MATHEMATICA POLICY RESEARCH

Assessing Children and Adolescents in Large Scale Surveys

Identifying Prevalence of Serious Emotional Disturbance

Sally Atkins-Burnett

1/29/2016

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 author, 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.

CONTENTS

A. Introduction 1
B. Features of items to consider 1
C. Strengths and challenges of different respondents 3
1. Child respondents 3
2. Teacher respondents 5
3. Parent respondents 8
D. Age-specific challenges 8
1. Infants and toddlers birth to three 8
2. Preschoolers 9
3. Elementary school 10
4. Adolescents 10
E. Sources of measurement error and other challenges in large scale surveys 13
1. Response scale 13
2. Response style and related sources of bias 15
3. Developmental differences 16
4. Abbreviating measures 16
5. Gender, cultural and linguistic differences 17
6. Mode Effects 17
7. Seam Effects 18
F. Pretesting and piloting 19
G. Methods for evaluating measures 21
1. Cognitive interviewing 22
2. Classical test theory approaches 23
3. Item response theory approaches 27
H. Applications of Rasch modeling in social, emotional and behavioral measurement 33
I. Summary of IRT Advantages 38
Conclusions 39
References 41

FIGURE

1 Differential item function in teacher-reported internalizing problem behaviors by grade 7

A. Introduction

The Substance Abuse and Mental Health Services Administration (SAMHSA) has set a goal of estimating the incidence of serious emotional disturbance (SED) in children (birth to age 18). To support meeting that goal, this paper will discuss best practices and what to consider in evaluating available items and measures (and/or in constructing measures) to assure that a survey will reliably and validly identify children with SED. Good measurement begins with well-defined constructs, and that task is already being undertaken by the Standing Committee on Integrating New Behavioral Health Measures into the Substance Abuse and Mental Health Services Administration's Data Collection Programs and SAMHSA. This paper focuses on how those constructs are operationalized and how to evaluate both items and measures.

This report begins with a brief discussion of features of items that need to be considered—whether the items operationalize and adequately represent the constructs and follow good survey practices. This is followed by discussion of the strengths and challenges associated with different reporters and then age-specific challenges. Next, different sources of error, with an emphasis on response biases such as acquiescence and satisficing, are discussed. This is followed by a discussion of what could be gained from pretesting and piloting, methods for evaluating items and scales and identifying if and how measurement could be improved. The report concludes with some recommendations for next steps, many of which will depend on the decisions made about the measures to use, availability of data, and selected analytic models.

B. Features of items to consider

After the selection of well-defined constructs that can be used to identify serious emotional disturbance in children of different ages, review of available measures is often the next step. The validity of the measure will depend on how well the items operationalize the constructs and how accessible the items are. The respondents should all understand the items in the same way. In

reviewing items, it is important to assess whether the items are clearly related to the construct and are written in plain language that would be easily understood in the same way by respondents from different economic, cultural, and linguistic groups included in the sample. In addition to the expert review by the committee and SAMHSA, it may be helpful to ask educational experts such as school psychologists and special educators working with culturally diverse students to review items that parents or students will complete.

Careful examination of how well the items operationalize the constructs of interest and follow good survey practice will provide the first indicator of how well the items may perform in a large, national sample, and may inform revisions before further empirical work. Some measures are developed by clinicians with strong knowledge of the constructs, but limited training in survey or measure development. Items following good survey practices do the following (Groves et al., 2009; Krosnick and Presser, 2010; Lippman, Anderson, Lippman, et al., 2014; Tourangeau and Bradburn, 2010):

- Use simple, familiar words, avoiding clinical terms—even those that are used in the vernacular
- Use simple declarative sentences; the more clauses involved in a sentence, the higher the likelihood of difficulty and differences in understanding
- Use specific words in unambiguous sentences
- Include exhaustive response options that are mutually exclusive
- Are worded positively; negation is more difficult to understand and in surveys can easily lead to double negatives when a response involves "never" or "not at all"
- Ask about one behavior or event at a time; items that are "double-barreled" (i.e., focus on more than one behavior or event) should be revised
- Avoid leading the respondent to a response
- Minimize social desirability by using response scales rather than dichotomies
- Use response scales that have clearly labeled scale points
- Use context as appropriate (reference groups and/or reference time periods) so that respondents are using similar yardsticks in reporting (but take care not to use faulty presuppositions such as "when you talk with your friends on the phone, do you …" assumes

that the respondent talks with friends on the phone, and needs an option indicating that this is not an activity that they do)

- Are ordered in a survey so that the initial questions help in building rapport; they should be pleasant and easy to answer

The combined set of items should include indicators and behaviors of the key dimensions that characterize the construct and allow measurement of severity (adequate representation of the construct). As discussed below, the descriptions will need to reflect the developmental differences in manifestation of SED. The description provided in the item and the type of response scale will need to be accessible to the intended reporter. As much as possible, the cognitive demand of the items should be limited so that lay individuals can easily respond.

C. Strengths and challenges of different respondents

When assessing features of items, it is also important to take into account the likely respondent. This section discusses three types of respondents for SED in children—children, teachers, and parents – and issues relevant to each that should be considered. The section following this will discuss developmental considerations that sometimes overlap with respondent issues.

1. Child respondents

Self-reports are problematic for young children. Children younger than eight years old tend to select the response that they like the most or that think will please the adult, rather than the one that is most indicative of them. Accordingly, measures of young children often have a negative skew as children select the most positive response.

Young children also have difficulty with understanding and differentiating emotions. Even though recent curricula strive to provide an emotion vocabulary (Domitrovich, Cortes, & Greenberg, 2007), young children are not always given a vocabulary for different feelings and so consider emotions on a range of bad/sad to good/happy. Emotion vocabulary develops across the

preschool years becoming more complex and differentiated (Fabes, Eisenberg, Hanish, & Spinrad, 2001), but children vary in rates of development. Emotion understanding is positively associated with general verbal ability and with gender (Bosacki & Moore, 2004). It is also important to consider that vocabulary development and general verbal ability typically vary by maternal education; this could introduce systematic bias into measurement.

In addition, young children are strongly influenced by immediacy in responding, resulting in less stability in measurement. If the day that they are responding, they feel happy then reports will reflect positive emotions and dispositions. If that day they are upset about something, that event will negatively color their responses.

For children 12 and older, the child is generally the best reporter, particularly of internal states. However, some children with serious emotional disturbance may have distorted thinking or may lie even when they have nothing to gain from lying. Obtaining information from multiple informants will provide a more accurate estimate of prevalence. The use of multiple informants is discussed further below in the next section on age-specific challenges.

Many of the child's experiences are not observed by any one reporter. For example, parents do not know all that happens at school and teachers do not know what happens outside the classroom or school setting. Peers often have greater knowledge of a child's behavior and expressed emotions, particularly in adolescence as children seek greater autonomy from adults. Peer sociometrics are good indicators of current and later school functioning (DeRosier & Thomas, 2003; Parker, Rubin, Price, & DeRosier, 1995). However, collection of peer sociometrics in a national survey is not feasible.

Across all age groups, children have more difficulty weighing options than adults would. Even when a child clearly understands what is being asked, making judgments about degree of

severity or frequency of behavior is likely to be challenging because it requires (1) recalling occurrences of that behavior or feeling, (2) weighing how often it occurred (and in some cases weighing duration as well), (3) considering how severe it was when it occurred, and then (4) making a decision about how to characterize the feeling or behavior on the given response scale. Children will vary in their ability to do this. Adolescents will have greater ability in responding, particularly when items are well-constructed following the survey practices discussed earlier.

2. Teacher respondents

Teachers have the advantage of having observed many children of the same age group and are more knowledgeable about the range of behaviors that are typically present within a given age range. Teachers observe children with peers and are often privy to comments from peers on the playground. Teachers also are likely to have some knowledge of the child's behavior on the bus or in the hallways.

While teachers bring advantages as respondents, their reports may also include construct-irrelevant variance that can result in poor estimations of SED prevalence. Teachers observe children in the context of the structured school environment and the level of support in a school or classroom will vary from school to school. Children may have far fewer problems when supported appropriately. In most schools, the teacher changes each year. With teacher change, environmental support and expectations also change. In addition, teachers vary considerably in their observational skills, and the ability to recognize when behaviors are indicative of greater problems, particularly behaviors associated with internalizing problems. Some teachers tend to give children "the benefit of the doubt" or see children's behavior as a reflection of their teaching and rate the children more positively.

The classroom characteristics may influence the teacher reports. Teachers often observe the child only in the classroom setting. Recent evidence supports the importance of classroom

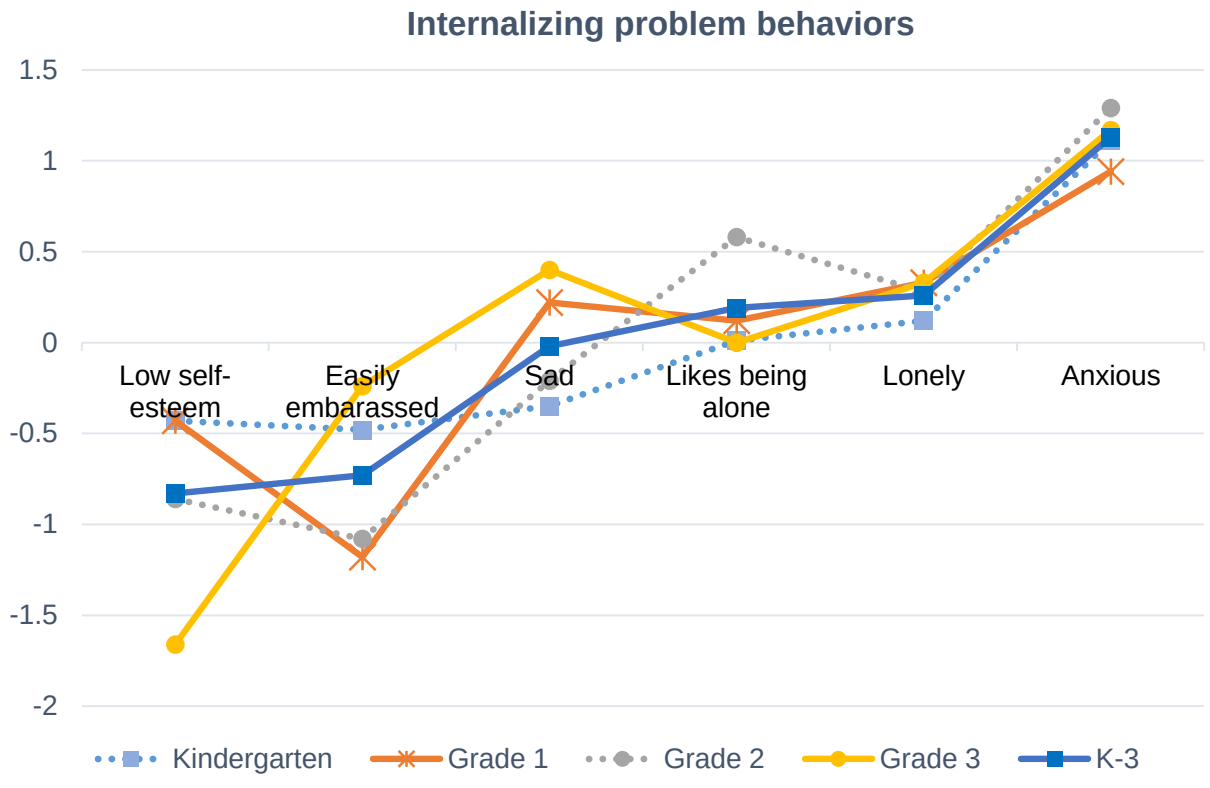
characteristics in shaping social interactions among children in that setting. Classrooms differ in emotional climate, school and classroom norms and behavior management, pedagogical approach, and the relational style of the adults with the children; these factors influence the child's functioning in that environment (Bierman, 2011) and would be reflected in teacher reports of the children.

