APPROVAL SHEET

Title of Dissertation: Information Gathered by Retrospective, Self-Report, Emotional Frequency Items in Children

Name of Candidate: Nicole Whyms Brocato
Doctor of Philosophy 2014

Dissertation and Abstract Approved: _____

Laura Stapleton, Ph.D.
Associate Professor
Human Development and Quantitative Methodology
University of Maryland – College Park

_____

Steven C. Pitts, Ph.D.
Associate Professor

Date Approved: _____

Curriculum Vitae

Name: Nicole Whyms Brocato

Degree and date to be conferred: Ph.D., 2014

Education

- *M.A., Psychology* - University of Maryland, Baltimore County, 2008
  Thesis title: Developing a questionnaire to measure social, tickle, and contagious yawning behavior in normally developing children and children with autism
- *B.A., Psychology* - University of Maryland, Baltimore County, 2005, magna cum laude

Publications and Presentations

- Provine, R., Cabrera, M., **Brocato**, N., & Krosnowski, K. (2011). When the whites of the eyes are red: A uniquely human cue. *Ethology*, *11*(5), 395-399.
- **Brocato, N.** (2010, April) *When a Google Search Isn't Enough…* In R. Siegried (Moderator), *Skills session: Expanding the evaluator's toolkit with technology*. Panel session conducted at the 33rd Eastern Evaluation Research Society Annual Conference: Expanding the evaluator's toolkit, Galloway, NJ.
- Provine, R., Kucharski, K., **Brocato, N.** (2009). Tearing: Breakthrough in Human Emotional Signaling. *Evolutionary Psychology*, *7(1)*, 52-56.
- Provine, R., Krosnowski, K., **Whyms, N**., Whipps, M., Webb, K. (2008). Sadness with and without tears: The tear effect. Poster presentation for the Society for Neuroscience, Washington, DC.
- Provine, R., Emmorey, K., Spencer, R., Mandell, D., **Whyms, N**. (2007). Emoting to people you can neither see nor hear. Poster presentation for the American Psychological Society, Washington DC.

Clinical Training History

- Assistant Clinical Care Manager, BHRS Program, Keystone Human Services, July 2013 – present
- Clinical Pre-Doctoral Intern, BHRS Program, Keystone Human Services, July 2012 – June 2013
- Children's Program Coordinator, Domestic Violence and Sexual Assault Center of Howard County (DVSAC), July 2009 – June 2012
- New Behaviors for Women Program Coordinator, DVSAC of Howard County, September 2010 – present
- Program Assistant, HIV/AIDS Community Collaborative, July 2008 – July 2009
- Domestic Violence Center of Howard County, supervised by Amy Bogdon-Abrams, LCSW-C, Sep 2008 – May 2009
- Medical Services Office, Baltimore City Circuit Court, supervised by Dr. Harriet Siegel-Miller, Sep 2007 – May 2008
- Springfield Hospital Center, McKeldin A Ward, supervised by Dr. Cheryl Zwart, Sep 2006 – May 2007

Teaching Experience

- Teaching assistant, September 2005-July 2008

Honors and Awards

- Phi Beta Kappa, 2005
- Distinguished Achievement Award in Psychology, 2004 & 2005
- Rosalie Tydings Business and Professional Women's Scholarship 2002
- Senatorial Scholarship 2004-2005
- Richard and Roselyn Neville Scholarship 2004
- Partial Scholarship to MBSR training at Omega Institute, 2008

Memberships and Affiliations

- American Psychological Association
- Maryland Psychological Association of Graduate Students
- Association of Family and Conciliation Courts
- Association for Contextual Behavioral Science
- Management team member, Howard County Child Advocacy Center

ABSTRACT

Title of Document: INFORMATION GATHERED BY RETROSPECTIVE, SELF-REPORT, EMOTIONAL FREQUENCY ITEMS IN CHILDREN

Nicole Whyms Brocato, Ph.D., 2014

Directed By: Laura Stapleton, Ph.D.
Associate Professor
Human Development and Quantitative Methodology
University of Maryland – College Park

Steven C. Pitts, Ph.D.
Department of Psychology

Retrospective emotional frequency appraisals are often used in clinical assessment measures, but their suitability for use with children has not been well studied. The aims of this project were to (a) examine whether items that use retrospective frequency structures gather more or less information than items that do not use such structures and (b) examine whether the information gathered by such items differs across children's ages. ***Method.*** Data were gathered from 9- to 12-year-old girls who participated in a larger study of a depression treatment protocol. Two sets of five pairs of items were sampled from two children's depression measures. The item pairs contained one item from each measure. One set of item pairs was matched for content and the use of retrospective frequency structures. The other set was matched for content only. ***Results***. For the first research question, information curves for the two item sets were generated using Samejima's (1969) Graded Response Model (GRM). Visual analyses of the information curves provided inconclusive results as to whether the presence of

retrospective frequency structures is associated with differences in item information levels. The second research question was conducted in two parts. For both, only data from the 9- and 12-year-old participants were analyzed. In the first part, confirmatory factor analysis was used to analyze measurement invariance across the two groups' responses. Theses analyses showed signs of measurement non-invariance in both item sets. The second part of the analyses was conducted by generating separate GRM information curves for the two age groups and conducting visual analyses of the information curves. These analyses showed that the model which had been used throughout the remainder of the study did not fit the 9-year-old group well. They also showed that the 12-year-old group's information curves varied more in height across measures and item sets than did the 9-year-old group's curves. ***Discussion.*** Although the findings failed to shed light on the effects of retrospective frequency structures on children's responding, they highlighted potential differences between the 9- and 12-year-old groups' factor structures and indicated that the 9-year-olds displayed decreased sensitivity to differences in item structure.

INFORMATION GATHERED BY RETROSPECTIVE, SELF-REPORT, EMOTIONAL FREQUENCY ITEMS IN CHILDREN

By

Nicole Whyms Brocato

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, Baltimore County, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in
Human Services Psychology
2014

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

Dedication

To my husband, who helped me remember that simple things like socks can be just as important as complicated things like dissertations.

Acknowledgements

Many thanks to Eliot Shimoff, who convinced me that psychology was my academic home; before Dr. Shimoff, I spent many years searching. I also owe many thanks to Robert Provine and Charlie Catania, who kept me fascinated, challenged, and supported during my later undergraduate studies and early graduate years. Stories and lessons from all remain central to my clinical and academic endeavors.

I am indebted to Steven Pitts for introducing me to the delights of quantitative studies and for answering many, many questions since then. I am equally indebted to Laura Stapleton for answering nearly as many questions, the honor of serving as primary dissertation chair, and tolerating endless revisions to methods and text. This project would not have been possible without either Dr. Pitts or Dr. Stapleton.

Without Robert and Barbara Deluty, graduate school may have been an insurmountable task. The Delutys made sure I believed in myself and the work. More important, they helped me remember all the wonderful things the world has to offer.

I was very fortunate to have an outstanding dissertation committee, which was completed by Laura Ting and Jason Schiffman. Both offered many thoughtful suggestions and frequent words of encouragement.

This project's focus on retrospective frequency structures was in part inspired by Kevin Stark, who generously donated the study's data.

Finally, I am endlessly grateful to my friends and extended family. They have believed in me and put up with me for an impressively long time. As important as the academic support has been, I would not have been able to benefit from that support without the many people who gave their love so generously.

Table of Contents

List of Tables

List of Figures

Introduction

Children are increasingly called upon to provide self-reports as part of the psychological assessment process. Assessment results are used to determine diagnoses and develop treatment options. Parents and other caregivers are often included in the assessment process because they can have insights into children's behavior that the children themselves lack (Mash & Hunsley, 2007). Given that clinical judgment alone is often outperformed by decisions based on measures and algorithms (Dawes, Faust, & Meehl, 1989; Kahneman & Klein, 2009), current best practice guidelines for the assessment of children recommend using a multimodal approach that includes interviews and standardized instruments with the child and at least one caregiver (Mash & Hunsley, 2007).

Children have historically not been included in the assessment process because they have not always been viewed as accurate self-reporters. In the early years of psychological assessment, parents were considered to be the best reporters on children's internal states and behavior (review by Mash & Hunsley, 2007). Although parents are still thought to be better reporters on children's behaviors than children themselves (Scott, 1997), children are increasingly considered to be accurate and consistent informants on their internal states (Amato & Ochiltree, 1987; de Leeuw et al., 2004; Grills-Taquechel & Ollendick, 2008; La Greca, 1990; Scott, 1997).

Retrospective frequency estimations of emotional experiences are among the most common tasks on standardized psychological assessment instruments because many diagnostic criteria rely on symptom frequencies over time (Brown, Williams, Barker, & Galambos, 2007; Christianson & Safer, 1996). Items that ask for retrospective frequency estimations of emotional experiences are items that ask participants how often over a

certain number of weeks in the past they have experienced a variety of emotions or emotion-specific experiences (such as crying). Responding to these retrospective emotional appraisal items requires at least two major tasks: generating retrospective emotional frequencies, and providing responses that are in the same format as the question asked.

Unfortunately for the accuracy of those retrospective reports, memory is a dynamic and reconstructive process, not a stable recording of events (Courage & Cowan, 2009; Flavell, Miller, & Miller, 2002; Kihlstrom, Eich, Sandbrand, & Tobias, 2000; Levine & Pizarro, 2004; Levine & Safer, 2002; Sudman, Bradburn, & Schwarz, 1996). Thus, retrospective frequency items are not merely gathering accounts of past experiences, but instead are reconstructions whose accuracy is affected by what information is being requested and the length of the time period for which estimates are being requested. Research with adults shows that retrospective frequency estimates of emotional experiences tend to be inaccurate, with many researchers believing that estimates are exaggerated for short periods of time, and underreported for longer periods of time (Shimmack, 2002; Thomas & Diener, 1990; Winkielman, Knauper, & Schwarz, 1998; Wirtz, Kruger, Scollon, & Diener, 2003).

Research on children's recall of emotional experiences, while far less extensive than research with adults, also seems to show that children's retrospective reporting accuracy is questionable. Children seem to have difficulty understanding the time concepts included in such items (e.g., Breton, Bergeron, Valla, Lepine, Honde, & Gauden, 1995), produce responses that are inconsistent with adults' observations (e.g., Baxter, Thompson, Litaker, Frye, & Guinn, 2002), and provide responses that are unstable in test-retest conditions (e.g., Schwab-Stone, Fallon, Briggs, & Crowther, 1994).

Despite growing evidence that retrospective self-reports about emotional experiences can be problematic, little research exists about the alternatives to retrospective structures or the differences in information gathered by items that use retrospective structures in comparison to items that do not use retrospective structures.

Participants' responses to self-report items are also a function of the structure of the item independent of the task it requests, such as item wording and the number of response options. A large body of literature addresses the sensitivity of adults' responses to features of item structure. Less literature exists on these response issues when assessing children, but the existing literature does strongly indicate that children's response stability over time, children's item comprehension, measure factor structure, and the ease of participant responding can be related to the use of complex or vague wording (e.g., Otter, Mellenbergh, & de Glopper, 1995), the use of negative wording (e.g., Marsh, 1986), the type and number of response options available (e.g., Rebok et al., 2001; Borgers, Hox, & Sikkel, 2004), and the vagueness of response option wording (Borgers, Hox, & Sikkel, 2003). Thus, any research that examines children's retrospective self-reports is affected by confounds associated with item structure.

Given that children's cognitive development is not as advanced as adults', it is not surprising that children's retrospective reports are inaccurate—if adults, with far more advanced cognitive skills, do not provide accurate responses, children can hardly be expected to do so. For those same reasons, it is also not surprising that their responding is sensitive to item structure. Developmental explanations for children's self-reporting abilities have often been founded on Piaget's stages of cognitive development, which describe children as not having the skills modern researchers believe are necessary for adult-like self-reporting until children are approximately 12 years of age and in what he

termed the stage of formal operations (Brainerd, 1978; Scott, 1997). Most researchers seem to agree that children certainly can provide self-reports prior to that age, but that items need to accommodate children's developing cognitive skills (e.g., Holaday & Turner-Henson, 1989). Children in the concrete operational stage (7 or 8 to 11 or 12 years of age) seem to be the youngest group of children to have the cognitive skills necessary to provide retrospective self-reports in the sustained question-and-answer format used by formal self-report measures (Borgers et al., 2000). Although it seems clear that older children are more accurate reporters than younger children, all children within the concrete operational stage tend to be treated as similarly competent self-reporters in the measurement literature (Amato & Ochiltree, 1987; Borgers et al., 2000; Fallon & Schwab-Stone, 1994; Woolley, Bowen, & Bowen, 2004). The result is a gap in the current understanding of differential reporting abilities across the ages in the concrete operational stage.

