow, J. C., Essex, M. J., Boyce, W. T., & Kupfer, D. J. (2003). A new approach to integrating data from multiple informants in psychiatric assessment and research: Mixing and matching contexts and perspectives. *American Journal of Psychiatry*, *160*(9), 1566-1577.

Kraemer, H. C., Periyakoil, V. S., & Noda, A. (2002). Kappa coefficients in medical research. *Statistics in Medicine*, *21*(14), 2109-2129.

Lubke, G., & Tueller, S. (2010). Latent class detection and class assignment: A comparison of the MAXEIG taxometric procedure and factor mixture modeling approaches. *Structural Equation Modeling*, *17*(4), 605-628.

Markon, K. E., Chmielewski, M., & Miller, C. J. (2011). The reliability and validity of discrete and continuous measures of psychopathology: a quantitative review. *Psychological Bulletin*, *137*(5), 856-879.

Merikangas, K. R., He, J. P., Burstein, M., Swanson, S. A., Avenevoli, S., Cui, L., ... & Swendsen, J. (2010). Lifetime prevalence of mental disorders in US adolescents: results from the National Comorbidity Survey Replication–Adolescent Supplement (NCS-A). *Journal of the American Academy of Child & Adolescent Psychiatry*, *49*(10), 980-989.

National Research Council (2013). *Nonresponse in Social Science Surveys: A Research Agenda.* Roger Tourangeau and Thomas J. Plewes, Editors. Panel on a Research Agenda for the Future of Social Science Data Collection, Committee on National Statistics. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Pelham, Jr, W. E., Fabiano, G. A., & Massetti, G. M. (2005). Evidence-based assessment of attention deficit hyperactivity disorder in children and adolescents. *Journal of Clinical Child and Adolescent Psychology*, *34*(3), 449-476.

Piper, B. J., Gray, H. M., Raber, J., & Birkett, M. A. (2014). Reliability and validity of Brief Problem Monitor, an abbreviated form of the Child Behavior Checklist. *Psychiatry and Clinical Neurosciences*, *68*(10), 759-767.

Polanczyk, G. V., Salum, G. A., Sugaya, L. S., Caye, A., & Rohde, L. A. (2015). Annual Research Review: A meta-analysis of the worldwide prevalence of mental disorders in children and adolescents. *Journal of Child Psychology and Psychiatry*, *56*(3), 345-365.

Power, T. J., Doherty, B. J., Panichelli-Mindel, S. M., Karustis, J. L., Eiraldi, R. B., Anastopoulos, A. D., & DuPaul, G. J. (1998). The predictive validity of parent and teacher reports of ADHD symptoms. *Journal of Psychopathology and Behavioral Assessment*, *20*(1), 57-81.

Shanahan, L., Copeland, W., Costello, J., & Angold, A. (2008). Specificity of putative psychosocial risk factors for psychiatric disorders in children and adolescents. *Journal of Child Psychology and Psychiatry*, *49(1)*, 34-42.

Sheldrick, R. C., Benneyan, J. C., Kiss, I. G., Briggs-Gowan, M. J., Copeland, W., & Carter, A. S. (2015). Thresholds and accuracy in screening tools for early detection of psychopathology. *Journal of Child Psychology and Psychiatry*, *56*(9), 936-948.

Silber, T. J., & Rosenthal, J. L. (1986). Usefulness of a review of systems questionnaire in the assessment of the hospitalized adolescent. *Journal of Adolescent Health Care*, *7*(1), 49-52.

Smith, B. H., Pelham, W. E., Jr., Gnagy, E., Molina, B., & Evans, S. (2000). The reliability, validity, and unique contributions of self-report by adolescents receiving treatment for attention-deficit/hyperactivity disorder. *Journal of Consulting and Clinical Psychology, 68,* 489-499.

Stone, L. L., Otten, R., Engels, R. C., Vermulst, A. A., & Janssens, J. M. (2010). Psychometric properties of the parent and teacher versions of the strengths and difficulties questionnaire for 4-to 12-year-olds: a review. *Clinical Child and Family Psychology Review*, *13*(3), 254-274.

Substance Abuse and Mental Health Services Administration, Center for Behavioral Health Statistics and Quality. *Serious emotional disturbance (SED) expert panel meeting: Operationalizing the SED definition for the production of national and state prevalence estimates, September 8, 2014a, Gaithersburg, MD* [meeting notes]. Rockville, MD: Author, in press.

Substance Abuse and Mental Health Services Administration, Center for Behavioral Health Statistics and Quality. *Serious emotional disturbance (SED) expert panel meeting: Instrumentation and measurement issues when estimating national and state prevalence of childhood SED, November 12, 2014b, Gaithersburg, MD* [meeting notes]. Rockville, MD: Author, in press.

Tourangeau, R., & Smith, T. W. (1996). Asking sensitive questions the impact of data collection mode, question format, and question context. *Public opinion quarterly*, *60*(2), 275-304.

Turner, C. F., Ku, L., Rogers, S. M., Lindberg, L. D., Pleck, J. H., & Sonenstein, F. L. (1998). Adolescent sexual behavior, drug use, and violence: increased reporting with computer survey technology. *Science, 280*(5365), 867-873.

U.S. Government (1993). Federal Register, 58(96), 29323-29520.

van de Looij-Jansen, P. M., & Jan de Wilde, E. (2008). Comparison of web-based versus paper-and-pencil self-administered questionnaire: Effects on health indicators in Dutch adolescents. *Health Services Research, 43(5)*, 1708-1721.

Villabø, M., Gere, M., Torgersen, S., March, J. S., & Kendall, P. C. (2012). Diagnostic efficiency of the child and parent versions of the Multidimensional Anxiety Scale for Children. *Journal of Clinical Child & Adolescent Psychology*, *41*(1), 75-85.

Voort, J. L. V., He, J. P., Jameson, N. D., & Merikangas, K. R. (2014). Impact of the DSM-5 attention-deficit/hyperactivity disorder age-of-onset criterion in the US adolescent population. *Journal of the American Academy of Child & Adolescent Psychiatry*, *53*(7), 736-744.

Wyatt, J. C. (2000). When to use web-based surveys. *Journal of the American Medical Informatics Association*, *7*(4), 426-430.

Weeks, M. F. (1992). Computer-Assisted Survey Information Collection: A Review of CASIC Methods and. *Journal of Official Statistics*, *8*(4), 445-465.