The amount of variance attributed to the classroom in teacher report measures is greater than classroom level variance in child self-reports, suggesting that a portion of the variance at the classroom level on teacher reports are a function of the teacher's response styles. A national study conducted in low income schools found differences in estimates of children's problems by mode and reporter (Camburn & Rowan, 2004). The study[1] collected externalizing problem behavior data from teacher-reported frequency ratings for individual children, and also individually administered a self-report to children in grades kindergarten through two asking children to report using a visual scale about how hard it was to pay attention and get work completed in reading and math classes, and whether they got into trouble often in these classes. Children in grades three through five responded to the same items in traditional rating scales. The teacher reports of children showed no consistent pattern of change in problem behaviors from kindergarten through grade 5, and HLM models of cross-sectional data attributed 28 percent of the variance to the classroom (teacher) and school. The child self-reports presented a different picture, albeit for different problems. Child reports indicated a decline from kindergarten to grade three in instances of "getting into trouble" and difficulty engaging in different subject areas. The longitudinal HLM models attributed only 8 percent of the variance to the classroom and school level (Camburn & Rowan, 2004).

[1] The study used a two cohort design. Children in kindergarten and the third grade in the same school were sampled and then followed for three years.

There is additional evidence that suggests teachers respond differently to items depending on the ages of children they teach. Examination of the item functioning of teacher reports of social skills and problem behaviors in a sample of kindergarten to third grade students found that the difficulty of endorsing items varied across the grades (Figure 1 illustrates differences in internalizing problems; Atkins-Burnett, 2002). For example, third grade teachers were more likely to endorse low self-esteem in children and less likely to endorse sadness than teachers of earlier grades. This differential item functioning is problematic in looking at incidence rates. Increases in anxiety at different ages, for example, may be a function of the reporter rather than a characteristic of the child.

Figure 1. Differential item function in teacher-reported internalizing problem behaviors by grade



Source: Atkins-Burnett, 2002; Atkins-Burnett, 2003

3. Parent respondents

Parents have the most complete knowledge of how the child has functioned across time and circumstances, though parents may not have a frame of reference for interpreting the appropriateness or severity of behaviors. As with all reporters, parents vary in how astute they are in observing children. However, parent reports of children's social and emotional skills and problem behaviors are more stable than teacher reports across the elementary school years (Van Horn et al. 2007).

D. Age-specific challenges

The information that different respondents are able to report also varies across the ages. In addition, some of the behavioral indicators of SED differ across the age groups. This section brings together issues related to the ages/developmental stages of the children and the respondents who are most likely to be able to rate the behaviors of interest at each stage.

1. Infants and toddlers birth to three

Assessing infant and toddler mental health and identifying disorder and risk is more challenging in the first years of life, but important for planning programs, services and policies (DelCarmen-Wiggins & Carter, 2004). Although more and more children receive child care in out of home settings for many of their waking hours, parents continue to be the most important reporters of the child's well-being. Parents of infants and toddlers are particularly vulnerable to problems related to type of response scale and what is used as the referent, particularly new parents who have not had much experience observing young children or knowledge of child development. However, this is a very important time period for intervening to support parents and children. Infant mental health specialists can obtain positive outcomes even in the face of challenging home situations (Zeanah 2009; Sameroff, McDonough, & Rosenblum, 2004)

Parents of infants and toddlers with limited knowledge of this period of life may make negative attributions about the child, for example, perceiving a crying infant as intentionally acting in opposition to them. These parental attributions are important because they may affect the parent-child relationship. A secure attachment in the first years of life is a consistent predictor of long term functioning (Sroufe, 2005). Children develop in relationships, not in isolation, and nurturing relationships are essential for positive development across domains (Shonkoff & Phillips, 2000; IOM & NRC, 2012).

In the first few years, it may be particularly helpful to include items assessing parental depression, stress, and other risks, as well as protective factors (Hamoudi, Murray, Sorensen, & Fontaine, 2015; Sameroff, Seifer, & McDonough, 2004). Negative early interactions, such as those involving a symbiotic negative emotional relationship between mother and child, can alter the gene expression in young children (Institute of Medicine, 2015). Research has consistently affirmed the cumulative effect of risk factors, and identified indicators that negatively impact early mental health (Sameroff, Seifer, & McDonough, 2004).

2. Preschoolers

Younger children are more susceptible to the influence of the environment and can behave quite differently at home and school, and parents may not regularly observe children interacting with peers or being asked to follow rules and adhere to routines. Therefore it is often helpful to have a teacher and a parent report at this age group. One way to address this is to have a parent ask their child's teacher or child care provider to complete the survey. However, because some children stay at home with a parent full-time, this approach (including both parent and teacher reports) would results in obtaining different information across children. An alternative approach is to survey only parents but to include in that survey items that address salient preschool-based

events (like suspension or expulsion from preschool settings) and that ask about the child's relationship with the teacher/caregiver and peers outside the home.

Differentiating anxiety and depression based on children's behavioral indicators is difficult for both parents and teachers when children are preschoolers and in early elementary school (Kindergarten through grade 2).

3. Elementary school

The surveys conducted with elementary students must address the difficulties that these children may have in reading and comprehending items (see earlier section). Parents are again the optimal reporter for this age group. They observe the child in multiple contexts with varying levels of structure and receive regular reports from the school about behavior as well as information from other adults who supervise the child in sports, recreational or after school settings. In addition, although parents may have less knowledge of child development than teachers, parents receive information about expectations for behavior in elementary school and receive information from teachers across the years. As noted above, longitudinal studies indicate that parent reports are more stable than teacher reports for this age group (Van Horn et al. 2007).

4. Adolescents

By adolescence, most children have the self-awareness and the cognitive sophistication to be able to weigh different sources of information and self-report emotional states and behaviors (Lippman et al., 2014; Caskey & Anfara, 2007). Although adolescents continue to vary in their ability to respond easily and accurately to questions, and may be subject to the same response biases and response style differences as adults are (discussed in more detail in the next section), they are usually able to respond to surveys as well as adults (Scott, 2000). Some adolescents will be more likely to implement Krosnick's (1991) satisficing theory ("I'll take a mental shortcut in responding to this question rather than devote a lot of cognitive effort."). Youth may respond to a

scale without reading the descriptors so the use of different scales with the same number of response options can cause problems in a survey. For more reliable information, Scott (2000) recommends interviewing in the home because at school children may view questions more as a test and they also have to contend with peers asking about what they said.

Social desirability is a particularly important issue with this age group[2] as youth try to fit in with others. Different attempts have been made to decrease social desirability responses. One method that has been tried is to present items worded as "some kids…other kids…" This approach mitigates the selection of socially acceptable responses, but appears to decrease rather than increase the validity of measures (Duckworth & Yeager, 2015).

Another method used to obtain the most valid responses even about sensitive issues has been to assure the child complete privacy by recording an interview that can be played back to the child on a personal listening device (prior large scale studies have used a walkman) and have the child record responses on a form that only has the response categories. In that way, only the child hears the questions (Fargas-Malet, McSherry, Larkin, & Robinson, 2010; Scott, 2000). Evidence from older respondents indicates that the use of self-administered modes reduces socially desirable responding (Moum, 1998). Audio recorded interviews can present some challenges in deciding the appropriate pace for delivering questions in the recording, but allowing children to pause and rewind as needed can help mitigate that concern. Children do appear to be sensitive to the quality of the voice (e.g., pitch, accent) (Scott, 2000), and even young adults respond in different ways to male and female voices asking sensitive questions (Dykema et al., 2012).In addition, the location of the interview at home versus school makes a

[2] Some researchers suggest that younger children may be less susceptible to social desirability bias than adults (Scott, 2000, p. 109)

difference for adolescents in reporting on sensitive topics, even when the highest level of privacy is provided (Griesler, Kandel, Schaffran, Hu, & Davies, 2008).

Audio Computer-Assisted Self-Interviewing (ACASI) is a more technologically-sophisticated approach to ensuring the child privacy in responding to an interview. The ACASI has been used in studies of adolescents around sensitive information including symptoms of depression (for example, National Longitudinal Study of Adolescent Health (Add Health); National Survey on Drug Use and Health (NSDUH, formerly NHSDA); Beck, 2010; Sedlak & Bruce, 2010) and also to investigate more subjective phenomena, such as attitudes and beliefs, of adults (Harmon et al., 2009). The interview can be programmed so that the child selects the voice and/or language of the interview. With the availability of touch screens, ACASI offers additional possibilities including ease of responding for children who are not yet proficient with a mouse or track pad, the potential to use graphically-displayed response options, and response options tailored to the child's characteristics, such as gender, age, and ethnicity. The ACASI on tablets has been used in studies of youth in residential placements to assess a variety of social-emotional and sensitive topics (Beck, 2010). The Middle Grades Longitudinal Study of 2017-2018 (MGLS:2017) conducted by the National Center for Education Statistics is field testing all student measures including measures of social-emotional development and executive functions on tablet computers and plans to use them in the main study.

For adolescents, time referents that align with school calendar years (August/September to May/June, summer vacation) are more helpful than the standard calendar year for obtaining reliable estimates. As with other respondents, recall is stronger for information that is more important from the respondent's perspective (such as the names of friends you hung out with last year) than for other information (such as the grades you received last year) (Groves et al. 2009).

It may also be helpful to interweave measures of functioning that are related to emotional problems, such as sleep (Alfano, Ginsburg, & Kingery, 2007; Ivanenko, Crabtree, & Gozal, 2005; Ribeiro et al., 2012), among the items asking about emotional issues more directly. Adolescents may be more willing to report physiological difficulties than emotional difficulties.

E. Sources of measurement error and other challenges in large scale surveys

As just alluded to, respondents are subject to retrieval problems as well as biases in responding. This next section will discuss some of the common sources of errors in surveys that are particularly relevant to identifying children with SED.

1. Response scale

The two most common response scales used in measures of problem behaviors (both internalizing and externalizing problem behaviors) ask about (1) how characteristic or true a certain attribute or behavior is and (2) how often a behavior or emotion is experienced. Across these two types of scales, the number of responses often range from 1 (“not at all”) to 7 (“extremely true” or “always”), though most measures use three- or four-point scales (e.g., “never,” “sometimes,” “often;” or “not at all true,” “rarely true,” “sometimes true,” “always true”). Reliability is usually stronger for longer scales, with 7-point scales often considered optimal (Kieruj & Moors, 2010).