This study proposes to use an existing data set to compare the information gathered by retrospective frequency self-report items to the information gathered by self-report items that do not use retrospective frequency structures, and to make these comparisons across ages in the concrete operational stage. The proposed study will use data collected from 9- to 12-year-old girls with two measures – the Beck Youth Inventory – Depression (BDI-Y; Beck, Beck, & Jolly, 2001) and the Children's Depression Inventory (CDI; Kovacs, 1992). These two measures both assess for depressive symptomatology in children, but differ quite a bit in their structure. The BDI-Y uses the same 4-point frequency response set for every single item stem (i.e., *never, sometimes, often, always*). The CDI does not use the common format of an item stem followed by a response set. Items on the CDI consist of three different statements, from which the

respondent selects one (e.g., *I am sad once in a while, I am sad many times, I am sad all the time*). Some of the items on the CDI make use of frequency response sets, and others do not (e.g., *I don't like to play outside, I sort of like to play outside, I like to play outside a lot*). Two sets of item pairs were selected from the BDI-Y and CDI for analysis in this study. One set consists of items matched across the BDI-Y and CDI for both content and the use of frequency structures. For instance, both measures include an item that addresses sadness frequency. The other set of items was matched across the measures for content, but mismatched for the use of frequency structure. For example, both instruments include an item about self-hate. On the BDI-Y, the response set consists of four frequency responses, but on the CDI the item consists of three statements about the extent to which the respondent hates him- or herself. By including both sets of item pairs, differential reporting to retrospective frequency items can be examined while controlling for the effects of item structure.

The analyses will rely mostly on comparing parameter estimates obtained from a form of item response theory (IRT) called the graded response model (GRM; Samejima, 1969). IRT has the benefits of allowing error to vary across items within a model and across latent trait levels, allowing for the simultaneous modeling of mixed-format item sets, and producing parameters that are not sample-dependent.

Literature Review

In the literature review that follows, findings from research with children and adults will be presented. More measurement research has been conducted with adults than with children, with the result that much of the research presented in this paper's literature review focuses on findings from studies with adult participants. The findings from studies conducted with adult populations will be briefly summarized to introduce

key measurement findings. Studies conducted with children will be discussed in more detail to highlight the operation of these measurement concepts in child populations. The review will focus on the areas of interest to this study—the use of retrospective frequency appraisals with children.

Generating Retrospective Emotional Frequencies – Findings from Studies with Adult Participants

When answering retrospective frequency items, respondents are not usually experiencing the emotions about which they are reporting. Instead, participants recall relevant emotional experiences and estimate their frequencies (Robinson & Clore, 2002b). Unfortunately for the validity of standardized assessment scores, memory is a dynamic, reconstructive, and somewhat error-prone process, not a stable recording of events (Courage & Cowan, 2009; Flavell et al., 2002; Kihlstrom et al., 2000; Levine & Pizarro, 2004; Levine & Safer, 2002; Sudman et al., 1996). Hypothetically, because emotions are so important in directing behavior, perception, and judgment (Levine & Pizarro, 2004), it seems reasonable to conclude that people would attend well to and therefore remember the frequencies of their emotions. On the other hand, emotional experiences fluctuate considerably over time and across situations; accurately remembering and integrating the subtleties of so many emotions may be an impossibly detailed task (Robinson & Clore, 2002b).

Not surprisingly, evidence for the accuracy of recalled emotions is mixed (e.g., Aaker, Drolet, & Griffin, 2008; Wirtz et al., 2003). Some studies find that negative emotions are remembered more intensely than positive, and some find the opposite (e.g., Parkinson, Briner, Reynolds, & Totterdell, 1995; Paz-Alonso, Larson, Castelli, Alley, & Goodman, 2009; Thomas & Diener, 1990; Wilson, Meyers, & Gilbert, 2003). Overall,

the research shows that people overestimate the frequency and intensity of positive emotional experiences more than they overestimate the negative ones (Parkinson et al., 1995; Thomas & Diener, 1990; Wilson et al, 2003). Even people who are depressed have been found to overestimate their positive affect; they simply do not overestimate their positive affect as much as non-depressed individuals (Ben-Zeev, Young, & Madsen, 2009). When events are highly emotional, people's confidence in the accuracy of their recall increases (Talarico, LaBar, & Rubin, 2004). However, evidence does not consistently support people's sense that highly emotional memories are remembered more accurately than other types of memories (Levine & Bluck, 2004; Levine & Pizarro, 2004). An exception to this generality is traumatic memories, which, when extremely painful, can be difficult to remember (Skowronski, Gibbons, Vogl, & Walker, 2004). The mixed findings about emotional recall have led some researchers to conclude that people tend to overestimate the frequency and intensity of all their emotional experiences (Thomas & Diener, 1990; Wirtz et al., 2003).

As the requested recall time frames stretch over longer time periods, people make increasingly large errors in their reporting, partly because of the methods they use to generate frequency estimates (Schimmack, 2002). Over the course of a week or so, people provide higher frequency ratings than they do for longer time frames (Winkielman et al.,1998). For these shorter time frames, people often provide estimates by counting remembered experiences. As time frames exceed two or three weeks, people find frequency estimation methods that rely on counting all pertinent experiences to be overwhelming (Levine & Pizarro, 2004; Robinson & Clore, 2002a; Robinson & Clore, 2002b; Schimmack, 2002). When asked to report on these longer time frames, participants reserve counting methods for only the most intense experiences (Schimmack,

2002). For more common experiences, people rely on semantic knowledge, which is knowledge based on scripts and beliefs rather than recalled details (Levine & Pizarro, 2004; Robinson & Clore, 2002a; Robinson & Clore, 2002b; Schimmack, 2002). The use of semantic knowledge seems to result in emotional experience underestimates (Schimmack, 2002; Winkielman et al., 1998). People explain their semantic knowledge-based responses by saying they provided the number of times such events usually happen, or thought of situations in which the event usually happens and then counted the number of times the situation occurred. They may even simply say that they chose an answer because it seemed like a good answer, but not be able to explain the strategy they used to select their answer (Levine & Pizarro, 2004; Robinson & Clore, 2002a; Robinson & Clore, 2002b; Schimmack, 2002).

When considered as a whole, the research on retrospective self-report accuracies paints a picture of such reports as being highly influenced by the subject that is being recalled and the length of time over which it is being recalled. In general, the frequency and intensity of emotions seems to be overestimated, as does the perceived accuracy of those estimates.

Children and Retrospective Frequency Appraisals

In theory, responding to retrospective frequency items should be more difficult for children because their cognitive abilities are not as developed as those of adults. The following sections provide a brief overview of children's development with respect to self-reporting before discussing research on children's ability to provide retrospective frequency appraisals. As will become clear in the reviews of measurement studies conducted with child populations, most self-report research is conducted with children who are at least 7 to 8 years of age, which is the lower bounaryd of Piaget's concrete

operational stage. This proposed study will focus on children within the concrete operational stage.

Piaget's theories of cognitive development. Although older children are usually more accurate reporters than younger children (Amato & Ochiltree, 1987; Borgers et al., 2000; Fallon & Schwab-Stone, 1994), all children within Piaget's concrete operational stage (7 or 8 to 11 or 12 years of age) tend to be treated as similarly competent self-reporters in the measurement literature (Borgers, et al., 2000; Woolley et al., 2004).

According to Piaget, children enter what he termed the *concrete operational* stage of cognitive development at approximately seven or eight years of age (Brainerd, 1978; Scott, 1997). Around eleven or twelve years of age, children enter the *formal operational stage*, where their self-reporting abilities become more like those of adults (Brainerd, 1978; Scott, 1997). Piaget theorized that prior to the concrete operational stage children do not possess the ability to take on others' points of view, a sense of time, and logical and deductive reasoning, all skills currently considered necessary for self-reporting (Scott, 1997). Modern researchers have adjusted their understanding of Piaget's theories somewhat by recognizing that development is uneven across children and may not exactly fit Piaget's age delineations. Modern researchers have also recognized the impact of children's sociocultural environments on their development and the theoretical benefits of incorporating insights from information processing theories (de Leeuw et al., 2004; La Greca, 1990; Scott, 1997). In the end, Piaget's theories are now used to ground the hypotheses that guide research efforts, but are supplemented with information processing insights and acknowledge the immense variability in children's developmental backgrounds.

General self-reporting skills and the concrete operational age. By the time children age into Piaget's concrete operational stage, they possess a number of abilities necessary for providing retrospective emotional appraisals, such as knowledge of basic emotional concepts, a psychological sense of self, and basic facility with time.

First and foremost, respondents must be familiar with the concepts described in an item. In the case of retrospective emotional appraisal items, emotions are one of the core components. Certainly by the time children are 7 or 8, they regularly use the emotion words *happy*, *sad*, *mad*, and *scared* (Flavell et al., 2002). The ability to conceptualize mixed emotions emerges somewhere around six to eight years of age, and develops throughout early adolescence (Flavell et al., 2002; Stone & Lemanek, 1990). Thus, children in the concrete operational stage are likely to understand the emotional concepts in emotional appraisal measures designed for children. Although children may be familiar with the emotional vocabulary of an item, children of all ages can have trouble identifying internal states (La Greca, 1990). Children younger than 8 or 9 years of age may have very little insight into internal states and need to rely on behavioral or somatic markers of emotions (Stone & Lemanek, 1990).

Self-reporting also requires a sense of selfhood (Flavell et al., 2002; Stone & Lemanek, 1990), especially for self-report items that request information about attitudes or estimates over a period of time. A sense of self is dependent on at least two cognitive skills: autobiographical memory and theory of mind. Autobiographical memory allows respondents to create and remember a history of their experiences. It seems to develop somewhere around two years of age, the same time that the cognitive self becomes stable and acts as an organizer for experiences (Courage & Cowan, 2009). The second major skill, theory of mind, begins to develop in preschool. Theory of mind skills allow

children to understand that people experience the world from different perspectives, and that the knowledge held by one person may be unknown to another. As they are better able to consider others' perspectives, children begin to more thoroughly distinguish themselves from others (Fivush, 2007).

Although some of the key skills necessary for a sense of self begin developing early in life, it is not until six or seven years of age that children begin to distinguish between physical and mental characteristics. At six or seven years of age, children still tend to define themselves in primarily concrete terms, such as their belongings or physical characteristics. At approximately eight years, children acquire a more global sense of self, but that selfhood is still highly sensitive to social input (Stone & Lemanek, 1990). It is not until approximately 12 years that the self becomes primarily psychological (Stone & Lemanek, 1990).

Facility with time is among the later self-reporting skills to fully develop, despite the fact that children start using time language as soon as they start talking, around 16 to 18 months of age (Fivush, 2007). By 3 years of age, children appear to have abstract, general, and temporally organized knowledge for recurring events. Children as young as 3 years are also able to provide temporally ordered accounts for events they commonly experience, such as the events that compose getting dressed or taking a bath. However, children at this age typically cannot order those events with respect to each other or other temporal landmarks (Hudson & Mayhew, 2009). At 4 to 5 years of age, children's understanding of time expands to events outside of themselves, and they learn to order the main events of the day. By 6 to 7 years, children can order seasonal events, and by 7 or 8, they can order days and months (Hudson & Mayhew, 2009). By approximately 12

years of age, children have enough facility with time that they can understand the concept of memory and their own use of memory techniques (Friedman, 2007).

In summary, children within the concrete operational stage possess a number of the skills necessary for generating retrospective self-reports on emotional states, such as insight into their emotional states, a psychological sense of self, and understanding of basic time concepts. The developmental literature does not, however, indicate that all children within this age group have the same ability to produce these frequency appraisals. This literature does seem to show that the skills children need to produce these estimates are still developing, which strongly implies that children within the concrete operational set likely vary in their ability to respond to these items.

Children and retrospective emotional frequency appraisals. Although the literature on children's retrospective self-reports is less extensive than that for adults, it does tend to show that, like adults, children do not produce entirely accurate retrospective reports. Children's ability to provide accurate retrospective self-appraisals has thus far been studied through four approaches: interviews in which children's understanding of time concepts is investigated, studies in which children's responses to items containing emotional terms are examined, interviews in which children's responses are compared to those of adults for the purposes of assessing accuracy, and test-retest studies that examine reliability.

Interview studies of time concept comprehension. Interview studies of comprehension have reported mixed findings about the extent to which children understand the sorts of time concepts used in self-report items—some report that children do understand these items, and some report that they do not. In interview studies of comprehension, children are typically asked to read self-report items aloud and then

describe their reasons for choosing a particular answer in a procedure called *cognitive process interviewing*. Depending on the type of cognitive process interview used, interviewers typically have an opportunity to ask participants a variety of probes about their item comprehension and response processes. Sometimes the time concepts examined in these interview studies include the type of frequency structures that will be examined in this study, such as how often children have had a particular experience. Sometimes these studies have examined other types of time concepts, such as how long an event lasted or when it started. Two studies that used interview methods to examine children's understanding of time concepts were identified and are reviewed here.