The mode that will be used to collect data also needs to be considered in selecting a response scale. For example, in a telephone interview each response option needs to be read aloud. To support the respondent, longer response scales should be administered as two questions (for example, first ask, “Would you say you do this seldom, sometimes, or often?”; if the respondent replies “often,” then ask “Is that often or very often?”). Krosnick and Pressner (2010) report that

administering longer response scales in two questions in a telephone interview is faster than going through all seven points on a scale in a single question.

The number of available responses on a scale and the type of responses influence how people respond. For example, some people tend to avoid extreme response (e.g., they do not select response options such as "never" or "always"); in these cases, a five-point scale effectively becomes a three-point scale. However, other respondents will primarily use the ends of the scale freely.

Frequency scales with vague quantifiers ("rarely," "sometimes," "often") can be problematic because respondents use different frames of reference for these modifiers (Schwarz et al. 1991). For example, one person might select "never" for something that happened once, while someone else may select "sometimes" for something with the same frequency. In addition, respondents often use a personal estimate of "average" for a behavior in selecting a response. If it is behavior that they consider low frequency in nature, they may interpret any response above "never" to be an indicator of moderate to high severity and so tend to select a less frequent category.

Use of specific frequencies (such as "not at all," "1-2 times," "3-5 times," "more than 5 times") within a recent time period helps in obtaining more accurate results. Accuracy decreases as memory demands increase. Remembering what happened yesterday is easier than remembering what happened in the previous months. However, particularly for children with SED, some behaviors and situations are highly salient for parents and so can be asked over a longer time period. For example, if children are suspended or expelled from recreational activities, child care, or bussing, parents have to make other arrangements for the child and so tend to remember how frequently this occurred.

Ordering of the response scale also needs to be considered. Respondents tend to select the first response that seems to apply to them and so may endorse different responses if a scale is reversed in presentation (i.e., if "very true for me" appears first rather than "not at all true for me") (Chan 1991).

Fielding a larger and/or well-specified response scale is usually better because you can combine scale points in analysis if they are not tapping differences, while you cannot increase options in analysis. However, the size of the response scale and the specificity of the response can affect some response styles. In addition, response scales that are well-specified, rather than bipolar scales (just anchored at the ends) help to decrease extreme response styles as discussed in the next section (Moors, Kieruj, & Vermunt, 2014).

2. Response style and related sources of bias

Respondents approach surveys and rating scales in different ways. It is important that measure developers are cognizant of these styles so that the selection and ordering of items encourages the respondent to consider each one without increasing fatigue. Particularly when a response scale is not well specified (e.g., using modifiers such as "little," "some," "most," "moderately," "strongly"), some respondents will tend to be more lenient in rating while others are more stringent in rating. In addition, some respondents may rate all items in a given section the same way (response sets and halo effects) without carefully considering each item. Individuals with lower social status (for example, lower income, lower education levels) and those with lower cognitive energy and ability tend to positively endorse items independent of the content. This acquiescence in responding is also more common when the survey is too long and the respondent is fatigued (Krosnick and Presser, 2010).

Extreme response style, that is, usually selecting the ends of a scale, is reported to be a stable trait in individuals independent of the length of the response scale or presence of a middle

category (Kieruj & Moors, 2010), or whether an agree/disagree or an item-specific question is used (Liu, Lee, & Conrad, 2015). In addition, extreme response style is more common in males, and in individuals with lower education and lower income levels (Meisenberg & Williams, 2008). The use of labels for every category mitigates the presence of extreme response styles, but end labeling and bipolar scales increase extreme response scales (Moors, Kieruj, & Vermunt, 2014).

Although, extreme response styles do not seem to be affected by the length of the response scale, longer response scales (8 or 9 categories) increase the likelihood that a midpoint response style will be observed (Kieruj & Moors, 2010).

3. Developmental differences

In selecting measures for children younger than five, the clinical value of behaviors for that age group needs consideration. For example, some behaviors that are indicative of serious emotional disturbance for older children and adolescents (for example, tantrums), are expected behaviors for very young children (e.g., birth to 3). The occurrence of these behaviors may fade more slowly for some children, particularly children with developmental disabilities.

4. Abbreviating measures

Large scale assessments are expensive. Long surveys, particularly if administered by interview, add not only expense but also burden for the respondent and may result in nonresponse to the survey. However, care must be taken in abbreviating scales to assure that constructs are adequately represented both across the dimensions of the construct and also maintain the range of difficulty (in endorsing items) to assure that you are representing the levels of the construct.

5. Gender, cultural and linguistic differences

Differences in social desirability for different groups as well as differences in understanding the meaning of items have been found with measures of social and emotional well-being (Harzing, 2006; He et al., 2015; Ivankeo et al., 2005; Sprachman, Atkins-Burnett, Aikens, & Caspe, 2010). Cultural differences in response styles have also been found (van Herk, Poortinga, & Verhallen, 2004; Chen, Lee, & Stevenson 1995; Marin, Gamba, & Marin, 1992). He and colleagues (2015) examined data from twenty different countries and found both gender differences and cultural differences. Males avoided negative self-description (denial) while females were higher in items placing themselves in a positive light (enhancement). With incidence rates of SED higher among males, this could lead to underestimates of SED from self-reports. The study also found cultural differences in response styles, particularly in the denial items, that were related to the affluence and values of cultures as well as personality traits (He et al., 2015).

6. Mode Effects

The research literature on mode effects for telephone interviewing versus face-to-face interviews is equivocal. Mode can affect some kinds of questions and not others, and the effects can include both linear and nonlinear effects (Jäckle, Roberts, & Lynn, 2010). More limited research has been conducted specifically with children (Vogl, 2013).

Telephone is less expensive than in person interviewing. It is also more easily monitored for quality assurance, and offers greater flexibility in the timing of the interviews. The respondent can complete the interview in the comfort of a familiar environment and does not need to look at someone when they answer more sensitive questions. It is less likely that the interviewer will influence a response by unmonitored facial expression or body language.

Vogl (2013) contends that telephone interviewing may also reduce the power imbalance when adults interview adolescents or children. However, she (Vogl, 2013) reported that children (ages 5-11) were more 'courageous" in face-to-face qualitative interviews than on telephone interviews. There may be less motivation to respond to a voice on the phone, and some people find that the phone requires greater effort to sustain concentration and is more fatiguing (Gillham, 2005; Irvine, 2011; Krosnick and Presser, 2010).

Telephone interviews with younger children would be difficult and potentially unreliable. Children less than 10 years old often need the visual feedback that they receive during an in-person exchange and many do not sustain conversations even with familiar relatives on the phone. In addition, the responses that have been used in child-reported measures have been scaffolded with visual aids (pictures or other visual representations) that allow the child to respond nonverbally by pointing.

7. Seam Effects

In panel studies seam effects occur between time periods. Respondents answer the question about the most recent month in ways that differ from previous months (usually with greater frequency in current month). Memory obviously has a role in this; therefore, the salience of the item as well as the specificity of the time period make a difference in the effect. For adolescents, questions that revolve around school-calendar related events (e.g., between the start of school and Thanksgiving break, or in the first marking period) can support memory for events. Seam effects will differ by the time periods used, length of time between data collection and the type of question. One study investigating seam effects in quantitative responses found decreased seam effects when questions about the same topic were distributed throughout the survey rather than grouped together (Conrad, Rips, & Fricker, 2009). The researchers also examined how providing

answers from an earlier interview affected the seam effect and found that this also reduced the size of the effect. However, these reductions in seam effect did not improve overall accuracy.

Differences between age groups—infant-toddler, preschooler, elementary age, adolescent—and forms can create a kind of seam effect whereby children continue to improve as the age until they are rated on a new form for the next age group. This may be evident when normative scores are used because the children at the ends of a form are not measured as well if classical test theory was used to create the measure (discussed further in the section on methods for evaluating measures). Some of the constructs being assessed differ in indicators across time periods. For example, soiling at ages 2 to 3 is very different from that same behavior at ages 10 to 11 and so does not hold the same meaning across forms. If this item is used for both age groups, it will have differential item functioning by age because the meaning of the behavior differs at each age even if worded exactly the same. Seam effects can be reduced by including some items that will have similar meaning across adjacent age groups and use them to link the forms for those age groups. This would be done using item response theory (IRT; discussed further below) and there are multiple methods for doing this[3].

F. Pretesting and piloting

Once SAMHSA has selected measures or items, it will be important to assure that reliable and valid information can be obtained for the purpose of identifying the incidence of SED. Depending on the measures selected, there may be data available that can be analyzed to examine the reliability and validity. However, the characteristics of items will change when placed in a different order or format, and when the items or response scales are altered. Empirical studies are needed to assure that the items continue to assess the constructs of interest in ways

[3] Chain linking and concurrent calibrations are the two methods that are often used in IRT.

that will inform estimates of SED prevalence at different age groups and among different linguistic and cultural groups. Pilot studies would need to include students who have already been identified as having SED, ideally in each age level and subgroup. If the scales have not previously been evaluated for stability, a test-retest reliability should be evaluated in the pilot studies.

The discussion below describes some of the work that should be done to pretest the items before piloting and then methods that can be used to examine the scales and items after piloting. The size of the sample needed for a pilot will depend on the analytic method that is selected.

If multimode data collection is being considered then a study of mode effects should also be conducted. Ideally this would occur after measure development is complete (that is, with the final measure after piloting). With a large enough sample, it could be incorporated into the pilot randomly assigning reporters to different modes.

Recruitment. Recruiting families with children who have serious emotional disturbance will be particularly challenging. Laws governing both education and medicine strongly protect the confidentiality of this group of children. Even in studies that are funded by the Office of Special Education Programs to research the effectiveness of the Individuals with Disabilities Education Act (IDEA 2004), the families of children with SED have a much lower consent rate. Schools are not allowed to disclose any information about students with disabilities without parental consent unless the study is funded by the Department of Education and intended to evaluate IDEA.

For the pretesting and the pilot, recruitment might take place through support groups such as the National Alliance on Mental Illness (NAMI) and the National Eating Disorders Association; parent newsletters from organizations such as the PACER center (www.pacer.org); state parent

training and information centers for families of children and youth with disabilities; and through websites such as www.conductdisorders.com. The endorsement of parent support organizations may help in obtaining parent cooperation. To determine how well the measures identify children with SED, the pilot would need to recruit individuals with children in other disability categories (in addition to children who are typically developing and those with SED) to assess differentiation. Researchers will need to keep in mind that comorbidity is high among children with disabilities, and SED may not be the child's assigned primary disability for educational purposes. The study might seek the endorsement of the Office of Special Education Programs (OSEP) and request permission to post invitations in schools that serve students with SED (inviting all students into the sample). For psychometric work, the sample does not need to be representative, but does need to have diversity and a distribution of the severity of the constructs. Convenience samples of sufficient size and purposive selection (to capture the diversity of socioeconomic status, culture and language as well as disabilities) will allow examination of the items and scales.