In the first study, two hundred forty 9- to 11-year-olds were asked to respond to structured interview questions about their mental health (Breton et al., 1995). Two psychiatrists evaluated the children's responses and ranked the children's understanding as successful, unsuccessful or doubtful. Based on those ratings, the researchers determined that the participants' understanding of items involving any sort of time concept was poor. Nine-year-olds understood about 26% of items, 10-year-olds about 24%, and 11-year-olds about 30%. Three main time categories were studied—symptom duration, period of time referred to by a question, and symptom frequency. Symptom duration referred to the length of time a symptom lasted. Specific period of time referred to by an item was the retrospective time scope of the item, usually somewhere between the most recent 6 to 12 months. Symptom frequency referred to the frequency with which a behavior or emotion was repeated. Of those three time categories, symptom frequency items were the most poorly understood with only 15% of items being understood by participants of any age. In contrast, children aged 9 to 11 understood approximately 38%

to 42% of all the questions, whether or not they included time concepts (Breton et al., 1995).

In the second study, researchers came to the opposite conclusion, finding that children did understand at least one time concept—period of time referred to by a question (Rebok et al, 2001). In this study, interviewers asked 5- to11-year-old children a variety of questions about a self-report measure of health. Some of the items in that measure asked participants to recall symptom experiences over "the past 4 weeks." According to the researchers, about 80% of children were able to report the "last time" a symptom or behavior had occurred in the past 4 weeks (Rebok et al.,2001, p. 68). Because children could produce an estimate, researchers concluded that children understood the item. No attempts to verify the accuracy of the children's reports were reported.

In summary, the two interview studies examining children's understanding of time concepts arrived at two different conclusions; one study found that children did not understand time concepts well, and the other study did. The second study, however, concluded that because children were able to produce reports, they understood time concepts; the researchers did not report efforts to corroborate this seeming evidence of comprehension through validity checks or through more detailed interviewing techniques. Neither study examined whether understanding varied across age or could be improved if the same questions were asked without the use of time concepts.

Studies examining items containing emotional content. Three studies examining children's responses to items measuring emotional content were found. The first two studies relied on statistical analyses and the third relied primarily on interviews.

In the first study, Cremeens, Eiser, and Blades (2007) conducted a study of possible response options for a quality of life measure designed for children. The measure consisted of 30 items divided across three scales: a 12-item ability scale (e.g., running, climbing, writing), a 12-item social scale (e.g., relationships with family and friends) and a 6-item mood scale (e.g., about feelings and general mood). The authors read the measure to 266 healthy children between the ages of 5 and 9 years who were recruited from schools. The children provided verbal responses to the items and were randomly assigned to one of three response option conditions: a 4-point bipolar (i.e., option extremes labeled with opposite wording such as *happy…sad*) set consisting of circles, a 4-point bipolar set consisting of faces, or a 4-point unipolar (i.e., options for the presence or absence of one emotion) thermometer-shaped set. Children were asked to respond to the items with reference to the prior week. Internal reliability estimates calculated using Crohnbach's alphas were lowest for the mood items for all three response option versions. Estimates ranged from .20 to .33 for the mood items, .45 to .70 for the social items, and .56 to .72 for the ability items. When participants completed the measure again one week later, they either used the same response option version they used originally or they used a different version. For all response option conditions, interclass correlation coefficients calculated for the test-retest responses were lowest for the mood items: .26 to .49 compared to the social items' range of .50 to .76 and the ability items' range of .44 to .70. The authors concluded that the mood items showed the lowest levels of internal and test-retest reliability, although they cautioned that the limited number of items in the scale could be causing the low estimates.

In the second study, Chambers and Johnston (2002) studied 60 children's responses to 18 items: 6 physical task items, 6 social objective task items, and 6

subjective task items. The physical task items asked children how they would compare a child completing a physical task (e.g., carrying library books) to other children completing different degrees of a similar task (e.g., carrying more or fewer library books). The social objective task asked children to compare the feelings of a child in a certain social situation (e.g., with a certain amount of candy) to children in similar settings of different degrees (e.g., with more or less candy). The subjective task was much like the social objective task, except that it asked children how they themselves would feel if they were in the position of the child in the scenario. The 60 children were between the ages 5 to 12 years, with 20 each in a 5- to 6-year-old group, a 7- to 9-year-old group, and a 10- to 12-year-old group. Thirty of the children (10 from each age group) were asked to use a 3-point response option set; the other 30 were asked to use a 5-point response option set. The extreme options for both response sets were "not at all" and "a lot." A three-way ANOVA (2 response option sets, 3 age groups, 3 task types) found significant effects for age group, task type ($p < .01$ for both effects), and the interaction between age group and task type ($p < .05$). Post hoc tests revealed that on the social objective task, the 5- to 6-year-old group provided extreme scores significantly more often than the other two age groups, and that the 7- to 9-year-olds provided extreme scores significantly more often than the 10- to 12-year-olds. On the subjective task, significant differences were found between the 5- to 6-year-old group and the remaining groups; no other significant effects were found. The post hoc tests did not find any differences between the age groups on the ability task. In the absence of significant findings for the number of response options, the authors concluded that the younger children's more frequent provision of extreme scores meant that younger children have

more difficulty with the emotional and perspective-taking tasks on the social objective and subjective tasks.

The interview study was conducted by Jacobson and colleagues (2012), who examined children's understanding of self-report items on a pediatric pain measure. The authors interviewed 34 children between the ages of 8 and 18 years (mean age of 13.8 years). The children all had a chronic pain condition and were recruited from various treatment centers. The interviews lasted between 30 and 60 minutes, and explored up to four domains on the measure, which included: anger, anxiety, depressive symptoms, fatigue, pain interference, peer relationships, and physical functioning. The authors analyzed the interview data using qualitative methods. They concluded that children between the ages of 8 and 12 years usually did not understand more abstract mood labels like *anger, anxiety*, or *depression*. The children in this age group instead reported feeling *mad, sad, excited*, or *upset*. The children usually reported these feelings in response to specific, short-term triggers like being teased and made reference to somatic symptoms (e.g., crying) when describing the feelings. In contrast, children between the ages of 13 and 18 years were better able to differentiate between anger, anxiety, and depression. These children were able to describe persistent forms of these moods and used more cognitively complex descriptors when describing their experiences of the moods (e.g., metaphors, reports about thought patterns and attentional capacity).

Considered together, this group of studies shows that children younger than 12 years-old do not seem to understand or report stably on some items that measure emotional content. However, only one of these studies examined differences between the reporting abilities of children younger than 12 years. The other two studies either did not examine differences across age groups or compared children younger than 12-years-old

to children 12 years of age or older. As such, the research base for differences in emotional self-reporting abilities in children 12-years-old and younger is limited.

Interview studies of accuracy. Interview studies that examine the accuracy of children's retrospective reports have used researcher, parent, and teacher observations as the evidence against which children's reports were compared. These studies have tended to show that children's estimated frequency rates differ quite a bit from the observed frequencies generated by the adults. Three such studies are detailed below.

The first study made use of data observable to the researchers and showed that fourth graders (who are about 9 years old) seem to have poor recall of their prior day's lunch menu. The 104 children were interviewed about 24 hours after they ate lunch. The lunches they ate were observed by the researchers. Children omitted about 51% of consumed items from their reports, and 39% of the items they reported eating were not actually consumed (Baxter, et al., 2002).

The next two reports were part of one research effort evaluating the quality of items on the National Assessment of Educational Progress (NAEP; Levine & Huberman, 2002; Levine, Huberman, & Buckman, 2002). The NAEP is a set of questions students answer when they take achievement tests. The questions are designed to measure "background information" (Levine & Huberman, 2002, p.1) and topics believed to be related to children's performance on these measures. The questions are available for public viewing on the National Center for Education Statistics' website. These questions vary widely in their scope, and include items about demographics, education and recreational activities at home, and educational activities at school. Time frames for the questions also vary widely. Some ask for frequency estimations over a day, e.g., *On a school day, about how many hours do you usually watch TV or videotapes outside of*

school? None, one hour or less, 2 or 3 hours, 4 or 5 hours, 6 hours or more. Others ask for estimates over a month, e.g., *How many days were you absent from school in the last month? None, 1 or 2 days, 3 or 4 days, 5 to 10 days, More than 10 days.* Still others ask for estimates over a range of time periods, e.g., *For this school year, how often do you solve mathematics problems with a partner or in small groups? Never or hardly ever, Once or twice a month, once or twice a week, Almost every day.* As with the three examples provided here, most of the questions provide 4 or 5 response options.

In the research detailed by one of the reports, when fourth and eighth graders (e.g., children who are typically between 9 and 13 years old) were asked how often they engaged in behaviors their parents would be able to observe at home (such as how often children read at home), their frequency estimates were different from those provided by their parents for approximately 53% of fourth-graders' responses and 33% of eighth graders' responses (Levine & Huberman, 2002). Time frame estimations (i.e., items that asked how long was spent in a particular activity) were almost as problematic, with parent-child discrepancies at roughly 42% for both fourth- and eighth-graders (Levine & Huberman, 2002). The precise details about the differences in parent and child reports were not provided; the authors noted only that any disagreement between parent and child report constituted a discrepancy.

Data from the other report on the NAEP study showed that teacher-student discrepancy rates were worse than parent-child discrepancy rates. Fourth-grade teacher-student discrepancy rates for behavioral frequency items that made use of a 4-point response scale were as high as 62%, and rates for eighth graders as high as 49% (Levine, Huberman, & Buckner, 2002). Through follow-up interview procedures, both fourth and eighth grade students were found to often misinterpret the time frames referred to in

items, for instance applying time frames from previous questions to the current question, or interpreting items literally (e.g., thinking that science lessons did not occur "almost every day" because they did not also occur on weekends; Levine, Huberman, & Buckner, 2002).

Supporting the quantitative evidence that children do not perform well on retrospective estimation tasks, when the fourth and eighth grade students in the NAEP study were asked in interviews to explain how they generated their frequency estimates, they often said that they simply guessed (Levine & Huberman, 2002). The researchers suggested that children tend to use guessing strategies for retrospective estimates because children lack the cognitive skills and the motivation necessary to conduct the calculations such items accurately require (Levine & Huberman, 2002).

Combined, these three studies demonstrate that children's retrospective reports of event frequency often do not concur with the reports provided by adult observers. If the reports produced by adults can be thought of as more accurate, then the adult-child discrepancy rates provide evidence that children do not produce accurate retrospective frequency accounts. All these studies focused on behavioral frequencies. No studies of emotional frequency accuracy with children could be found. These studies were also very limited in the sample ages, including only fourth- and eighth-graders.

Test-retest studies of reliability. The most common study format involving children's retrospective estimates reviewed is the test-retest format. Studies making use of the test-retest design, like other studies reviewed thus far, call into question the accuracy of children's reports. In these studies, the evidence comes in the form of poor test-retest reliability. The three studies detailed below studied test-retest reliability of children's retrospective reports using the Diagnostic Interview for Children-Revised

(DISC-R), a highly structured interview designed to provide diagnostic information about a variety of children's mental health issues. Questions on the DISC-R typically begin with a stem question about the presence of a symptom. Respondents are usually able to answer only *yes* or *no*. If the symptom is endorsed, participants are then asked a series of *yes/no* "contingent" questions designed to clarify whether the symptom meets diagnostic frequency, duration, and intensity criteria (Fisher, Lucas, Lucas, Sarsfield, & Shaffer, 2006; Shaffer, Fisher, Lucas, Dulcan, & Schwab-Stone, 2000).

In the first study, 6- to 11-year-olds were administered the DISC-R, and then re-administered the interview between 7 and 18 days later. The resultant kappa coefficients were poor (i.e., <.40) for items that required considerations of symptom duration or onset. Kappa coefficients for items that did not require the use of such time concepts tended to be slightly higher and fair (i.e., .40-.58; Schwab-Stone et al., 1994).

In a separate test-retest study with the DISC-R, one hundred fourteen 6- to 12-year-old children produced low kappa coefficients for items that make use of a variety of time concepts, not just frequency estimates. Polytomous response option questions that asked for frequency estimates had kappa coefficients of .08 (4 items), duration items had kappa coefficients of .06 (10 items), and date of onset had kappa coefficients of .27 (14 items), all of which are low values (Fallon & Schwab-Stone, 1994). In contrast, items that did not make use of time concepts had kappa coefficients of about .33 (227 items). These results must be interpreted with some caution because so few items made use of time concepts. This study was one of the few to examine age differences in performance. The researchers found that there was no interaction between age and question attribute, which the authors defined to include content (i.e., behavior, emotion, or thought), whether comparison to other children was requested (e.g., *compared to other children, how often*

do you…), sentence length, and the presence of time concepts. Instead, kappa coefficients were found to increase linearly with age across all question attributes studied (Fallon & Schwab-Stone, 1994).

The third study again demonstrated poor performance of items that contain time concepts (Lucas et al., 1999). In this third study, neither the ages nor the grade levels of the 247 child participants were reported. After the initial administration, the researchers administered the DISC-R again after approximately two weeks. When items were endorsed at baseline but not endorsed at follow-up, they were said to have *attenuated*. A series of t-tests was conducted to compare the proportion of items with a particular item characteristic that was attenuated at follow-up to the proportion of items without that characteristic that was attenuated. Item characteristics studied included several timing elements. The *t*-tests showed that items containing a time element were more attenuated at follow-up than items that did not contain a timing element ($p < .01$). Five specific forms of time concepts were explored: frame (whether the symptom occurred within a specific period), frequency (how often the symptom occurred), timing (when the symptom started), duration (how long the symptom lasted), and intensity (how intense the symptom was during a particular period). Of those elements, items containing frequency elements showed the greatest degree of attenuation, significant at the $p < .001$ level. Timing, duration, and intensity were also attenuated at significance levels of at least $p < .05$. Frame was the only time concept that did not show significant levels of attenuation (Lucas et al., 1999).

Taken together, the test-retest studies provide evidence that children's retest reports on the DISC-R tend to be quite different from their initial reports. Items that make use of time concepts usually have quite poor test-retest statistics, with retrospective

frequency items performing the most poorly of all the time concepts studied. Of these three studies, only one examined age differences. All of them studied the effects of time concepts on test-retest reliability, but none of them attempted to compare the reliability of items with time concepts to the reliability of items without time concepts while controlling for potential confounds, such as differences in content or other differences in item structure (such as the number of response options or the amount of item wording).

Summary. With one exception, the literature on the accuracy of children's self-reports indicates that children between the ages of 9 and 12 years—the age range being examined in this study—appear to supply inaccurate or unreliable responses to self-report items that make use of a time or emotional concept, particularly retrospective frequency items. These data support implications from the developmental literature that children in the concrete operational stage will have difficulty generating retrospective emotional frequency appraisals.

Missing from these studies is a thorough examination of age differences in reporting abilities. These studies also do not compare responses to frequency items in a manner that controls for potential confounds, such as differences in item content or differences in other features of item structure (such as the number of response options or the amount of wording).

In the next set of sections, the second major topic of this literature review will be presented – the relation between item structure and participants' responding. Findings from studies conducted with adults will be presented first followed by those conducted with children.

Generating Self-Reports that are Consistent with Item Structures – Findings from Studies with Adult Participants

Answering self-report questions consists of multiple cognitive tasks, not just the process of completing the task requested in an item, such as a retrospective frequency appraisal. A variety of theories have been developed to explain the cognitive tasks necessary for generating responses (Courage & Cowan, 2009; Hastie, 1987; Schwarz, 1997, 2007; Sudman et al., 1996; Tourangeau & Rasinski 1988; Tourangeau, Rips, & Rasinksi, 2000; Wanke & Schwarz, 1997). These theories vary somewhat in their details, but can be briefly summarized as follows:

1. Interpret the item (both the question and the response options)
2. Retrieve relevant beliefs, feelings, and facts
3. Use the information from step 2 to render an appropriate judgment
4. Combine the judgment in step 3 with concerns about the response's impact to select a response

Interpreting items involves the comprehension of the language in the item and the task being requested by an item. For the purposes of this study, retrieving relevant beliefs involves drawing upon the memories necessary to generate retrospective emotional frequency appraisals. Rendering a judgment is the process that includes identifying response options that are consistent with the memories recalled. Selecting a response encompasses the actual choosing of a response from among those available, a process which can be affected by participants' concerns about the effect of providing that particular response.

The execution of all these tasks is a linguistic process—reading, interpreting, recalling, and analyzing memories requires the use of language. Each person who reads and responds to a particular item is a unique individual with a series of unique experiences which are described and categorized by each person into nodes on an

idiosyncratic informational web (Hastie, 1987). Since each person's informational web is unique, the interconnections among the nodes on those webs are unique as well. The result is that people can read the same self-report item but interpret it in different ways, call upon different types of memories, and find different types of response options to be appropriate for the judgment at hand.

Many facets of a self-report measure can be adjusted in an attempt to reduce idiosyncratic responding and promote response accuracy. For instance, the way that items are written, are ordered in relation to each other, and the setting in which they are administered can all be controlled by writers of self-report measures. This particular study will focus on aspects of items' structures, such as wording, number of response options, and type of response options. It will not address features of item sets (such as item order) or administration.

Item wording. Shortcomings in item construction sometimes produce just a simple misunderstanding of the item's task: a participant may misread a word or response item, recall information about the wrong time period, or retrieve information about the wrong opinion. Dillman (2000, p. 53) provides an example of a question tested for a U.S. Census question that proved problematic:

> How many people were living or staying at this residence on Saturday, March 3rd, 2000? To make sure each person in the United States is counted only once, it is very important to:
>
> **Include** everyone who live here whether related to you or not, and anyone staying temporarily who has no permanent place to live;

> **But do not include** anyone away at college, away in the Armed Forces, in a nursing home, hospice, mental hospital, correctional facility, or other institution.

When participants were interviewed about their responses to the question, some respondents reported that the question was asking participants to estimate the total number of people in the United States. Comprehension problems are more likely when item wording is vague (e.g., How many times have you been on vacation?), overly complicated (e.g., Notwithstanding any interest in evolutionary bioethics, what is your considered opinion on the foreseeable utility of stem cell research?), technically inaccurate (e.g., How often do you calculate Box's *M* test during an ANOVA?), or double-barreled (e.g., To what extent do you think that your department chair is a caring and effective leader?; Dillman, 2000).

In addition to understanding the literal meaning of an item, participants must also infer what the item authors intended an item to emphasize (Schwarz, 2007). Consider the question "How are you today?" If asked by a friend, this question is likely to evoke very different responses than those evoked if the question is asked by a nurse in an emergency room. To sort out connoted meanings, participants refer to facets of the items themselves, to earlier items, and to the context in which the items are being asked (Krosnick, 1991; Schwarz, 2007; Tourangeau et al., 2000). One item feature to which respondents often refer is the available set of response options (Schwarz, 1997). Consider, for instance, the stem, "How much do you like eating vegetables?" Now consider the response set, "I never eat vegetables," "I sometimes eat vegetables," and "I eat vegetables with almost every meal." This response set implies that the frequency with which people eat vegetables indicates how much people like vegetables. As respondents consider

subsequent items, they may now be more likely to use behavioral frequency as an indicator of whether or not they like something.

Response option types. With the tasks of item understanding and information recall roughly complete, participants must next transform the recalled information into an answer that fits the response options available. Several types of response options are frequently seen on psychological assessments. Serial lists, Thurstone scales, frequency scales, and Likert-types scales are among the more common (DeVellis, 2003). Each of these response types has benefits and drawbacks. This study will focus on Likert-type and frequency scales.

Likert-type scales are often used to capture degrees in endorsement (Dillman, 2000). Likert-type scales consist of a range of response options that move between two extremes. The two extremes of the scale can reflect varying degrees of the same valence (a *uniploar* scale) or cover two separate valences (a *bipolar* scale; Krosnick & Fabrigar, 1997). The points on a Likert-type scale are typically labeled, and respondents are usually directed to select one of the available options. An example of a unipolar Likert-type scale is:

How do you feel about luxury cruises (*circle one option*)?

1	2	3	4	5
I can barely tolerate them	They're not too bad	They're pretty fun	I like them a whole lot	They're the absolute best!

An example of a bipolar scale is:

How do you feel about luxury cruises (*circle one option*)?

1	2	3	4	5
I hate them	I don't like them	I'm completely indifferent	I like them	I love them

In the example of the unipolar scale, respondents were provided with a scale that described varying degrees of liking something. In the bipolar scale example, the participant was provided with response options that covered both disliking and liking.

The bipolar scale example also contained another common feature of Likert-type scales: a *neutral midpoint*. Neutral midpoints are supposed to reflect a naturally existing fulcrum between two opposing ends of a continuum (e.g., like and dislike) (Ayidiya & McClendon, 1990; Krosnick & Fabrigar, 1997; Raajimakers, van Hoof, 't Hart, Verbogt, & Vollebergh, 2000). However, seemingly opposing opinions are not necessarily opposite ends of a single continuum. Factor studies consistently find that positively and negatively worded items form different scales, even when the negatively worded items are simply negative variations of the positive items (for instance, "I like pie," and "I don't like pie;" Weems, Onwuegbuzie, & Collins, 2006). Since bipolar scales may not really reflect two opposing ends of one construct, neutral midpoints may not really exist. Instead, neutral midpoints may be a way of people saying that they do not have an opinion about something one way or the other. If people have no opinion to report, what is called neutrality may actually be missing data.

Frequency scales are similar to Likert-type scales, the prime difference being that frequency scales measure frequency rather than opinion or valence. Continuing with the luxury cruise example, an example of a frequency scale question is:

How often in a typical year do you go on a luxury cruise (*circle one option*)?

1	2	3	4	5
Once a year or less	Three times a year	Four times a year	Once every other month	Once a month

Frequency scales often use language that is more vague than counts within a certain time frame, especially when the information being requested is difficult to recall. Events that happen frequently or to which people do not attend are difficult to count because people do not notice them enough to remember them (Kihlstrom et al., 2000). Seatbelt usage is a good way to demonstrate the difficulties of counting unnoticed and common events. Which question would be easier to answer – a version that asks for a count?

How many times a month do you usually buckle your seatbelt (*circle one option*)?

1	2	3	4	5
0-20 times	21-40 times	41-60 times	61-80 times	81 or more times

Or a version that asks for a vague frequency?

How often during a month do you usually buckle your seatbelt (*circle one option*)?

1	2	3	4	5
Never	Rarely	Sometimes	Often	Always

Vague quantifiers have the benefit of making questions easier for respondents to answer. The drawback is that vague quantifiers are somewhat relative terms, which means that they may function differently across subject matters and across people. When

people say that they always brush their teeth, for instance, they are probably saying that they brush their teeth twice a day. On the other hand, when people say that they always keep their hair trimmed, they probably mean that they get their hair cut once every month or two. These variances in response option interpretation can become more extreme as the subject being investigated varies widely across respondents. For instance, children might be asked how often they study by using the response options *never*, *sometimes*, *often*, and *always*. Children might be prone to reporting *always*, although it is likely that some children spend very little time completing homework while others might spend several hours a day on it. In the end, researchers making use of instruments with vague quantifiers can never be sure just what people mean when they use those terms and whether term usage is consistent across respondents.

Response option wording. Participants often find Likert-type and frequency scales easier to use when there are just the right number of response options and response options are clearly labeled (Krosnick & Fabrigar, 1997; Weems, 2004). Labels are more helpful when the number of scale options and the detail in the labels is not too great (Krosnick & Fabrigar, 1997). Unfortunately, nothing more precise than "clearly labeled" and "the detail is not too great" can be offered in the way of labeling guidelines—no precise formula exists for determining just how much verbiage is appropriate in a scale's response options. The proper amount of verbiage is unique to each measure, and is a function of the item's task and the number of response options. Fortunately, more concrete guidelines exist for the optimal number of response options for use with adults. Five to seven options seems to be the most reliable number of options for unipolar scales; bipolar scales seem most reliable with seven options (Krosnick & Fabrigar, 1997; Weems, 2004)

Children and Item Structure

Like those of adults, the answers children provide to self-report questions are affected by item structure, which includes features of the item stem such as wording, negative wording, and length; and by response option structure, which includes features such as the type of response options used, the number of response options, the use of neutral midpoints, and response option wording. The sections below include reviews of research on each of these topics.

Item stem wording. Children's comprehension and use of language requires that items make use of simple, concrete language. Language comprehension difficulties are more pronounced in younger children. The discussion below presents studies that highlight some of the language comprehension issues that may be present in the measurement of children, and also discuss several studies of changes in measurement comprehension across different age groups.

Holaday and Turner-Henson (1989) were among the first authors to discuss the importance of using age-appropriate language when assessing children. Based on a review of developmental literature and their own experiences of assessing children, Holaday and Turner-Henson noted that children can interpret items quite differently than the authoring adults intended. Concrete interpretation was offered as among the more common forms of item misinterpretation. Holaday and Turner-Henson provided an example on concrete interpretation in which "school-aged" (p. 249) children were asked whether they had been on a school field trip that year. A number of the children had answered "no" because they had been on a class field trip, not a school field trip.

In a review of methods for interviewing children, Scott (1997) reported results from some of her own interviews that were conducted as part of the British Household

Panel Study (BHPS). The BHPS includes an annual survey of 5000 households selected to be representative of the national population. Children were involved in the BHPS for the first time in 1994. The age of child participants ranged from 11 to 15 years of age. Individual interviews conducted in children's homes were used to pilot questions for subsequent group interviews. The pilot interview consisted of 83 structured items with an open response format, and lasted about 30 minutes. These interviews revealed that children had difficulty with depersonalized language. For instance, when asked to answer questions about "people my age" (p. 346), children would sometimes try to answer the question by guessing the age of the interviewer asking the question. Neither the number of children who participated in these pilot interviews nor the number who displayed this type of comprehension difficulty was reported.