G. Methods for evaluating measures

This section will discuss the use of cognitive interviewing and different analytic approaches to investigate the validity of the items and measures. Both classical test theory and IRT will be briefly described. Classical test theory has been used to develop most survey measures in the last century, however, it has shortcomings relative to measuring social, emotional and behavior measures that will be discussed. The section on IRT summarizes the theory and introduces some of the different IRT models that are available for use in analysis. The next section will provide more in-depth discussion of how Rasch models can inform understanding of social, emotional and behavioral measures.

1. Cognitive interviewing

Cognitive interviewing provides a window into how respondents think about and respond to the items (Lippman et al, 2014). Cognitive interviewing answers such questions as: What behaviors, contexts, or time period do respondents consider when answering? Do they understand the question in the same way as other respondents and as intended? Do respondents use the same or different points of reference? Do they have a different understating of the words in the item? Cognitive interviewing can also evaluate how parents use the response scale. Why did they select "once a month" instead of "2-3 times a month?" What did they consider in selecting their response? Special attention should be given to items that seem to elicit a socially desirable response.

Cognitive interviewing can be especially valuable when adapting or translating items. It can answer the question: Do the words and response scales mean the same thing across different age, cultural, or linguistic groups? It is critical that the clinical or technical meaning of an item on a scale remains in any adaptation or translation. Any non-English version should be carefully evaluated to assure that the sentences have the same meaning, not only across English and the non-English versions, but also across different dialects of a non-English language.

Even within English-speaking samples, the lessons of cognitive interviewing are important. Just because a respondent speaks the language of the instrument, he or she may not fully understand what is being asked. This is particularly true for measures developed by clinicians using clinical terms that have entered the vernacular; those terms may not have the intended meaning for respondents, particularly for respondents with less education. For example, a study in Los Angeles County (Vogel et al, 2008) examined how parents from different cultural and linguistic backgrounds understood the items on English and Spanish measures of social-emotional development and problem behaviors. Using cognitive interviewing techniques, the

researchers found that the parents' understanding of terms differed for a number of items. For example, an item that asked whether the child was "anxious" and "tense" was described by some parents as asking about whether the child was hyperactive or had muscle tension problems. An item asking whether the child "clings to you" was interpreted as a positive behavior by some parents indicating that the child hugged often and was affectionate.

Cognitive laboratory work can include more than one-to-one interviewing. In addition to the individual interviews, the Los Angeles County study included focus groups and card sorts (Sprachman, Atkins-Burnett, Aikens, & Caspe, 2010a; 2010b). The card sort involved providing parents with index cards each displaying an item from different standardized rating scales. The instructions asked parents to consider each item and sort it into one of three envelopes: (1) questions that are **easy** to answer, (2) questions that are **confusing and/or hard** to answer, (3) questions that you are **uncomfortable** answering.

After completing the task, parents discussed social skills and problem behaviors that were of importance or concern to them, and responded to questions about the behaviors that they saw on the cards. For example, all groups reported difficulty understanding "Is difficult to comfort when upset." The parents explained that they were not sure if this was asking how difficult it was for the parent to comfort the child or whether the child was having difficulty with being comforted. Some difficulties were specific to a cultural group. For example, the concept of individual ownership differs across groups and it is not established for many cultures at the preschool age period so parents from those cultures struggled with an item about stealing.

2. Classical test theory approaches

In the last century, most social scales were developed using only classical test analysis. However, as will be discussed in this section, classical test theory has limitations for measures of social and emotional development. It assumes that errors are normally distributed among persons

with constant variance and have an expected value of zero. In classical test theory the observed score is equal to the true score (the latent variable) plus error. In the application of classical test theory, the standard error of measurement is provided for the particular population rather than individual scores. The score obtained on a measure applies only to that test or to items on a parallel form with equivalent item properties. The item difficulties and discrimination are omitted from the model and are justified by their impact on various group statistics (variances and reliabilities) and their relationship to other measures (Embretson, 1999; Embretson & Hershberger, 2014).

While not an issue for large nationally representative samples, for smaller pilots and validity studies it is important to remember that, in classical approaches, the item parameters are influenced by the sample distributions. The indicators of item difficulty and discrimination (i.e., p values and biserial correlations in classical test theory) are influenced greatly by the sample distributions (Embretson, 1999; Embretson & Hershberger, 2014; Reise, 1999; Smith, 1999). The estimate of item difficulty is the proportion of responses that are endorsed (p-value) by the individuals who were in the sample. If an item is administered to a group that is high in the construct being assessed, the p-value might be very high and the item would appear to be very easy/common behavior. If the sample is low in that construct, the p-value is lower and the item would appear to be difficult to endorse or a very rare behavior as most externalizing problem behaviors are for typically developing children.

In a similar manner, the level of the trait assigned to any given person (often referred to as person ability) is dependent upon the difficulty of the sample of items included in the assessment. Person ability in classical test theory is based on the proportion of items answered correctly (or endorsed) by the individual. If the sample of items is relatively easy (easily

endorsed), the person ability will appear high on the trait. A third grader taking a first grade reading test would receive ability scores on that test that would be quite high, even if their reading ability in third grade books was low. Classical test theory tried to mitigate this problem by developing norms for different samples (e.g., by grade or age). Tests would include items targeted for a particular age or grade and individuals would receive a score based on the specific normative group to which they were being compared. This is problematic for social, emotional and behavioral survey measures. Reporters on these surveys may use their own frame of reference in rating the items.

Use of confirmatory models in structural equation models (SEM) has improved what we can learn about the validity of measures and what may be affecting the measurement from a classical approach. SEM can include analyses of the invariance of the thresholds and item parameters and provides reliability based on the model estimates (though still dependent on the sample so the heterogeneity of the sample on the latent trait is important). Under specific conditions, results from a confirmatory factor analysis (CFA) with categorical variables and an IRT analysis would be essentially the same (Thomas, 2011). The item thresholds and loadings in CFA are analogous to the item parameters in IRT (discussed in the next section). Exploratory and confirmatory factor analyses can provide important information about the functioning of a measure and the associations of the variables. CFA is easily incorporated with concurrent regression analyses (that is, structural equation modeling).

When multidimensionality is present, hierarchically structured CFAs and confirmatory bifactor models (sometimes referred to as hierarchical or nested-factor models; bifactor IRT models are discussed further in the next section on item response theory approaches) can be estimated. However, there are restrictions inherent in some CFA models when used to

investigate associations between the constructs and other variables such as sociodemographic characteristics or life outcomes (Brunner, Nagy, & Wilhelm, 2012). Higher order factor models have proportionality constraints[4] that limit looking at associations with other constructs, and large cross-loadings in CFAs will overestimate general factor loadings and underestimate group factor loadings (Brunner, Nagy, & Wilhelm, 2012; Reise, 2012).

Nested factor or bifactor models do not constrain the variance ratios and so are more useful in examining many psychological constructs, though subtest factors are specified as orthogonal, that is, the items can only load on the general factor and one other factor. Bifactor models are increasingly used to investigate psychological constructs such as depression (Gomez & McLaren, 2014), posttraumatic growth (Konkoly, Kovács, and Balog, 2014), perceived competence (using self-reports on the Self-Description Questionnaire –I; Morin, Arens, & Marsh, 2015), and attention deficit hyperactivity disorder (ADHD) and oppositional defiant disorder (ODD; Burns, Moura, Beauchaine, & McBurnett, 2014).

The next section discusses item response theory and the contributions that it makes to evaluating and constructing measures. Item response theory uses different assumptions and methods of estimation than CFA and is more developed in terms of providing information about items and scales that help in refining measurement (Thomas, 2011). IRT, and in particular use of Rasch models, has been useful in examining measures of social and emotional functioning (Atkins-Burnett, 2002; DiStefano, Greer, Kamphaus, & Brown, 2014; McDermott, Watkins, Rovine, & Rikoon, 2013; Schmitt, 2014), including measures used in large scale national surveys (Blumberg, Carle, O-Connor, Moore, & Lippman, 2008; Raczek et al., 1998).

[4] The proportion of variance attributable to the subtests must be the same as the variance attributable to the domain-specific first order factor.

3. Item response theory approaches

Item Response Theory (IRT) is widely used in test development, and has much to contribute to the development of social, emotional, and behavioral measures. Though previously applied to the measurement of academic and cognitive ability (Embretson & Hershberger, 2014), in the past few decades IRT has been applied to measures of personality, affect, and behavior (Blumberg, Carle, O-Connor, Moore, & Lippman, 2008; DeRoos & Allen-Meares, 1998; DiStefano, Greer, Kamphaus, Brown, 2012; Gumpel, Wilson, & Shalev, 1998; Pollack, Atkins-Burnett, Najarian, & Rock, 2005; Reise, Moore, & Haviland, 2008), and is beginning to be used more in clinical assessment (Thomas, 2011). IRT estimates the probability of a correct response on an item based on the ability or trait level of a particular person and the characteristics of a particular item. Unlike the Classical Test Theory which assumes normally distributed variables, IRT assumes a non-linear model with Bernoulli, multinomial, or Poisson sampling. This allows the use of the type of data collected when studying social behaviors and reports of emotional well-being. IRT models provide item characteristic curves (ICCs) that "describe how the probability of responding to an item in a specific way changes as a function of the examinee's position on a latent trait variable" (Reise, 1999, p. 220). A respondent has a 50% probability of endorsing an item when their ability/trait level is the same as the item difficulty level. On items above their trait level they would have a decreased likelihood of endorsing those items. On items below their trait level, they would have an increased likelihood of endorsing that item. For example, a child with ADHD at a trait level equivalent to the difficulty of paying attention in a one-to-one conversation would have a strong probability of receiving a high rating on difficulty paying attention in a large, noisy group.

IRT models are grouped into three categories: Rasch models, 2- parameter logistic models, and 3-parameter logistic models. The Rasch models, sometimes called the one parameter logistic

models, estimate the ability level of the person and the difficulty level of the items and hold the discrimination constant across items (based on the average). In the 2-parameter logistic models, the ability of the person is estimated along with two characteristics of an item: the item difficulty and the item discrimination. In the 3-parameter models, an additional characteristic of an item, the probability of answering correctly by guessing (a "guess" parameter, the lower asymptote parameter), is also estimated (Embretson & Hershberger, 2014; Van der Linden & Hambleton, 1997). The 3-parameter model is seldom used outside of tests that involve multiple choice. The error in a psychological rating scale is as likely to be in the upper asymptote as the lower asymptote and so this model is not an appropriate one for analysis of ratings of social and emotional functioning and behavior.

Bifactor IRT models. In recent years bifactor IRT models (Gibbons & Hedeker, 1992; Gibbons et al. 2007; Gibbons et al. 2008; Gibbons, Rush, & Immekus, 2009; Reise, Moore, & Haviland, 2010; Reise, Morizot, & Hays, 2007; Reise & Waller, 2009; Yung, Thissen, & McLeod, 1999) have been rediscovered as an "effective approach to modeling construct-relevant multidimensionality in a set of ordered categorical item responses." (Reise, 2012, p. 667). For example, bifactor models mitigate the problem of local dependencies, effectively accounting for the residual correlations (Thomas, 2011). The bifactor model allows examination of the potential for subscales (Reise, Morizot, & Hays, 2007), whether interpretable results can be obtained with a unidimensional scale in the presence of some multidimensionality (Reise, Moore, & Haviland, 2010), and has also shown greater prediction to external outcomes compared with higher order models (Chen, West, & Sousa, 2006).