The three studies described below demonstrate that older children can have fewer difficulties understanding item stems than younger children. Otter, Mellenbergh, and de Glopper (1995) asked 635 nine-year-olds and 570 fourteen-year-olds to participate in a test-retest study of a reading literacy survey. After the initial administration, surveys were administered a second time at five to eight week intervals. Each age group was administered a subset of the 207 items that comprised the original survey. The 9-year-olds were administered 81 items and the 14-year-olds were administered 122 items. Researchers rated the items on their degree of clarity and complexity using a dichotomous scale: items were rated as either being clear or ambiguous and as being either simple or complex. Items were rated as ambiguous if they contained vague frequency options or could be interpreted in multiple ways. Items were rated as complex if the information requested had to be calculated, as in an estimate of frequency.

Four coefficients were generated for each item to assess response stability between the two administration times: correlations (polychoric and tetrachoric), Kendall's tau-b, the contingency coefficient (which is an adaptation of a phi coefficient that allows for more than 2 x 2 comparisons), and the proportion of diagonal elements in the contingency table. The number and type of response options were not provided, but appear to have ranged between dichotomous and polytomous because the four coefficients used to assess response stability between the two administration times are designed for use with dichotomous and polytomous ordinal and nominal data. To evaluate the relation between item characteristics and response stability, the four coefficients for each item were simultaneously regressed on the two independent variables of clarity and complexity using multivariate analysis of variance (MANOVA). A MANOVA was conducted for each age group. No significant interactions between clarity and complexity were found for either age group, but clarity and complexity were significant as main effects in both groups. Not all coefficients were significantly related to clarity and complexity in each age group. Cohen's *d* effect sizes for significant effects of clarity on the coefficients ranged from 1.01 to 1.79 in the 9-year-old group and from 0.28 to 0.51 in the 14 year-old group. For complexity, Cohen's *d* effect sizes for significant relations ranged from 1.08 to 2.75 in the 9-year-old group and from 0.44 to 3.75 in the 14 year-old group. The authors concluded that although clarity and complexity in stem wording can both reduce retest stability, the effects of clarity are smaller in 14-year-olds.

In the second study to be presented here, Borgers and Hox (2001) examined the relation between a variety of item structure features and item non-response. The data were a combined set of data from 5 educational research studies that utilized a total of 37

scales (348 items) and included responses from a total of 3492 children between the ages of 8 and 18. Items were coded on 26 different dimensions, one of which included item stem ambiguity. The researchers stated that categorization of items as ambiguous was in this study a somewhat subjective rating. Data were treated as being hierarchically ordered, with items being nested within respondents. Each of the 26 item dimensions was entered one at a time into a logistic regression where the presence or absence of an item response was treated as the dichotomous dependent variable. Item dimensions that were significant in the individual analyses were then simultaneously entered into a hierarchical regression. In the course of their analyses, the authors found that participants' years of education interacted significantly with item ambiguity ($p < .01$). Ambiguous items were found to be associated with higher rates of non-response, with increased years of education decreasing rates of non-response.

In the third study, Rebok et al. (2001) interviewed 60 children ranging in age from 5 to 11 years using 32 items from a health survey. Each item was presented to children twice using different stem wording. In addition to being asked how they chose their answers, children were probed about their understanding of key terms in the items (e.g., healthy, worried, on a dare). Comprehension was evaluated using a 3-point coding scheme and two judges. Judges gave 1 point to interview responses that demonstrated poor or no understanding, 2 points were given to responses that showed some understanding, and 3 were given to responses that showed clear understanding. Inter-rater reliability in coding was 78%. Age was found to be inversely related to stem comprehension ($r = -.70$). Eight- to 11-year-olds showed poor understanding of about 3.5% of stems, while 5-year-olds showed poor understanding of about 50% of stems.

As the above studies demonstrate, item comprehension appears to be greater in older children and to be improved by the use of clear and simple item wording. However, the participant ages in these studies were not focused on the concrete operational stage. Rather, large age ranges were examined (e.g., 8 to 18 years of age), children in the concrete operational stage were compared to children from other age groups (e.g., 9-year-olds compared to 14-year-olds), or the concrete operational age group was treated as a single age group. The next selection of studies will include discussions of another important aspect of item wording – that of negative wording.

Negative wording. Similar to findings with adult populations, research with children demonstrates that measures that include negatively-worded items are associated with a different latent factor structure than measures that include only positively-worded items. Response stability over time seems not to be affected. Below are descriptions of three studies that investigate the effects of negative wording on participants' response selection.

In the earliest of the three studies, Benson and Hocevar (1985) administered a survey measuring attitudes toward school integration to 522 fourth, fifth, and sixth graders, grades which cover age ranges of approximately 9 to 11 years of age. The survey consisted of 15 items, each with 5 response options. Each child was administered one of three survey versions. Items on one version were all positively worded, items on another version were all negatively worded, and items on the third version included 8 positively worded items and 7 negatively worded items. The negative items were reworded versions of the positive items, and they were reverse coded for data analysis. Several factor analyses were performed on the three versions of the survey. In the first, the measures that included only positively and negatively worded items were entered into a

confirmatory factor analysis (CFA). All item parameters were allowed to be free in the first model, and in the second the loadings between the positive and negative versions of each item were constrained to be equal. The difference in fit between the models was significant ($p < .01$), with the constrained model showing poorer fit. The researchers also entered the measure which included both positively and negatively worded items into a CFA model, this time loading all the positively worded items onto one factor and the negatively worded items onto another. All but one item loaded significantly onto its hypothesized factor.

Soon after the Benson and Hocevar (1985) study, Marsh (1986) also published research about the factor structures underlying positively- and negatively-worded items. This study produced slightly different results. Instead of demonstrating that the negatively-worded items formed their own independent factor, this study demonstrated that negatively-worded items load onto both the content factor they are intended to measure and a separate factor associated only with negatively-worded items. Marsh administered a self-concept measure to 658 second to eighth graders, a grade range which typically covers ages 7 to 13 years. The measure consisted of 76 items each with a 5 point response option set. The measure was designed to assess 8 scales. Each scale included one or two negatively-worded items and eight positively-worded items. Marsh conducted a set of four CFAs. In the first, the model included eight factors to represent the eight measure scales. Both positive and negative items were loaded onto the factor associated with the scale they were intended to measure. In the second model, a ninth factor was included, and negatively-worded items were only allowed to load onto the ninth factor. In the third model, the negative items were allowed to load onto both the factors associated with the measure scales and onto the ninth factor. Of those models, the third (in which

negative items were set load onto both their associated scale factors and a ninth factor) fit the best, indicating that negatively worded items are associated with different factor structures than positively worded items.

Marsh (1986) also conducted a separate set of analyses using corrected item-total correlations on the same data that examined the relation of negatively-worded items with the other items in the measure scales. Cronbach's alpha coefficients did improve somewhat when negatively-worded items were removed from the scales. Differences when the negatively-worded items were removed were about .04 on average. Correlations demonstrated that positively- and negatively- worded items were poorly related in the younger age ranges, with the correlation improving as age increased. For second graders, the correlation between positive and negative items was r = -.02, and by fifth grade was r = .59. Marsh concluded that this analysis demonstrated the cognitively demanding nature of negatively-worded items, reasoning that as cognitive capacities increase, the ability to provide consistent answers across different item constructions also increases.

The third study also examined corrected item-total correlations, and was the only one of the three to examine response stability over time. Borgers, Hox, and Sikkel (2004) utilized a telepanel survey of 2000 households to contact potential participants for a test-retest study. A total of 222 children between the ages of 8 and 16 years participated in the first wave of the study. Of those 222 initial participants, 91 participated in the retest wave of the study, which occurred between 3 and 8 weeks after the initial wave. The items were extracted from larger existing measures and included items about self-esteem and well-being. The researchers altered the wording of the original items so that half of the participants received measures with negatively worded items. The authors found no effect

of negative item wording on corrected item-total correlations or on response stability over time.

The above studies suggest that negatively-worded items form different latent factor structures than items that are positively worded, but are as stable over time as those obtained through positively-worded items. Whether negatively-worded items are well correlated with positively-worded items measuring the same concepts is unclear; one study found statistically significant poor corrected item-total correlations and another found no significant effects. The study that found significant effects on corrected item-total correlations noted that correlations were much higher in older participants. That study was the only one to examine differences between age groups. In the next and final section on item stem structure, item length will be discussed.

Item length. With the exception of two studies, most research on item length shows no significant effects of the length of the item stem on the response obtained. The two publications that claim support for the effects of item length reach different conclusions. In the first publication, Holaday and Turner-Henson (1989) state that, based on their experiences with interviewing children and theoretical information processing literature, longer questions can provide children with more memory cues and give them more time to produce accurate responses. In the second study, Breton et al. (1995) compared comprehension rates of items on the DISC-R between 9-, 10-, and 11-year-old participants. Two hundred forty children each took one quarter of the DISC-R so that a total of 60 DISC-Rs were completed. The participants were asked to provide answers and then were probed about their response process and understanding of the items. Two independent judges rated whether or not the children understood the items. Items with 1 to 9 words were rated as being understood significantly more often than items with 10 to

19 words, which themselves were found to be understood significantly more often than items with 20 or more words ($p < .01$ for all comparisons).

The remainder of the studies reported no effect of item length on item non-response or response stability over time. The research by Borgers and Hox (2001), discussed in the section on item stem wording, was one of four studies to find no effect. Unlike the other studies, which relied on test-retest methods, Borgers and Hox used data aggregated from 5 previously conducted research efforts. The data included a total of 37 scales (348 items) and included responses from a total of 3492 children between the ages of 8 and 18. Items were coded on 26 different dimensions, one of which was the number of words in the item stem, and another of which was the number of sentences in the item stem. Item non-response was regressed on the 26 item dimensions. Significant item dimensions were then entered into a multilevel logistic regression that treated items as nested within respondents. No effect of either number of words or number of sentences was found.

In the first of several test-retest studies to be discussed in this section, Fallon and Schwab-Stone (1994) also found no effect of item length on response stability. The DISC-R was administered to 114 children between the ages of 6 -to 12-years-old at one to three week intervals. Because the DISC-R makes use of a polytomous response scale, the authors were able to generate kappa coefficients for each item. These kappa coefficients were regressed on a number of item features, including sentence length. Sentence length was not found to have a significant effect on kappa coefficients.

The next test-retest study also used the DISC-R. Lucas et al. (1999) administered the measure to 247 parent-child pairs and administered again 2 weeks later. When items were endorsed at baseline but not endorsed at follow-up, they were said to have

attenuated. A *t*-test was conducted to examine whether items that were longer than or equal to 15 words were more likely to have attenuated than items that were less than 15 words. No differences in attenuation rates between the two groups of item length were found.

In the final test-retest study to be discussed, the Diagnostic Interview for Children and Adolescents – Revised (DICA-R) was administered to one hundred nine 7- to 17-year-olds twice at approximately 11 day retest intervals (Perez, Ascaso, Massons, & de la Ossa Chaparro, 1998). Much like the DISC-R, many items on the DICA-R make use of polytomous response sets. Kappa coefficients were generated for each item, and then regressed on a number of question characteristics including question length. No significant effects of question length were found.

The above studies demonstrated that the published literature on item length provides differing findings. One publication theorized that longer items should increase comprehension, while another found that comprehension decreased as item length increased. The remaining studies used test-retest formats and found that length had no effect on response stability over time.

Summary of item stem research. In summary, the studies presented above demonstrated that various facets of item stem structure such as the use of complex or negative wording can affect item comprehension, latent factor structures, and response stability over time. While the studies presented above included participants in the concrete operational age range, differences across age groups were rarely studied. The next several sections will discuss the effects of response option structure and will focus on type of response options, number of response options, the presence of a neutral midpoint, and response option wording.

Response option types. As discussed in the section on adult responding, multiple types of response options exist. Three types used with children are horizontal visual analog scales (VAS), polytomous vertical response sets, and polytomous horizontal sets. The structure of VASs varies, but usually involves a continuous line on which participants mark their answers. Sometimes only the extremes of the scale are labeled and sometimes options between the two extremes are labeled. Polytomous vertical response sets present multiple response options listed in a vertical line. Polytomous horizontal response sets typically present a horizontal line of numbers or pictures. Three studies discussed below compared participants' performance using these different response option structures.

Rebok et al. (2001) conducted several studies exploring the use of different item response structures. When 35 children between the ages of 5 and 11 years were asked whether they preferred a horizontal or a vertical VAS (after using both formats), all children reported preferring the horizontal VAS. Responses between the horizontal and the vertical scales were reported as being consistent with each other, although rates of consistency were not reported. In another of their studies, Rebok et al. administered 25 self-report items about health to 19 children between the ages of 5 and 11 years. Each participant was asked to respond to two variations of the items that differed in the response structures offered. Six different response options were tested. The first five made use of pictures demonstrating both extremes of the available response option set: VAS with only the extremes labeled, VAS with extremes and middle labeled, three labeled equal sized circles, 4 labeled equal sized circles, and 4 labeled graduated circles. The sixth response option set offered 4 labeled graduated circles but only one picture centered over the circles to demonstrate the response options. Participants reported

preferring the circles over the VAS (74%) and the graduated circles over the circles of the same size (68%). Rates of consistent responding also favored the circles over the VAS. Consistency of responding across response option types was at least 80% for approximately half of the matched items that differed in their use of graduated or same sized circles, and was at least 80% for only 25% of the items when any form of a VAS was used. The authors did not specify how they assessed consistency for the VAS response options.

Respondent impressions about and response consistency between VAS scales and vertical polytomous structures were examined by van Laerhoven, van der Zaag-Loonen, and Derkx (2004). The researchers asked 120 children between the ages of 6 and 18 years to complete the same seven questions three times using three different sets of response options. The children completed the three versions in succession, but were not allowed to review their answers to a version once it was completed. One version used a 5-point frequency vertical scale, another used a VAS consisting of a horizontal number line labeled from 1 through 10, and the third used a VAS that was a horizontal line with only the extreme ends labeled. After completing the set of 3 questionnaires, the participants were asked to use a rating scale of 1 to 10 to indicate how much they liked each type of response option and how easy they found each type of response option. Wilcoxon tests revealed that the frequency scale was scored as significantly easier and more preferred than either VAS scale ($p < .005$). Consistency of responses across the three response option types was found to be relatively high, with correlation values ranging from .76 to .82.

de Tovar and colleagues (2010) recruited 131 children between the ages of 5 and 15 years to self-report on pain levels after surgery. All children were asked to provide

pain ratings using two scales. One scale was a VAS with 10 numbered points arranged horizontally. The children slid a marker along the scale to indicate their choice. The other scale consisted of 5 black and white drawings of faces exhibiting increasing amounts of pain; the study authors assigned the scale 10 points at 2-point intervals. The children were divided into three age groups of approximately equal sizes: 5 to 6 years, 7 to 10 years, and 11 to 15 years. Nine of the youngest children provided the maximum possible scores compared with 3 children in each of the other two age groups. The correlations between the scale scores were .71, .66, and .88 for the youngest, middle, and oldest age groups, respectively. Seventy percent of the children preferred the faces scale over the sliding scale.

The above three studies demonstrate that children seem to prefer response option sets that provide multiple, discrete response options over sets that use a horizontal line. Whether responses are consistent across response option types remains in question. One study showed better consistency between option sets with discrete categories than between option sets that do not provide discrete categories. The other two showed high consistency irrespective of response option types. No studies were located that compared the item structures being used in this study: horizontal numeric polytomous (on the BDI-Y) versus fully labeled vertical polytomous (on the CDI). The next set of studies will discuss literature that examines how many response options gather the most accurate data.

Number of response options. Several studies have examined the effects of offering different numbers of response options when administering self-report questionnaires to children. The findings from the studies seem to indicate that the optimal number of response options is between four and five.

Borgers and Hox (2001), whose study was cited in the section above on item length, tested differences in non-response rates for 8- to 18-year-olds with items that varied in their numbers of available response options. Using data that were gathered in the course of five earlier studies on educational research, the number of response options was found to be a significant coefficient in a logistic regression predicting rates of response ($p < .05$). Three response options were associated with response rates of 88%, four options with 92% response rates, five with 85% response rates, seven with 74% response rates, and ten with 65% response rates. The results pointed to four response options as the most likely to maximize item response rates.

Borgers, Hox, and Sikkel (2004), in a study cited in the section above on negative item wording, investigated the effects of the number of response options on corrected item-total correlations and response stability over time with 8- to 16-year-olds. Fisher-Z transformed corrected item-total correlations for the two measures used in the study were regressed on number of response options, and this parameter estimate was significant at the $p < .05$ level ($\beta = 0.05$). Also significant ($p < .05$) in the regression was the squared number of responses options ($\beta = -0.02$). The authors plotted the number of response options and the predicted corrected item-total correlations, and found that the predicted corrected item-total correlations increased until the number of response options reached six and then began to decline. On the basis of this finding, the authors concluded that internal stability was maximized when approximately six response options are offered. However, the authors found a slightly different number of optimal response options when they examined response stability over time. Multilevel logistic regressions in which the proportional differences between time 1 and time 2 measurements were regressed on the number of response options did not produce a significant effect of the number of response

options ($\beta = -0.02$). An identical model which regressed the absolute difference between the two measurement waves did find a statistically significant ($\beta = -0.19$, $p < .05$) effect of the number of response options. The authors again plotted their predicted results, which this time were the predicted differences between the time one and time two measurement scores over a range of available response option numbers. This time, the plot revealed that increasing numbers of response options decreased consistency between the two measurement times. The inconsistencies peaked at approximately seven or eight options, and then began to decline again. Comparing the internal consistency analyses with the retest analyses, the authors concluded that approximately four options maximizes internal consistency and minimizes retest inconsistencies.

Rebok et al. (2001), whose study was already discussed as exploring different types of response options, also asked participants in one of their studies to report their preferences for a four or five point response option set. In the course of asking 60 participants aged 5 to 11 years old to explain their response processes to 32 questionnaire items, the authors presented six items in both a four point and five point scale. Participants were asked to report which they preferred and which was easier. Sixty-two percent of the participants said the five point option was easier, and 67% said they liked it better.

The above three studies concluded that response sets consisting of between four and five response options were associated with minimized item non-response, maximized scale internal consistency, and reduced response instability over time. When asked, participants reported preferring five options over four. One difference between a four and five point scale that has not yet been addressed is the potential presence of a midpoint in

a five point scale. The next set of studies will discuss the effects of midpoints in response option sets.

Midpoints in response option sets. As discussed above, midpoints are typically intended to represent neutral fulcrum between the two ends of bipolar scales. The two studies presented below present evidence that the use of neutral midpoints may not differ much across age ranges, but may have a small effect on response stability over time.

The first study was conducted by Raajimakers, et al. (2000). The data for the study were extracted from a larger written panel survey on adolescent development. The data were responses to 31 political attitude questions that used a 5-point Likert-type response scale (*strongly agree*, *agree*, *neither agree nor disagree*, *disagree*, and *strongly disagree*). Participants included 1883 adolescents and young adults between the ages of 12 and 24 years. The participants were paneled once in September of 1991 and again in September of 1994. Rates of midpoint use were treated as a dependent variable in an ANOVA, where participant ages were treated as a categorical independent variable. Twelve- to 14-year-olds were entered into the ANOVA as one age group, 15- to 17-year-olds in another, 18- to 20-year-olds in the third, and 21- to 24-year-olds in the last. There were no significant differences between the age groups in rates of midpoint usage in either the 1991 panel ($\eta^2 = .01$) or the 1994 panel ($\eta^2 = .00$). Differences between the two panels were also minimal, with no age group changing their usage of midpoints by more than 4% between 1991 and 1994.

The second study includes another component from the Borgers et al. (2004) work, discussed in sections above. Borgers et al. asked ninety-one 8- to 16-year-olds to participate in a test-retest study at 3 to 8 week intervals. Participants received one of 12 formats of a 19-item questionnaire. The questionnaires varied between using 2 and 7

response options. Questionnaires with an odd-number of response options were written to include neutral midpoints. The relative differences between the time one and time two responses were entered into a multilevel logistic regression as a dependent variable (i.e., the absolute difference between time one and time two divided by the total number of response options). In that model, items were considered to be nested within children. Neutral midpoints demonstrated a small but significant effect on the differences between responses at the two measurement waves ($\beta = 0.22$, $p < .05$). Because the dependent variable was based on absolute difference values, it is unknown whether the time two values were higher or lower than the time one values. All that can be determined is that the two values are different.

In summary, the two studies above demonstrate that the presence of neutral midpoints seems to have little effect on the stability of participants' responses over time. However, large age ranges were collapsed in each of these studies, perhaps masking differences across age groups. In the next and last section on response option structure, response option wording will be discussed.

Response option wording. Two studies were identified that discussed the effects of response option wording. The first, by Borgers, Hox, and Sikkel (2003) made use of - the same ninety-one 8- 16-year-olds and 3 to 8 week retest design as the Borgers, et al. (2004) study mentioned several times above. In this component of the study, participants were asked to respond to 34 items about daily activities, reading habits, and parental educational support. The items all made use of a 5-point response option scale, but participants received one of three possible versions of the response scales. One version consisted of fully labeled vague options: *never*, *sometimes*, *often*, and *very often*. The second version consisted of vague response options labeled only at the two endpoints:

never and *very often*. The final version consisted of fully labeled concrete options: *never*, *less than once a month*, *about once a month*, *once a week*, and *daily*. The absolute difference between the measure administrations was entered as a dependent variable into a multilevel regression wherein items were nested within children. The only significant effect was for the interaction between age and type of labeling ($\beta = -0.02$, $p < .05$). To the authors' surprise, response stability with children younger than 10 years was not affected by the different labeling conditions. Response stability in children older than 10 years was higher under the fully labeled response option condition (irrespective of wording vagueness). The authors concluded that the younger children could not take advantage of the additional information offered by the fully labeled concrete option set, and that the vague and partially labeled response option conditions were difficult for all respondents. The poorer performance of the younger children was attributed to poorer cognitive skills in children younger than 10 or 11 years of age.

While the use of concrete versus vague response option labels was not significantly related to response stability, the three response option sets did form different factor structures. The authors attempted to model all three response option sets as multiple groups in a single confirmatory factor analysis model. Responses from all participants were used to generate the models. Any models that constrained the item parameters to equality fit the data poorly, which led the authors to conclude the presence of different factor structures across the three labeling conditions. Because model fit modification indices were the smallest for the fully labeled clear options and large for the other two formats, the authors concluded that vague response options result in poor relations between items and latent factors. Differences in model structure across age groups were not explored. On the basis of the factor analysis results, the authors

concluded that fully labeled concrete response options provide the best quality data in children's self-report items.

In the second study, Betts and Hartley (2012) recruited 187 children aged 9 to 11 years old to participate in a study of the effects of response option order and wording on children's responding. The authors created 5 scenarios about two children cheating on a school test, and asked the participants to respond to questions about two of them. Each set of questions consisted of six items. The authors created four possible response option sets for the items. In all sets, the response options for three items were negatively worded and the response options for the other three items were positively worded. All response option sets consisted of 6 horizontally-arranged numbers (1 to 6) with unipolar anchor wording. Across the four response option sets, the authors varied whether the most extreme anchor word endorsing the item was placed on the left or the right and they also varied whether the highest number was placed on the left or the right. The four resultant response option sets were: most extreme and highest number on the left, most extreme and lowest number on the left, least extreme and highest number on the left, and least extreme and lowest number on the left. One to two months later, 130 of the children completed the same procedure that they had during the first administration. The authors analyzed the data using a series of ANOVAS and found that at both times children provided the highest ratings when the most extreme anchor words and largest response numbers were on the left ($p < .01$ for both effects). They did not find significant effects for age.

To review, the first of these studies demonstrated that fully labeled options seem to be associated with better response stability over time for children older than 10 years of age, but not for children younger than 10 years old. One possible explanation for these

findings may be that the cognitive skills of children under the age of 10 years are not as developed as the cognitive skills of older children. This study also indicated that items with concrete response options may be better related to the latent factor and thereby produce more accurate data. The second study showed that children's scores on measures that use horizontally-oriented response option sets are higher when the positive anchors and larger scale numbers are located on the left.

Age and use of response option set anchors. Four studies were found which examined children's use of the extreme ends of response option sets. Three found evidence supporting the hypothesis that younger children are more prone to using extreme response options than younger children. Two of those studies were cited elsewhere in this paper. The fourth study did not find that younger children were more likely to use the extreme ends of a response option set. It, too, has been cited elsewhere in this paper.

The first study (Chambers & Johnston, 2002; cited in the *Studies examining items containing emotional content* portion of the *Children and Retrospective Frequency Appraisals* section) asked 60 children between the ages of 5 and 12 years to respond to 18 items spread across 3 scales: one that measured their perceptions about a physical task, one that measured their perceptions about how someone else might feel in a given situation, and one that measured how they might feel in a particular situation. The authors found that the 5- to 6-year-old group provided more extreme scores than the other ages to the 2 scales that assessed how they and others might feel in a hypothetical situation. The authors also found that the 7- to 9-year-old children provided more extreme scores than the 10- to 12-year-old children to the task measuring other people's, but not their own, feelings.

The second study already cited (de Tovar et al., 2010; cited in the *Response option wording* portion of the *Children and Item Structure* section) compared children's responses to a VAS scale and scale that used pictures for the response option set. The 131 participants ranged in age from 5 to 15 years. Nine of the children between the ages of 5 and 6 years provided the maximum possible scores on the measures, compared with three children in the 7- to 10-year-old group and three children in the 11- to 15-year-old group.