IRT bifactor models differ from CFA bifactor models in how the parameters are estimated and the methods used to evaluate the models. IRT models use the entire response matrix in

calibrating parameter estimates (Reise, 2012). An IRT bifactor model requires item parameter invariance across groups. Bifactor models help to minimize problems found with other analytic approaches in testing for multigroup equivalence (Byrne & van de Vijver, 2010) or resolving dimensionality issues (Reise, Morizot, & Hays, 2007). However, they usually require larger sample sizes than a Rasch model.

Rasch models. Rasch models (one-parameter IRT) are used more often with social, emotional, and behavioral measures than the two- and three-parameter models[5]. In Rasch models the log odds of the probability of a correct response is a function of the difference between the person's ability or the person's level of the trait and the difficulty of the item. In the simplest dichotomous model[6], this is expressed as

$$\log\left(\frac{P_{ni}}{1-P_{ni}}\right) = B_n - D_i$$

in which P_{ni} is the probability of person n correctly answering or endorsing item i, B_n is the measure of person n, D_i is the difficulty of answering or endorsing item I, and log is the natural logarithm. The item discrimination is held constant across the items. Applying the Rasch model to the data allows researchers to construct a system of invariant linear measures, estimate the accuracy of the measures (standard errors), and determine the degree to which these measures and their errors are confirmed in the data using the fit statistics (Linacre & Wright, 2000). A FACETS Rasch model (Linacre, 2010) is used when other factors, such as raters, influence the

[5] Two- and three-parameter models allow the discrimination parameter to vary across items and as a result the item characteristic curves can cross making it difficult to interpret the individual scores relative to the level of the behaviors.

[6] For more complete discussion of the Rasch model, see Wu & Adams, 2007 (available online) or Bond and Fox, 2015, Smith & Smith, 2004, Wilson, 2005.

measurement. FACETS can be estimated to examine how other factors are affecting measurement, such as differences in rater leniency, differential use of categories, and response sets (Englehard, 2013).

Most Rasch models assume unidimensionality, that is, a single dimension is being measured, though multidimensional models have been developed (Wu, Adams, & Wilson, 1997) and used in particular with measures of academic achievement on international studies such as PISA and TIMSS. The multidimensional models are more complex than those that have been used with social, emotional and behavioral data.

Some analytic programs (for example, Winsteps; Linacre, 2003; Linacre & Wright, 2000) offer factor analyses of the residuals that allow the researcher to examine whether additional dimensions are evident in the data. For example, analysis of responses from 317 college students on a self-report instrument measuring Attention Deficit Hyperactivity Disorder (ADHD), Smith and Johnson (2000) applied a Rasch rating scale model and found that seven of the 24 items exhibited misfits greater than one. A factor analysis of the residuals identified secondary variables measuring impulsivity and hyperactivity. Analysis of the secondary dimensions indicated that some students could be high in Hyperactivity while their overall score on the primary variable of Impulsivity/Hyperactivity was below the referral point. Likewise, some students were high on Impulsivity, but low on the overall Impulsivity/Hyperactivity scale. Scores on the secondary dimensions indicate that these individuals may need help with some specific behaviors or aspects of the disorder even when they do not exhibit the full range of the disorder. The fit statistics and analysis of residuals in Rasch analysis identified this important pattern of differences.

Rasch models can help determine what is measurable on a linear scale, which data are useful in describing the latent trait and which are not, how the respondents used the categories in the measure, and whether different groups of respondents utilized the categories of the measures in different ways (Smith, 1992). The major advantages of the Rasch model are the item and sample invariance properties and the interval measurement scale (Hashway, 1998), and the ability to obtain stable estimates with sample sizes less than 200.

As previously stated, Rasch measurement includes models that can be used with a variety of types of data, including dichotomous, Poisson counts, partial credit, ranks, ordered categories (graded response models), and rating scales (Wright & Mok, 2000). For most measures of behavior, a dichotomous (present/absent; yes/no), partial credit, graded, or rating scale models are used. If the step difficulties (between categories) are assumed to be the same across items (e.g., the distance between a rating of 1 to 2 is the same for all items), then the rating scale model may be applied. The partial credit and rank models do not require that the step difficulties are the same (Wright & Mok, 2000). A Rasch graded response model was used in the development of a measure of social competence that was used in the 2003 National Survey of Children's Health (Blumberg, Carle, O-Connor, Moore, & Lippman, 2008). Graded response models are used with data based on ordered categories.

Rasch programs produce a variety of statistics that allow one to examine a measure's reliability and validity, and item characteristics. In Rasch models, estimates of item difficulty and each person's level of the trait are placed on the same scale. The item difficulty estimates the likelihood that an individual will rate highly on the skills or behaviors represented by each scale item. A map illustrating the ordering of item difficulties is produced. This ordering should be consistent with theoretical definitions of the trait. For example, more severe aggressive behaviors

should be harder to endorse than less severe aggressive behaviors. This would be a measure of the construct validity of the scale.

The infit mean square statistic indicates the extent to which the responses to an item are consistent with the hierarchical position of that item in the scale. Infit mean square statistics that are close to 1.0 indicate that individuals are responding to the item in a way that is consistent with the item's location in the scale (Smith & Smith, 2004; Wright & Masters, 1982). For example, in a nonclinical sample, if a child is given a high rating on a particular item, he should receive high ratings on the easier to endorse items below it in the scale, and have a low probability of receiving positive ratings on the more difficult to endorse items above it. Thus, the infit mean square statistic captures the extent to which response patterns are consistent with the difficulty-based rank-ordering of the item obtained in a Rasch analysis. High infit mean square statistics indicate that there is something other than the trait influencing the responses on that item. It may be that the wording of the item is not understood in the same way by all respondents, or that another trait is influencing the responses on a given item. A high level of misfit means that the item as currently written is contributing poorly to the measurement of the defined trait (Bond & Fox, 2015; DeRoos & Allen-Meares, 1993; Smith, 1992; Wright & Masters, 1982). Taken together, the reliability estimates, item difficulties, and fit statistics provide evidence for evaluating the validity and reliability of measures produced by the Rasch model.

The next section will discuss the use of Rasch models to provide evidence of validity and identify some of the common problems in survey measurement of social, emotional, and behavioral functioning.

H. Applications of Rasch modeling in social, emotional and behavioral measurement

Rasch modeling has been used to examine the reliability and validity of different measures of social and emotional functioning (Fogarty, Bramston, & Cummins, 1997; Fox & Jones, 1998; DeRoos & Allen-Meares, 1993; DeRoos & Allen-Meares, 1998; Selner-O'Hagan, Kindlon, Buka, Raudenbush, & Earls, 1998; Zaporozhets et al., 2015), to test the dimensionality (Muller & Wetzel, 1998; Smith, 1999; Smith & Johnson, 2000), to evaluate response scales (Fox & Jones, 1998; Linacre, 1998; Zhu & Kang, 1998; Zhu, Updyke, & Lewandowski, 1996) and response styles (Chen, Lee, & Stevenson, 1995;De Jong, Steenkamp, Fox, & Baumgartner, 2008; Fox & Jones, 1998), and to examine gender (McRae, 1991) and cross-cultural differences in manifestations of social and emotional traits (Wilson, 1994; Zhu & Kang, 1998). Clinical applications of Rasch models have also been discussed (Elliot et al., 2006; Ludlow & Haley, 1996: Reise & Waller, 2009; Smith & Johnson, 2000; Smith, Lawless, Curda, & Curda, 1999; Thomas, 2011; Wright & Masters, 1982).

Evaluating construct validity. As noted earlier, the Rasch model uses responses on each of the items to estimate the probability that items will be endorsed by respondents with different levels of the trait (construct). If the items are measuring the construct reliably, the hierarchy of item endorsability should be consistent with theoretical assumptions about the construct. Inconsistency between theoretical suppositions about the construct and the ordering of the items may indicate a need to refine the theory about the trait, or alternatively, may indicate that the items are understood by respondents in ways that are different than those assumed by the item developers. Further examination of the fit of the items to the model and examination of a factor analysis of the residuals (both of the items and the persons) helps elucidate whether the items are unidimensional and whether they are understood by the respondents in ways that are similar

across respondents. The fit statistics, ordering of the item difficulty and the examination of person fit (and differential item functioning by group) provide initial evidence of the construct validity of these items for measuring the construct.

Response Scale Categories. Rasch models allow researchers to examine the use of the rating scale categories (Engelhard, 2013; Fox & Jones, 1998; Linacre, 1999; Lopez, 1996; Zhu & Kang, 1998; Zhu, Updyke, & Lewandowski, 1997). Rasch models can be used to identify the optimal rating scale categories, identify persons or items that misfit, and provide a means of measuring response style. Measures of constructs in the social and emotional domains often use frequency and Likert-type scales. For example, a four-point frequency scale from "never," "rarely," "sometimes", "almost always" is frequently used as a response scale. For respondents, the psychological distance from "rarely" to "sometimes" may be greater than the distance from "never" to "rarely." Rasch models estimate the distance between categories or steps. Fit statistics on the step calibrations (and graphs of the category probabilities, frequency distributions over categories) can indicate whether two categories are distinct for the respondents and whether they are used by respondents. A variable map with category threshold locations allow researchers to evaluate whether threshold are distinct (should be differences between location) and reflect the order of the categories (Engelhard, 2013; Linacre, 1999). Examination of the item and person separation and fit statistics under alternative scoring can help researchers investigate how the scales were interpreted by the respondents and adopt the most reliable response scale. For example, Fox and Jones (1998) examined different category models and compared the item and person separation as well as the fit statistics to determine the best (i.e., most reliable) scoring scheme for the measure Attitudes toward Child Molestation.

Response Style. Respondents may have different styles of responding. When individuals do not hold strong opinions about an issue, there is a tendency to answer with more agreeable responses (Fox & Jones, 1998). Rasch models can be used to create measures of acquiescence. Response styles may vary according to group membership and differences in response styles can be investigated through the creation of measures of acquiescence. Measures of response style help in differentiating category utilization styles from response to content. When analyzing data, measures of acquiescence can be entered into regression models to control for response style (Fox & Jones, 1998).

Examining gender differences. McRae (1991) used a Rasch analysis with items designed to measure self-esteem to assess both the unidimensionality of the items and sex differences in responses to the items. His initial analysis detected some multidimensionality in the items. He removed the misfitting item from the scale. He analyzed different models in order to test the hypotheses that the relative severity of the items and/or the distance between responses varied by sex. These hypotheses were rejected. Examination of the trait level by sex was not significantly different. This result was contrary to the results of a regression analysis that used a simple sum of the scores on the items assuming unidimensionality. When the summed scores were regressed on sex, women were significantly lower on self-esteem. This result was due to the effect of sex on that single item identified by the Rasch model as not fitting a unidimensional model. The effect was very evident in this scale due to a limited number of items (N=3). However, this study highlights the importance of the metric and of assessing the fit of the model. The conclusion reached by regression analyses with simple summary measures did not reflect the fact that the differences by sex were dependent upon a single item.