A third study also found that younger children are more likely to use the extreme ends of a response option set. Rubie-Davies and Hattie (2012) analyzed children's responses to a 135-item measure designed to assess children's beliefs about several domains related to their schooling. The three major domains of the measure were self-concept, motivation, and class climate. The self-concept domain measured children's beliefs about their peer relations, reading, math, school, and their personal qualities. The motivation scale measured children's sense of competence, self-efficacy, mastery, performance, interest, and utility regarding math and reading. The class climate domain measured children's beliefs about their academic competence, satisfaction with school, and their teachers' and peer's academic and personal support. All items used a numeric 5-point bipolar Likert-type scale where 1 was an extreme negative and 5 was an extreme positive. A total of 2298 children participated, and they ranged in age from 7 to 12 years old. Using Wilks' Lambda, the authors found that children's grade year in school was significantly related to their use of the extreme upper end of the scale, with younger children more likely to use the extreme positive response than older children ($p < .01$). Differences between specific age groups were not explored.

A fourth study (Betts & Hartley, 2012; cited in the *Response option wording* portion of the *Children and Item Structure* section) found results that contrast with the

prior three studies. The authors (Betts & Hartley, 2012) examined the effects of response option wording and order on children's responding. The authors created four response option sets which varied in the location of the most extreme endorsement anchor word and the ordering of the response option numbers. One hundred thirty children between the ages of 9 and 11 years answered 12 items about two hypothetical scenarios, and answered the same items again one to two months later. Using a series of ANOVAs, the authors found significant effects for the ordering of the anchor words and response option numbering ($p < .01$ for both effects), but they did not find an effect for age. In this study, younger children did not provide more extreme scores than older children.

The studies reviewed on children and the use of extreme response options tended to show that younger children are more likely than older children to provide extreme responses, particularly to items that include emotional content. The one study that provided results inconsistent with these findings may not have shown age effects because they studied children within the limited age range of 9 to 11 years. In contrast, the remaining studies included a broader range of ages and younger children. Those studies tended to find significant effects between the youngest participants and participants in the 9- to 11-year-old age range. Because of the limited number of studies, however, firm conclusions about children's age and the use of the extreme ends of response sets are premature.

Summary of response option research. In summary, the studies described above demonstrate that for children in the concrete operational stage differences in response option structure can be associated with differences in response stability over time, latent trait structure, item comprehension, and use of response option extremes. The optimal number of response options for children in the concrete operational stage appears to be

four or five. Responses to the fully labeled concrete response option sets seem to be more stable over time in children older than 10 and to be more strongly related to the latent factor. Vague and complex wording can be associated with higher rates of response instability over time. Children 10 years of age and older may be more sensitive to differences in response option structure, possibly because their cognitive skills are more sophisticated. Younger children may be more likely to provide extreme scores, particularly positive ones.

Summary and Rationale

Children's self-reports on their internal states are increasingly considered an important component of diagnosis and assessment. Gathering these reports often takes the form of retrospective frequency appraisals of emotional experiences. Literature on these reports is more extensive with adults, and shows that adults tend to produce inaccurate accounts. Children, whose self-reporting skills are not as advanced as those of adults, also seem to have difficulty with these reports, as evidenced in studies demonstrating that children do not always understand the time concepts described in these items, provide reports that are inconsistent with observations of researchers and other adults, and provide unstable responses to such items in test-retest procedures. The scant literature in this area seems to show that children's facility with these items increases with age. However, some methodological limitations with these studies prevent any strong conclusions from being drawn. First, most of the studies that included retrospective frequency structures included those structures as part of a larger set of items. As a result, these studies often included a limited number retrospective frequency items and did not control for confounds such as differences in item content or structure. Second, only a few of these studies examined differences across age ranges. Instead, children from large age

groups tended to be treated as similarly competent reporters or only a very limited age range.

Research on the relation between item structure and children's responding, while larger than the literature on children's response to retrospective frequency items, is also slim. Findings from adult research strongly indicate that the responses adults choose can be affected by a number item stem and response set features. Findings from research with children also show that children's responding can be affected by features of item structure. Items that make use of negative wording or vague response options can form different factor structures from positively worded items or concretely labeled response options. Response stability over time is improved by the use of 4 to 5 fully labeled response options and wording that is simple and concrete. Whether the amount of wording affects any aspect of responding remains unclear. While there was some evidence that older children's responses are generally more affected by differences in item structure than those of younger children, there was little work in this area. As in the literature encompassing retrospective frequency reports, the limited research on differences in responding across ages often included large ranges of ages or compared children inside and outside of the concrete operational stage.

This study aims to supplement existing research by examining the functioning of retrospective frequency structures in comparison to other structures. This study also aims to examine that functioning across age groups within the concrete operational stage. Children within the concrete operational stage will be the target population because children in this age group are often asked to respond to retrospective frequency items, but what little is known about these children's abilities to respond to these items indicates

that these items may perform poorly in this age group, and do so differentially across the group's age levels.

The associated research questions for this study are:

1. Are retrospective frequency self-report items associated with different information than self-report items that do not use retrospective frequency reports?
2. Is age related to the information gathered by self-report items?

Answers to Research Question 1 will help assessors and instrument developers identify whether different item structures vary in the amount of information they gather and whether the locations on the latent at which they gather the most information differ. This information, by extension, will ultimately help match item structures to measurement goals. Answers to Research Question 2 will help researchers design self-report items that are appropriate to children's self-reporting abilities, and also help clarify differences in self-reporting abilities across 9- and 12-year-olds.

Method

The data to be used in this study were gathered as part of a larger study examining outcomes for a manualized treatment of depression for school-aged children. The study included a number of measures, including two measures designed to assess depression in children: the Beck Inventory for Youth-Depression (BDI-Y; Beck et al., 2001), and the Children' Depression Inventory (CDI; Kovacs, 1992). This study will use data gathered with these two depression measures. Although the measures assess depression symptoms, this study is a study of the psychometric properties of specific item structures, not a study of depression measures. The measures were chosen because they contain items that are matched for content but make use of varying item structures.

Because the original study was examining response to treatment intervention, the measures were administered to study participants multiple times. This study only uses data gathered from the initial participant screening process.

Participant Characteristics

Participants were 1659 girls in fourth through eighth grades from 13 middle and elementary schools in suburban Texas. The area's public school district serves approximately 22,000 students and includes elementary, middle, and high schools. Three Christian private schools and one private alternative education school are located in the school district; total enrollment for these schools was not available. Using mutually exclusive categories based on 2011 state school district information, approximately 20% of the district is African-American, 30% is White, 40% is Hispanic, and the remaining students are from other racial/ethnic backgrounds. Those same state data show that approximately 50% of the students in the region are economically disadvantaged (i.e., are eligible for free meals under the National School Lunch Program), and 15% have limited English proficiency. The per capita average income in the area is approximately $25,000, and the median home income is approximately $70,000. Economic information about the participants in this study was not available.

Of the 13 schools included in this study, three of the schools were middle schools, and the remaining schools were elementary schools. Intraclass correlation coefficients (*ICCs*) based on the CDI's and BDI-Y's summed scores for all items in the measures showed that students' responses on the depression inventories were not clustered by schools—the *ICC* for both the CDI and BDI-Y responses was less than .01. At levels close to 0, the absence of correlated data can be concluded (Killip, Mahfoud, & Pierce, 2004).

Unlike the state data reported above, in this study race data were collected separately from ethnicity. Just under 2% of participants reported themselves as being of American Indian racial heritage, 6% reported being Asian, 19% African-American or Black, just under 1% Native Hawaiian or other Pacific Islander, and 34% White. Participants were allowed to select multiple race options, and an additional 9% reported being from multiple racial backgrounds. A total of 11% selected an option to decline reporting racial information, and data were missing for 20% of participants. To the ethnicity item, approximately 40% of participants reported being Hispanic, 60% reported not being Hispanic, and data were missing for less than 1% of participants.

Because 40% of the sample reported being Hispanic, mean responses and factor structures were compared across the Hispanic and non-Hispanic groups to determine whether the two groups provided different mean responses and whether the groups' responses had similar factor structures. Differences in either could indicate that the data should not be analyzed as representative of a single sample. The analyses of mean differences were conducted using *t*-tests on the full set of BDI-Y and CDI items. According to the *t*-test results, the groups' responses differed significantly on the BDI-Y, although the effect size was small. CDI *t*-test results showed that groups' responses did not differ significantly. Analyses of factor structure were conducted only on the items included in this study. Factor structure analyses failed to provide evidence that the groups' data have different factor structures. Given these results, the Hispanic and non-Hispanic groups are treated as a single sample throughout the remainder of these analyses. (Full details of the *t*-test and factor structure analyses are available in Appendix A.)

Sampling Procedures

Researchers and school staff explained the study to female students in two grades of participating elementary and middle schools, with grades ranging from 4th to 7th. Teachers collected parental consent and student assent forms. The data were collected over 7 different administration times between October 2002 and February 2007. These 7 cohorts were collapsed and used as one sample in the analyses here.

Measures

The analyses for this study were conducted on a subset of items from the CDI (Kovacs, 1992) and the BDI-Y (Beck et al., 2001). Psychometric properties of the two measures are explained below. Prior to providing psychometric information, the process of selecting the subset of items used in this study is explained first.

Item selection. Whereas all the BDI-Y items use retrospective frequency structures, the CDI items use a variety of response option structures, for instance frequency, degree of preference, and degree of experience intensity. Items on the BDI-Y all use a *never, sometimes, often, always* structure. CDI items use complete sentences and may be ordered from least to most or in the reverse direction. CDI item response sets that were not ordered in the same direction as the BDI-Y item response sets were reverse coded.

Because both measures assess for depression, a number of item pairs across the measures address similar topics. As a result, there are 5 pairs of items across the BDI-Y and CDI that are nearly identical in content and consistent in their use of frequency scales, and another set of 5 item pairs that are nearly identical in content but do not appear in a frequency format on the CDI.

The first set of item pairs consists of five items on each measure that are matched for content and their use of a frequency structure. The five item pairs in this set address the topics of feeling sad, crying, social withdrawal, sleep quality, and loneliness. They are BDI-Y items18, 17, 16, 5, and 8; and CDI items 1, 10, 12, 16, and 20[1]. As a hypothetical example of these item pairs, the BDI-Y item might contain the item stem *I eat apples* followed by the response set *never, sometimes, often, always*. The matching CDI item would consist of three sentences such as *I never eat apples*, *I sometimes eat apples*, *I eat apples all the time.*

The second set of item pairs consists of another 5 items on each measure that are matched for content, but these item pairs are mismatched for their use of frequency structures. The five item pairs in this set address the topics of beliefs about the future, self-loathing, self-blame, suicidal ideation, and feeling loved. They are BDI-Y items 20, 15, 7, 4, and 6; and CDI items 2, 7, 8, 9, and 25. As a hypothetical example of these item pairs, the BDI-Y item might contain the item set *I watch TV* followed by the response set *never, sometimes, often, always*. The matching CDI item might contain the three sentences *I really like TV, TV is OK, I hate TV*.

Conclusions about the outcomes of using frequency response options based only on the five pairs of items mismatched for frequency scales would be inconclusive. If the pairs of items are found to function similarly, the conclusion could be easily made that the presence of a retrospective frequency task makes little impact on the response provided. However, the possibility that the items function similarly due to some unforeseen contributions of item structure could not be ruled out. Similarly, if the items

[1] Due to copyright restrictions, the items cannot be reproduced in this document.

were found to function in a dissimilar fashion, it would be impossible to determine whether differences in overall item structure or the use of retrospective frequency structures led to the results. Analyzing a second set of items matched for both content and the use of frequency structures will help to isolate the effects of frequency structures from the effects of other item structure components. If these items were analyzed without also examining the previous set of five items, similar functioning in these items would seem to indicate that the effects of the retrospective frequency structures trumped the effects of the overall item structures. However, the possibility that the similarity in functioning was actually due to (rather than inhibited by) overall item structure could not be ruled out. In a similar vein, if the items were found to function differently, it would be very difficult determine the causes of those differences—there could have been no effects of retrospective frequency structure, or there could have been retrospective frequency structure effects that were compensated for by other factors about item structure.

Although the combined results of the two analysis sets will not be incontrovertible, their combination will be more informative than either set alone. The most definitive evidence for the effects of retrospective structures over and above the effects of other item features would come from analyses finding that the item pairs matched for content only function dissimilarly, while the items matched for both content and frequency structures function similarly. The most definitive evidence for the effects of item structure over the presence of retrospective frequency structures will come from analyses that find that the pairs of items function dissimilarly in both sets of analyses.

CDI. The Children's Depression Inventory (CDI; Kovacs, 1992) is a 27-item self-report instrument designed for use with children between the ages of 6 and 17 years. The CDI can be read to children or completed by children independently and requires

approximately 15 minutes to complete. The items are intended to reflect symptom areas of negative mood, interpersonal difficulties, low self-esteem, loss of interest, and general competence at tasks.

Each item on the CDI consists of three statements. Participants are asked to select the statement that most accurately reflects their experiences over the past two weeks. Items vary in the type of scale used. Some item statements are based on frequencies, while others are based on degree of endorsement. Items are ordered into two vertical columns that span two pages, and each item is enclosed by a box.