Examining cultural differences. Rasch analysis has been used to assess the construct validity of instruments used cross-culturally (Wilson, 1994; Zhu & Kang, 1998). Wilson used both structural equation modeling (SEM) and the Rasch partial credit model to examine the construct validity of the three subscales of Quality of School Life (QSL; Williams & Batten, 1981), an affective measure administered to high school students in Australia and the United States. The SEM approach indicated a problem with parameter invariance across samples on only one of the three subscales (the Teachers subscale) and affirmed the unidimensionality of each of the subscales for the two samples. The Rasch partial credit model was much more sensitive to the differences between the two samples. The Rasch analysis also indicated a fit problem for the Teachers subscale with the U. S. sample, questioning the unidimensionality of this subscale. In all three subscales there were statistically detectable differences in parameter estimates between samples on about half of the items at one or more steps. The majority of these differences represented similar item curves with a shift in difficulty. When the item statistics from the U. S. sample were used to anchor a reanalysis of the Australian sample, the means and standard deviations of the sample were similar for the anchored and unanchored estimates in each of the subscales with the anchored estimates yielding lower mean person estimates.

Clinical cut-off scores. The equal interval scale makes it easier to establish clinical cut-off scores. Smith and Johnson (2000) used a standard setting procedure with Rasch measurement to determine the cut-off points for referrals on a self-report instrument that is administered to college students to screen for ADHD. The researchers asked a convenience sample of eight university staff and faculty clinicians to complete the form by indicating how a student with significant symptomatology would be expected to respond on each of the items. A frequency response scale ranged from 1 (“Never/Not At All”) to 4 (“Very often/Very Much”). The data

from the judges were analyzed using a Rasch rating scale model. Despite relative inexperience of the clinicians used as judges (only one of the judges had participated in ten or more diagnoses), a reliable solution was generated based on the results of six of the judges. The Rasch analyses indicated that the response pattern for two of the judges misfit the model. The model was recalibrated on the remaining six judges. The resulting estimated standard was more stringent than a standard previously estimated using normative information (two standard deviations above sample average) as the standard (Smith & Johnson, 1998 cited in Smith & Johnson, 2000) for the Inattention subscale and more lenient than the normative based standard for the Impulsive/Hyperactive subscale.

Assessing multidimensionality. Linacre (1998) contends that "Rasch analysis followed by a factor analysis of residuals was always more effective at both constructing measures and identifying multidimensionality than direct factor analysis of the original response-level data" (p. 282). Additional dimensions in the data may be due to differences in response styles or the presence of more than one trait in the data. When the dimensions are correlated for the majority of the sample, a factor analysis of the original response level data (rather than the residuals) will not reveal the fact that there is more than one dimension in the data. If these dimensions differ in their relationship to other variables, treating them as a single dimension could obscure relationships with other variables. Cheong and Raudenbush (2000) note

> "High intercorrelations among subscales are a necessary but not sufficient condition to assert unidimensionality. It is also essential that the subscales relate identically or, at least, very similarly, to theoretically linked covariates" (p.9).

DeRoos and Allen-Meares (1993) examined the unidimensionality of the Children's Depression Inventory (CDI) and identified four misfitting items. Further examination of these

items led the researchers to identify problematic wording and conceptual differences in these items. The items used different time frames for rating performance. Though responses on these items might be correlated with depression, these items measured time-relative perceptions of academic competence and are not indicators of three degrees of depression.

I. Summary of IRT Advantages

Rasch offers a measurement model that can be used to develop linear interval scales that measure change and support interpretation of trait levels. The main features of Rasch models include the following:

- Difficulty of the item and person trait levels are estimated using data from the entire matrix. When the data meet the requirements of the model, the estimates are not as dependent upon the sample of items or the sample of individuals as they are in Classical Test Theory.
- Item difficulty and person abilities are placed on the same scale. Item difficulty and ability (trait level) are expressed as log odds. A person with ability equivalent to item difficulty will have a 50% chance of endorsing that particular item (or getting it correct in the case of a test).
- Person abilities (trait levels) can be estimated (and interpreted) even when data are incomplete.
- Rasch models assess the unidimensionality of items. Misfitting items are identified. Misfitting items are items that do not fit a linear interpretation of the construct. These items are influenced strongly enough by other factors that an increase in endorsement of that item does not clearly indicate an increase in the ability level of the latent trait.
- Rasch models provide a linear interval scale when the assumptions of the model are met. Standard errors are provided for each score.
- Rasch analysis can help to identify problems with items, response categories that need to be collapsed or refined, detect multiple dimensions, and identify differential item functioning by subgroups of interest.

With classical approaches, receiver operating curves (ROC) are used to determine the optimal cut point for identifying clinical groups from nonclinical groups. With Rasch approaches, benchmarking is used to set clinical levels based on all of the information in the measure (Bond & Fox, 2015) and the clinical group can be identified so that the distribution is

clear. A combination of Rasch benchmarking confirmed with ROC would provide stronger evidence of the incidence of SED among those who are beyond the cut point.

With the increased interpretability due to the invariance of item structure and the linear interval scale, Rasch models are the recommended measurement models for analysis of social, emotional, and behavioral item functioning. Rasch models have been developed for the kinds of data found in measures of social, emotional and behavioral functioning. Rasch provides many different indices of the reliability and validity of the items and measure that will inform decisions about inclusion and need for revisions. Rasch also provides a method for setting cut points that could be confirmed with ROC, if desired. Multidimensional Rasch models or bifactor models are available when there is multidimensionality.

Conclusions

Similar to the development of any measure, I recommend that SAMHSA follow these steps in order to obtain a strong measure for use in a large scale survey:

1. Clearly define the constructs of interest
2. Operationalize what you would observe that would indicate the presence of the construct(s) of interest.
3. Create or select items/scales that capture the indicators of the construct(s)
4. Evaluate how well the items follow good measurement and survey practices and capture the construct (expert review)
5. Conduct cognitive laboratory work to determine whether individuals with differing backgrounds interpret the items and the scale in similar ways. Are they drawing on the same types of information in responding? Also evaluate whether the response scale is clear and will be interpreted in similar ways across individuals.
6. Collect pilot data for analysis of the psychometric properties of the measures and to identify areas needing revision as well as to determine benchmarks or cut points for identifying children with SED, and test mode effects, if necessary. If planning to abbreviate measures or test alternative indicators or versions of a measure, assure adequate sample sizes receive each version and use clinical opinion as well as data from analysis to determine advisability of abbreviating the measure. You want to assure that constructs are adequately represented.

References

Alfano, C. A., Ginsburg, G. S., Kingery, J. N. (2007). Sleep-related problems among children and adolescents with anxiety disorders. *Journal of the American Academy of Child & Adolescent Psychiatry,* 46 (2), 224-232.

Atkins-Burnett, S. (2002). *Measuring social competence in the early elementary years.* Ann Arbor, MI: Bell & Howell.

Atkins-Burnett, S. (2003). *Measuring social competence in the primary grades.* Presentation at the Society for Research in Child Development Biannual Meeting, Tampa, FL, April 25, 2003.

Atkins-Burnett, S., &. Meisels, S. J. (2001). *Measures of Socio-Emotional Development in Middle Childhood.* Working Paper Series (NCES 200103). Report submitted to the U.S. Department of Education, Office of Educational Research and Improvement. Washington, DC: National Center for Education Statistics.

Beck, A. J. (2010). *Sexual victimization in juvenile facilities reported by youth, 2008-2009*. Darby, PA: Diane Publishing.

Bierman, K. L. (2011). The promise and potential of studying the "invisible hand" of teacher influence on peer relations and student outcomes: A commentary. *Journal of Applied Developmental Psychology*, *32*(5), 297–303. doi:10.1016/j.appdev.2011.04.004

Blumberg, S. J., Carle, A. C., O'Connor, K. S., Moore, K. A., & Lippman, L. H. (2008). Social competence: development of an indicator for children and adolescents. *Child Indicators Research*, *1*(2), 176-197.

Bond, T. & Fox, C. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences (3rd).* Mahwah, NJ: LEA.

Bosacki, S. L., & Moore, C. (2004). Preschoolers' understanding of simple and complex emotions: Links with gender and language. *Sex roles*, *50*(9-10), 659-675.

Brunner, M., Nagy, G., & Wilhelm, O. (2012). A tutorial on hierarchically structured constructs. *Journal of Personality*, *80*(4), 796-846.

Burns, L.G., Moura, M. A., Beauchaine, T. P., & McBurnett, K. (2014). Bifactor latent structure of ADHD/ODD symptoms: predictions of dual-pathway/trait-impulsivity etiological models of ADHD. *Journal of Child Psychology and Psychiatry*, 55(4), 393-401.

Byrne, B. M., & van De Vijver, F. J. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence. *International Journal of Testing*, *10*(2), 107-132.

Camburn, E., & Rowan, B. (2004). Factors contributing to problem behaviors in elementary schools: a hierarchical analysis. CPRE Paper presented at the Biennial Meeting of the Society for Research on Adolescence in March 2004 in Baltimore, MD.

Caskey, M. M., & Anfara, V. A., Jr. (2007). *Research summary: Young adolescents' developmental characteristics*. Retrieved 9/8/2015 from pdxscholar.library.pdx.edu

Chan, J. C. (1991). *Response-order effect in Likert-type scales*. *Educational and Psychological Measurement,* 51, 53`-540.

Chen, C., Lee, S.-Y., & Stevenson, H. W. (1995). Response style and cross-cultural comparisons of rating scales among east Asian and north American students. *Psychological Science*, *6*(3), 170-175.

Chen, F. F., West, S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research, 41*(2), 189-225.

Cheong, Y. F., & Raudenbush, S. W. (2000). Measurement and structural models for children's problem behaviors. *Psychological Methods*, *5*(4), 477-495.

Cheung, G. W., & Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. *Journal of Cross-Cultural Psychology*, *31*(2), 188-213.

Christensen, P., & James, A. (Eds.). (2008). *Research with children: Perspectives and practices*. New York: Routledge.

Conrad, F. G., Rips, L. J., & Fricker, S. S. (2009). Seam effects in quantitative responses. *Journal of Official Statistics*, 25(3), 339-361

De Jong, M. G., Steenkamp, J. B. E., Fox, J. P., & Baumgartner, H. (2008). Using item response theory to measure extreme response style in marketing research: A global investigation. *Journal of marketing research*, *45*(1), 104-115.

DelCarmen-Wiggins, R., & Carter, A. S. (2004). *Handbook of infant, toddler, and preschool mental health assessment*. Oxford University Press, USA.

Denham, S. A., Mitchell-Copeland, J., Strandberg, K., Auerbach, S., & Blair, K. (1997). Parental contributions to preschoolers' emotional competence: Direct and indirect effects. *Motivation and emotion*, *21*(1), 65-86.

DeRoos, Y. S., & Allen-Meares, P. (1993). Rasch analysis: Its description and use analyzing the Children's Depression Inventory. *Journal of Social Service Research*, *16*(3-4), 1-17.

DeRoos, Y., & Allen-Meares, P. (1998). Application of Rasch analysis: exploring differences in depression between African-American and white children. *Journal of Social Service Research, 23*(3-4), 93-107.

DeRosier, M. E., & Thomas, J. M. (2003). Strengthening sociometric prediction: Scientific advances in the assessment of children's peer relations. *Child Development, 74*(5), 1379-1392.

DiStefano, C., Greer, F. W., Kamphaus, R. W., & Brown, W. H. (2014). Using Rasch Rating Scale Methodology to Examine a Behavioral Screener for Preschoolers At Risk. *Journal of Early Intervention, 36*(3), 192-211.

Domitrovich, C. E., Cortes, R. C., & Greenberg, M. T. (2007). Improving young children's social and emotional competence: A randomized trial of the preschool "PATHS" curriculum. *The Journal of Primary Prevention, 28*(2), 67-91.

Duckworth, A. L., & Yeager, D. S. (2015). Measurement matters: assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher,* 44(4), 237-251.

Dykema, J., Diloreto, K., Price, J. L., White, E., & Schaeffer, N. C. (2012). ACASI gender-of-interviewer voice effects on reports to questions about sensitive behaviors among young adults. *Public opinion quarterly, 76*(2), 311-325.

Elliot, R., Fox, C. M., Beltyukova, S. A., Stone, G. E., Gunderson, J., & Zhang, X. (2006) Deconstucting therapy outcome measurement with Rasch analysis: The SCL-90-R. *Psychological Assessment,* 18(4), 359-376.

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment, 8*(4), 341.

Embretson, S. E., & Hershberger, S. L. (Eds.). (2014). *The new rules of measurement: What every psychologist and educator should know*. Mahwah, NJ: Psychology Press.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Psychology Press.

Engelhard, G., & Stone, G. E. (1998). Evaluating the Quality of Ratings Obtained from Standard-Setting Judges. Educational and Psychological Measurement, 58(2), 179-196.

Engelhard, Jr., G. (2013). *Invariant measurement: Using Rasch models in social, behavioral, and health sciences.* New York: Routledge Academic.

Fabes, R. A., Eisenberg, N., Hanish, L. D., & Spinrad, T. L. (2001). Preschoolers' spontaneous emotion vocabulary: Relations to likability. *Early Education and Development, 12*(1), 11-27.

Fargas-Malet, M., McSherry, D., Larkin, E., & Robinson, C. (2010). Research with children: methodological issues and innovative techniques. Journal of Early Childhood Research, 8, 175-192.

Fogarty, G. J., Bramston, P., & Cummins, R. A. (1997). Validation of the Lifestress Inventory for people with a mild intellectual disability. *Research in Developmental Disabilities*, 18(6) 435-456.

Fox, C. M., & Jones, J. A. (1998). Use of Rasch modeling in counseling psychology research. *Journal of Counseling Psychology*, 45(1), 30-45.

Frank, J. W., Binswanger, I. A., Calcaterra, S. L., Brenner, L. A., & Levy, C. (2015). Non-medical use of prescription pain medications and increased emergency department utilization: Results of a national survey. *Drug and alcohol dependence*, *157*, 150-157.

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., Kupfer, D. J., Frank, E., Grochocinski, V. J., & Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement*, *31*(1), 4-19.

Gibbons, R. D., Rush, A. J., & Immekus, J. C. (2009). On the psychometric validity of the domains of the PDSQ: An illustration of the bi-factor item response theory model. *Journal of psychiatric research*, *43*(4), 401-410.

Gibbons, R. D., Weiss, D. J., Kupfer, D. J., Frank, E., Fagiolini, A., Grochocinski, V. J., Bhaumik, D. K., Stover, A., Bock, R. D., & Immekus, J. C. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services*, *59*(4), 361-368.

Gibbons, R.D. & Hedeker, D. (1992). Full-information item bi-factor analysis. *Psychometrika*, 57, 423-436.

Gillham, Bill (2005). *Research interviewing: the range of techniques.* Berkshire UK: Open University Press.

Gomez, R., & McLaren, S. (2014). The Center for Epidemiologic Studies Depression Scale support for a bifactor model with a dominant general factor and a specific factor for positive affect. *Assessment*, 1073191114545357.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., Tourangeau, R. (2009). Survey methodology. Hoboken, NJ: John Wiley & Sons.

Gumpel, T., Wilson, M., & Shalev, R. (1998). An item response theory analysis of the Conners Teacher's Rating Scale. *Journal of Learning Disabilities*, 31(6), 525-532.

Haley, S. M., Ludlow, L. H., & Coster, W. J. (1993). Pediatric Evaluation of Disability Inventory: clinical interpretation of summary scores using Rasch rating scale methodology. *Physical Medicine and Rehabilitation Clinics of North America*, *4*, 529-529.

Hamoudi, A., Murray, D. W., Sorensen, L., & Fontaine, A. (2015). Self-Regulation and Toxic Stress: A Review of Ecological, Biological, and Developmental Studies of Self-Regulation and Stress. OPRE Report # 2015-30, Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Harzing, A. W. (2006). Response styles in cross-national survey research a 26-country study. *International Journal of Cross Cultural Management*, *6*(2), 243-266.

Hashway, R. M. (1998). *Assessment and evaluation of developmental learning: qualitative individual assessment and evaluation models*. Westport, CN: Praeger.

He, J., van de Vijver, F. J., Espinosa, A. D., Abubakar, A., Dimitrova, R., Adams, B. G., ... & Villieux, A. (2015). Socially desirable responding enhancement and denial in 20 countries. *Cross-Cultural Research*, *49*(3), 227-249.

Institute of Medicine (IOM) and National Research Council (NRC) (2015). *Transforming the workforce for children birth through age 8: A unifying foundation.* Washington, DC: The National Academies Press.

Institute of Medicine (IOM) and National Research Council (NRC) (2012). *From Neurons to Neighborhoods: An Update: Workshop Summary.* Washington, DC: The National Academies Press.

Irvine, A. (2011). Duration, dominance, and depth in telephone and face-to-face interviews: a comparative exploration. *International Journal of Qualitative Methods*, 20(3), 202-220.

Ivanenko, A., McLaughlin Crabtree, V., & Gozal, D. (2005). Sleep and depression in children and adolescents. *Sleep medicine reviews* 9(2), 115-129.

Jäckle, A., Roberts, C., & Lynn, P. (2010). Assessing the effect of data collection mode on measurement. *International Statistical Review*, *78*(1), 3-20.

Kieruj, N. D., & Moors, G. (2010). Variations in response style behavior by response scale format in attitude research. *International journal of public opinion research*, 22(3), 320-342.

Konkolÿ Thege, B., Kovács, É., & Balog, P. (2014). A bifactor model of the Posttraumatic Growth Inventory. *Health Psychology and Behavioral Medicine: an Open Access Journal*, *2*(1), 529-540.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5, 213-236.

Krosnick, J. A., & Presser, S. (2010). Question and questionnaire design. In In P. V. Marsden & J. D. Wright (Eds.) *Handbook of survey research, second edition*, (pp. 263-313). Bingley, UK: Emerald Group Publishing.

Lafreniere, P., Masataka, N., Butovskaya, M., Chen, Q., Dessen, M. A., Atwanger, K., Schreiner, S., et al. (2002). Cross-Cultural Analysis of Social Competence and Behavior Problems in Preschoolers. *Early Education and Development*, *13*(2), 201–220.

Leonard Burns, G., Moura, M. A., Beauchaine, T. P., & McBurnett, K. (2014). Bifactor latent structure of ADHD/ODD symptoms: predictions of dual-pathway/trait-impulsivity etiological models of ADHD. *Journal of Child Psychology and Psychiatry*, *55*(4), 393-401.

Linacre, J. M. (1998). Detecting multidimensionality: which residual data-type works best? *Journal of outcome measurement*, *2*, 266-283.

Linacre, J. M. (2003). *A user's guide to WINSTEPS: Rasch-model computer programs.* Chicago: MESA Press.

Linacre, J. M. (2010). *A User's Guide to FACETS Rasch-Model Computer Programs Program Manual 3.70.* Chicago: MESA Press.

Linacre, J. M., & Wright, B. D. *(2000). WINSTEPS: Multiple-choice, rating scale, and partial credit Rasch analysis. Computer software]*. Chicago: MESA Press.

Linden, W. J., & Hambleton, R. K. (1997). Handbook of modern item response theory. *New York*.

Lippman, L. H., Moore, K. A., & McIntosh, H. (2011). Positive indicators of child well-being: a conceptual framework, measures, and methodological issues. *Applied Research in Quality of Life*, *6*(4), 425-449.

Lippman, L., Moore, K. A., Guzman, L., Ryberg, R., McIntosh, H., Ramos, M., Caal, S., Carle, A., & Kuhfeld, M. (2014). *Flourishing children: Defining and testing indicators of positive development*. New York: Springer.

Liu, M., Lee, S., & Conrad, F. G. (2015). Comparing extreme response styles between agree-disagree and item-specific scales. *Public Opinion Quarterly*, Advance Access August 14, 2015, Retrieved from http://poq.oxfordjournals.org/content/early/2015/08/13/poq.nfv034.short?rss=1

Lovaglio, P. G. (2012). Benchmarking strategies for measuring the quality of healthcare: Problems and prospects. *The Scientific World Journal*, pp. 1-13.

Love, J. M., Atkins-Burnett, S. Vogel., C., Aikens, N., Xue, Y., Mabutas, M., Carlson, B., Martin, E.S., Paxton, N., Caspe, M., Sprachman, S., & Sonnenfeld, K. (2009) "Los Angeles Universal Preschool Programs, children served, and children's progress in the preschool year: final report of the First 5 LA universal preschool child outcomes study." Report submitted to First 5 LA. Princeton, NJ: Mathematica Policy Research.

Ludlow, L. H., & Haley, S. M. (1996). Displaying change in functional performance. In G. Englehard, Jr. & M. Wilson (Eds.) *Objective Measurement into Practice Volume 3.* Pp. 3-18, Stamford, CT: Ablex Publishing.

Marín, G., Gamba, R. J., & Marín, B. V. (1992). Extreme response style and acquiescence among Hispanics: The role of acculturation and education. *Journal of Cross-Cultural Psychology*, *23*(4), 498-509.

McAllister, Sue. "Introduction to the use of Rasch analysis to assess patient performance." *International Journal of Therapy and Rehabilitation* vol. 15, no. 11, 2008, pp. 482-490.

McDermott, P. A., Watkins, M. W., Rovine, M. J., & Rikoon, S. H. (2014). Informing context and change in young children's sociobehavioral development–The national Adjustment Scales for Early Transition in Schooling (ASETS). *Early Childhood Research Quarterly*, 29(3), 255-267

McDermott, P., Watkins, M. W., Rovine, M. J., & Rikoon, S. H. (2013). Assessing changes in socioemotional adjustment across early school transitions: New national scales for children at risk. *Journal of School Psychology*, 51(1), 97-115.

McRae, Jr., J. A. (1991). Rasch measurement and differences between women and men in self-esteem. *Social Science Research,* 20, 421-436.

Meisels, S. J., Atkins-Burnett, S., Nicholson, J., & West, J. (1996). *Assessment of social competence, adaptive behaviors, and approaches to learning.* Working Paper Series. Report submitted to the U.S. Department of Education, Office of Educational Research and Improvement. Washington, DC: National Center for Education Statistics.

Meisenberg, G., & Williams, A. (2008). Are acquiescent and extreme response styles related to low intelligence and education? Personality and Individual Differences, 44, 1539–1550.

Moore, K. A., & Lippman, L. H. (Eds.). (2006). *What do children need to flourish?: Conceptualizing and measuring indicators of positive development* (Vol. 3). Springer Science & Business Media.

Moors, G., Kieruj, N. D., & Vermunt, J. K. (2014). The effect of labeling and numbering of response scales on the likelihood of response bias. *Sociological Methodology*, *44*(1), 369-399.

Morin, A. J., Arens, A. K., & Marsh, H. W. (2015). A bifactor exploratory structural equation modeling framework for the identification of distinct sources of construct-relevant psychometric multidimensionality. *Structural Equation Modeling: A Multidisciplinary Journal*, (ahead-of-print), 1-24.

Moum, T. (1998). Mode of administration and interviewer effects in self-reported symptoms of anxiety and depression. *Social Indicators Research*, *45*(1-3), 279-318.

Müller, M. J., & Wetzel, H. (1998). Dimensionality of depression in acute schizophrenia: a methodological study using the Bech-Rafaelsen Melancholia Scale (BRMES). *Journal of psychiatric research*, *32*(6), 369-378.

Parker, J. G., Rubin, K. H. Price, J. M., & De Rosier, M. E. (1995). Peer relationships, child development, and adjustment: A developmental psychopathology perspective. In D. Cicchetti & D. Cohen (Eds.), *Developmental psychopathology, Vol. 1, Theory and methods*, New York: Wiley.

Pollack, J.M., Atkins-Burnett, S., Najarian, M., and Rock, D.A. (2005). *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS–K), Psychometric Report for the Fifth Grade* (NCES 2006–036). Washington, DC: NCES. Department of Education. Washington, DC: National Center for Education Statistics.

Raczek, A. E., Ware, J. E., Bjorner, J. B., Gandek, B., Haley, S. M., Aaronson, N. K., ... & Sullivan, M. (1998). Comparison of Rasch and summated rating scales constructed from SF-36 physical functioning items in seven countries: results from the IQOLA Project. *Journal of Clinical Epidemiology*, *51*(11), 1203-1214.

Reise, S. P. (1999). Personality measurement issues viewed through the eyes of IRT. In S. Embretson & S. L. Hershberger (Eds.) *The new rules of measurement: What every psychologist and educator should know*. (pp. 219-242). Mahwah, NJ: Psychology Press.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*(5), 667-696

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual review of clinical psychology*, *5*, 27-48.

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of personality assessment*, *92*(6), 544-559.

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research*, *16*(1), 19-31.

Ribeiro, J. D., Pease, J. L., Gutierrez, P. M., Silva, C., Bernert, R. A., Rudd, M. D., Joiner, T. E. (2012). Sleep problems outperform depression and hopelessness as cross-sectional and longitudinal predictors of suicidal ideation and behavior in young adults in the military. *Journal of affective disorders*, 136(3), 743-750.

Rips, L. J., Conrad, F. G., & Fricker, S. S. (2003). Straightening the seam effect in panel surveys. *Public Opinion Quarterly*, 67, 522-554.

Sameroff, A. J., McDonough, S. C., & Rosenblum, K. (Eds.). (2004). *Treating Parent--infant Relationship Problems: Strategies for Intervention*. Guilford Press.

Sameroff, A., Seifer, R., & McDonough, S. C. (2004). Contextual contributors to the assessment of infant mental health. *Handbook of infant, toddler, and preschool mental health assessment*, 61-76.

Schmitt, B. A. (2014). *Developing a Teacher Rating Scale of Preschool Student Behavior for Use in an RTI Decision-Making Framework* (Doctoral dissertation, University of Minnesota).

Schwarz, N., Bless, H., Bohner, G., Harlacher, U., & Kellenbenz, M. (1991). Response scales as frames of reference: The impact of frequency range on diagnostic judgements. *Applied Cognitive Psychology*, *5*(1), 37-49.

Scott, J. (2000). Children as respondents: The challenge for quantitative methods. In P. M. Christen & A. James (Eds.) *Research with children: Perspectives and practices*, (pp. 98-119). New York: Psychology Press.

Sedlak, A. J., & Bruce, C. (2010). Youth's Characteristics and Backgrounds: Findings from the Survey of Youth in Residential Placement. Juvenile Justice Bulletin. *Office of Juvenile Justice and Delinquency Prevention*.

Selner-O'Hagan, M. B., Kindlon, D. J., Buka, S. L., Raudenbush, S. W., & Earls, F. J. (1998). Assessing exposure to violence in urban youth. *Journal of Child Psychology and Psychiatry*, *39*(2), 215-224.

Shonkoff, J. P. & Phillips, D. A. (Eds.). (2000). *From Neurons to Neighborhoods: The Science of Early Childhood Development*. Washington, DC: National Academies Press.

Smith Jr, E. V., Lawless, K. A., Curda, L., & Curda, S. (1999). Measuring change in efficacy. *Popular Measurement*, 2(1), 31-33.

Smith, E. & Smith, R., Eds. (2004). *Introduction to Rasch Measurement.* Maple Grove, MN: JAM.

Smith, E. & Smith, R., Eds. (2004). *Introduction to Rasch Measurement.* Maple Grove, MN: JAM.

Smith, E. V., & Johnson, B. D. (2000). Attention deficit hyperactivity disorder: scaling and standard setting using Rasch measurement. *Journal of Applied Measurement,* 1(1), 3-24.

Smith, P. B. (2004). Acquiescent response bias as an aspect of cultural communication style. *Journal of Cross-Cultural Psychology, 35*, 50-61.

Smith, R. M. (1992). *Applications of Rasch measurement.* Chicago: MESA Press.

Smith, R. M. (1999). *Rasch measurement models: interpreting WINSTEPS, BIGSTEPS, and FACETS output.* Chicago: MESA Press.

Sprachman, S., Atkins-Burnett, S., Aikens, N. & Caspe, M. (2010a). Meaning in the method: pretesting methods for a diverse population." Data Collection Methods, Issue Brief no. 3. Princeton, NJ: Mathematica Policy Research,

Sprachman, S., Atkins-Burnett, S., Aikens, N. & Caspe. M. (2010b). *Card sorts, focus groups, and cognitive interviews: three methodologies to improve question development for different cultural and linguistic groups*. Presentation to the American Association for Public Opinion Research Annual Conference, Chicago, May 2010

Sroufe, L. A. (2005). Attachment and development: A prospective, longitudinal study from birth to adulthood. *Attachment & human development, 7*(4), 349-367.

Stone, G. (2004). Chapter 11, Objective Standard Setting in Understanding Rasch Measurement, Lawrence Erlbaum, Edited by Richard Smith.

Stone, G. E. (2000). Objective standard setting (or truth in advertising). *Journal of applied measurement*, 2(2), 187-201.

Thomas, M. L. (2011). The value of item response theory in clinical assessment: A review. *Assessment, 18*(3), 291-307.

Tourangeau, R., & Bradburn, N. M. (2010). The psychology of survey response. In P. V. Marsden & J. D. Wright (Eds.) *Handbook of survey research – second edition*, (pp. 315-346), Bingley, UK: Emerald Group Publishing.

Van der Linden, W. J., & Hambleton, R. K. (1997) *Handbook of modern item response theory.* New York: Springer.

Van Herk, H., Poortinga, Y. H., & Verhallen, T. M. (2004). Response styles in rating scales evidence of method bias in data from six EU countries. *Journal of Cross-Cultural Psychology*, *35*(3), 346-360.

Van Horn, M. L., Atkins-Burnett, S., Karlin, E., Ramey, S. L., & Snyder, S. (2007). Parent ratings of children's social skills: Longitudinal psychometric analyses of the Social Skills Rating System. *School Psychology Quarterly*, *22*(2), 162-199.

Vogel, C. A., Aikens, N., Atkins-Burnett, S., Sama Martin, E., Caspe, M., Sprachman, S. C., & Love, J. M. (2008). "Reliability and Validity of Child Outcome Measures with Culturally and Linguistically Diverse Preschoolers: The First 5 LA Universal Preschool Child Outcomes Study Spring 2007 Pilot Study." Princeton, NJ: Mathematica Policy Research.

Vogl, S. (2013) Telephone versus face-to-face interviews: mode effect on semistructured interviews with children. *Sociological Methodology,* 43(1), 133-177.

Waller, N. G. (1999). Search for structure in the MMPI. . In S. Embretson & S. L. Hershberger (Eds.) *The new rules of measurement: What every psychologist and educator should know*. (pp. 185-218). Mahwah, NJ: Psychology Press.

Waruru, A. K., Nduati, R., & Tylleskär, T. (2005). Audio computer-assisted self-interviewing (ACASI) may avert socially desirable responses about infant feeding in the context of HIV. *BMC medical informatics and decision making*, *5*(1), 24.

Williams, T., & Batten, M. (1981). *The quality of school life*. Melbourne, Australia: Australian Council for Educational Research.

Wilson, M. (1994). Comparing attitude across different cultures: two quantitative approaches to construct validity. In M. Wilson (Ed.) *Objective measurement into practice, Volume 2,* (pp. 271-292). Stamford, CT: Ablex Publishing.

Wilson, M. (2005). *Constructing measures: An item response modelling approach.* New York: Routledge.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement.* Chicago: MESA Press.

Wright, B. D., & Mok, M. (2000) Rasch models overview. *Journal of Applied Measurement,* 1(1), 83-106.

Wu, M., & Adams, R. (2007). *Applying the Rasch model to psycho-social measurement*. Melbourne, AU: Educational Measurement Solutions. Retrieved September 9, 2015 from http://media.metrik.de/uploads/incoming/pub/Literatur/von%20Winfried/RaschMeasurement_Complete.pdf

Wu, M.L., Adams, R.J., and Wilson, M.R. (1997). ConQuest: Multi-Aspect Test Software, [computer program] Camberwell: Australian Council for Educational Research. www.acer.edu.au/quest

Yung, Y.-F., Thissen, D., & McLeod, L. D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika*, 64, 1999, 113-128.

Zaporozhets, O., Fox, C. M., Beltyukova, S. A., Laux, J. M., Piazza, N. J., & Salyers, K. (2015). Refining change measure with the Rasch model. *Measurement and Evaluation in Counseling and Development, 48*(1), 59-74.

Zeanah, C. H. (Ed.). (2009). *Handbook of infant mental health*. New York: Guilford Press.

Zhu, W., & Kang, S. J. (1998). Cross-cultural stability of optimal categorization of a self-efficacy scale: A Rasch analysis. *Measurement in Physical Education and Exercise Science,* 2, 225-241.

Zhu, W., Updyke, W. F., & Lewandowski, C. (1996). Post-hoc Rasch analysis of optimal categorization of an ordered-response scale. *Journal of outcome measurement, 1*(4), 286-304.