The CDI measures depressive symptom severity by summing respondents' scores across all items. Total scores range from 0 to 54 and are converted into *T* scores (mean of 50, standard deviation of 10) for scoring, normed within scoring groups. Two scoring groups exist for each gender: 7 to 12 and 13 to 17 years of age. The study proposed here makes use of data collected only from girls in the younger scoring group.

Kovacs (1992) reported that the norming population consisted of 1,266 school students ranging in age from 7 to 16 years. The sample included 592 boys aged 7 to 15 years old and 674 girls aged 7 to 16 years old. The mean score for girls aged 7 to 12 years was 9.00. The mean for the 7- to 12-year-old group of boys and girls was 10.5 and the standard deviation was 7.3. Gender-specific standard deviations were not reported. Total score ranges were also not reported, although the scoring form provides information about raw scores and maximum *T* score values. For girls aged 7 to 12 year, the highest reported raw score was 44, which is associated with a *T* score of 99.

Reviews of psychometric literature on the CDI by Craighead, Smucker, Craighead, and Ilardi (1998) and Carle, Millsap, and Cole (2008) typically find alpha levels ranging between the approximately .75 to .95for the CDI total score. Note,

however, that the CDI test score is not being used in the proposed research; only individual item responses were used.

BDI-Y. The Beck Inventory for Youth – Depression (BDI-Y; Beck et al., 2001) is a 20-item self-report instrument designed for use with children between the ages of 7 and 18 years. The measure requires approximately 5 to 10 minutes to complete. The items are intended to reflect symptom areas of negative thoughts, feelings of sadness, and physiological symptoms.

Each item on the BDI-Y uses the same 4-point frequency response scale: *never*, *sometimes*, *often*, and *always*. All items are placed on the left of the page in a horizontal array, and all responses are on the right. The measure is one page in length. Participants are asked to circle the frequency word that best describes their experiences over the previous two weeks.

The BDI-Y measures depressive symptom severity by summing respondents' scores across all items. Total raw scores range from 0 to 60. Separate scoring norms were developed for boys and girls in three different age ranges: 7 to 10, 11 to 14, and 15 to 18 years of age. In the current study, the relevant scoring norms are those for the 7- to 10- and 11- to 14-year-old girls.

According to Beck et al. (2001), the BDI-Y was developed using a standardization sample of 800 children between the ages of 7 and 14 years from multiple regions across the United States. The participants were selected to represent the ethnic composition of the general United States population.

Sample means were not reported by Beck et al. (2001) in the BDI-Y manual. Scores associated with *T* score ranges were reported. In the female 7 -to 10-year-old group, a raw score of 16 is associated with a *T* score of 50; in the 11- to 14-year-old

group, a raw score of 13 is associated with a *T* score of 50. Total raw score ranges were not reported, although the manual provides information about raw scores and maximum *T* score values. In the female 7- to 10-year-old group, a raw score of 60 (the maximum possible score) is associated with a *T* score of 97; in the 11- to 14-year-old group, a raw score of 53 or higher is associated with a *T* score of 100. Standard errors in the female group were 2.86 for the 7- to 10-year-old group and 2.35 for the 11- to 14-year-old group.

Beck et al. (2001) reported that Cronbach's alphas for the summed scores across the 20 items in the standardization sample were .91 and .90 for females and males, respectively, in the 7- to 10-year-old age group, and .91 and .92 for females and males, respectively, in the 11- to 14-year old age group. Seven-day test-retest correlations based on the standardization sample were .81 and .79 for females and males, respectively, in the 7- to 10-year old age group and .91 and .92 for females and males, respectively, in the 11- to 14-year-old age group.

Validity for the BDI-Y was assessed using a sample of 128 children between the ages of 7 and 14 years using correlations between BDI-Y and CDI scores. Correlations were evaluated for the group as a whole, and not stratified by age or gender. The overall correlation between BDI-Y and CDI scores was $r = .72$, $p < .01$ (Beck et al., 2001). Note, however, that the BDI-Y total score is not being used in the proposed research; only individual item responses will be used.

Research Design

The study in which the data were gathered was a study of depression treatment for girls. Participants were screened into the study over a 5-year period. The study used multiple rounds of screening to determine whether respondents were eligible to participate in the treatment component of the study. The data used in this research project

were extracted from this 5-year, cross-sectional set of screening data. Only the initial screening data were used; subsequent administrations were not included. The screening procedures used in the original treatment study are described below.

The CDI and BDI-Y were administered during the first round of screening. Researchers administered the measures to groups of 5-100 assenting students during the school day. Approximately 1 researcher was present for every 10 students in an administration group. Questionnaires were read aloud to students in the 4th grade groups and to any participants who requested the measures be read.

In the first year of the study, participants who scored below a cutoff score of 16 on the CDI (i.e., one standard deviation above the mean) were screened out of the study. Participants who returned scores of 16 or higher were administered the CDI again one week later. Participants who still had elevated scores were administered the Kiddie Schedule for Affective Disorders and Schizophrenia for School-Age Children IV (K-SADS, 1996) and given the option to join the study if found to have a diagnosis of depression. Participants whose symptoms did not meet a diagnosis of depression were screened out of the study. In subsequent years, participants who returned elevated scores on the initial administration of either the CDI or BDI-Y were immediately administered a brief symptom interview based on *DSM-IV-TR* criteria (American Psychiatric Association [APA], 2000). Participants who did not report elevated symptom levels were screened out of the study, while participants who reported elevated symptom levels were administered the K-SADS and given the option to join the study if found to have a diagnosis of depression.

Both the first year's and the later years' screening structures administered the BDI-Y and CDI to participants multiple times if participants screened into and joined the

longitudinal study. If participants did not screen into the study, they were administered the measures only at the first screening. Analyses for this study will use only data from the first screening, which contains scores from participants who did and did not join the treatment study and were thus not necessarily diagnosed with a depressive disorder.

Analysis

The primary form of analysis in this study is a type of item response theory model called the graded response model (GRM; Samejima, 1969). The analyses were conducted using the statistical software Mplus Version 6.1 (Muthén & Muthén, 2010). Mplus was used to conduct confirmatory factor analyses. In a few instances, statistics generated by item response theory models were compared using dependent samples *t*-tests using SAS software (9.2, 2008). SAS software was also used to transform CFA parameters into IRT parameters. Before listing the analyses that were conducted for each research questions, an overview of IRT and GRM is provided immediately below.

Item response theory (IRT). IRT models treat responses to items as a function of properties of the respondents and properties of the items. The next several paragraphs will introduce some of the basic premises of IRT models. Unless stated otherwise, the paragraphs below that introduce IRT are drawn from DeVellis (2003), Embretson and Reise (2000), Fletcher and Hattie (2004), Hambleton and Jones (1993), Reeve and Mâsse (2004), and Reise and Henson (2003).

Respondents are thought to possess some true degree of the latent variable being measured, sometimes called a latent ability or a latent trait level. In this study, the latent variable of interest is depression. IRT models assume that these latent ability levels are distributed normally in the population. In mathematical terms, this assumption translates to the assertion that the distribution of the population is symmetric around the mean and

that 99% of the population possesses a level of the latent ability that falls within -3 to 3 standard deviations from the mean. To what extent this assumption holds true for depression will be discussed in a later section on testing assumptions.

Under an IRT framework, items can be modeled on how strongly they are related to the latent variable and at which level(s) of the latent variable they best capture information. Items can express any combination of strength of relation to the latent variable and level of the latent variable. For instance, items can be weakly or strongly related to the latent variable and they can capture information at the low or high end of the latent variable continuum. The latent trait level at which an item gathers the most information is referred to as the item's difficulty level.

This simultaneous modeling of respondent and item properties lends IRT several distinct advantages. First, error is allowed to vary across items within a model, instead of the usual practice in which error is treated as a constant across all items in a model. When individual items are combined into a summary model to estimate a single latent trait level for individuals, the result is an error estimate that varies across levels of the latent trait.

Because IRT parameter estimates are generated for each individual item entered in the model, IRT models have the benefit of allowing for items of mixed format (such as with 3- or 4-point response options) to be entered into the model without deleterious effects on the parameter and ability estimates. The modeling of individual items in conjunction with respondent characteristics also means that item parameter estimates generated from one sample are generalizable to other samples. In other words, parameter estimates are not sample-dependent.

Numerous IRT models exist for various hypothesized relations between item responses and the latent trait level of the participant. For instance, some models can

capture theories about how much participants guess when responding to items and others are designed for use with binary response options. The above-described probabilistic relation between a response and the latent trait level of the participant holds true across the various IRT models. However, the various models estimate item and person characteristics differently in order to capture the different theorized relations between response probabilities and trait levels.

Before discussing the model used in this study, GRM (Samejima, 1969), the next several sections will discuss the mathematical foundations upon which IRT models are built. A review of these foundations will make GRM more understandable. GRM is a variation of a two-parameter logistic model. The two-parameter logistic model will be described first, followed by a discussion of GRM. The two-parameter logistic model and the GRM will both be thoroughly reviewed because the item parameters estimated by these models will be used to answer the proposed research questions.

Logistic model foundation of IRT. Logistic models are based on dichotomous dependent variables. In the simplest version of an IRT model, the dichotomous dependent variable is whether or not a participant endorses the item. To increase the range of the dependent variable's value in the IRT equation and center that value around 0, this dichotomy is entered into the IRT model as a natural logarithm of the odds of responding versus not responding:

$$\ln[P/(1-P)] \qquad [1]$$

There are two independent variables. One is the person's latent trait level, which is called *theta* and symbolized as θ. The second variable is the item's difficulty level, which is symbolized as β.The item difficulty level is the value of the latent trait at which it is equally likely that someone whose ability is that same value will endorse the item as it

that they will not endorse the item. For instance, if the difficulty of an item is 1, a person whose true level of the latent trait is 1 has a 50% chance of endorsing that item. People whose latent trait levels are less than 1 are less than 50% likely to endorse the item, while people whose latent trait levels are more than 1 are more than 50% likely to endorse the item. When the item's difficulty is subtracted from the person's latent trait level, the resulting formula is as follows:

$$\ln\left[\frac{P}{1-P}\right] = \theta - \beta \quad [2]$$

According to this formula, when the person's latent trait is equal to the item's difficulty level, the log odds are 0, thus the odds are 1, and the probability of endorsement is .50. Interpreted another way, a 50% probability of endorsement indicates that respondent's underlying trait level is equal to the difficulty level of the item.

If the antilog of equation 2 is taken, the equation can be written as follows:

$$P(X|\theta,\beta) = \frac{\exp(\theta-\beta)}{1+\exp(\theta-\beta)} \quad [3]$$

In formula 3, the left side of the equation continues to refer to the probability that an item will or will not be endorsed. It now also stipulates that the "item" is referred to as *X* and the probability of endorsing or not endorsing *X* is a function of θ and β.

The models discussed below are elaborations on this basic model.

Two-parameter logistic models. The two-parameter logistic model is the model from which the GRM was derived (Samejima, 1969). The two-parameter logistic model is named as such because it includes two parameters for item features instead of one (i.e., just β). The second item parameter represents the strength of the relation between the item and the latent construct. This second parameter is called the item's discrimination,

and is symbolized as α. Adding in the α parameter to equation 3 produces the formula for the two-parameter logistic model:

$$P(X|\theta,\beta,\alpha) = \frac{\exp[(\alpha(\theta-\beta))]}{1+\exp(\alpha(\theta-\beta))} \quad [4]$$

The dependent variable is now the probability (P) of endorsing an item (X) given the variables and parameters in the model (θ, β, α). The formula is still founded on a subtraction of the item difficulty from the participant's trait level (θ-β), but the difference in the two values is now weighted by the discrimination parameter (α).

In the two-parameter model, the value of the item difficulty parameter is still defined as the point on the latent variable continuum where the probability of endorsing the item is .50. Figure 1 demonstrates two different response probability curves with different difficulty values. Latent trait levels are plotted along the x-axis. The probability of endorsing an item is plotted along the y-axis. The curve drawn with the solid line crosses the .50 probability level at a difficulty level of approximately .80. Since the latent trait is normally distributed with 0 as the mean, a latent trait level of 0.80 can be interpreted as a latent trait level of 0.80 standard deviations above the mean. In other words, respondents at a true latent trait level of .80 standard deviations above the mean have a 50% chance of endorsing this item. The dashed curve represents an item with a difficulty level of 2.80 standard deviations above the mean. (For the sake of convenience, it will be assumed from this point forward that latent trait levels are always interpreted as standard deviations above or below the mean.) For this item, respondents with a true latent trait level of 2.80 have a 50% chance of endorsing the item. Note that respondents with a latent trait level of 2.80 have more than a 90% chance of endorsing the item drawn with the solid line. In contrast, respondents at a trait level 0.80 have less than a 10%

chance of endorsing the item drawn with the dashed line. This difference in response probabilities highlights the differences in the difficulty levels of the items—the dashed line represents a "harder" item that is unlikely to be endorsed by respondents at lower trait levels. Likewise, the item represented by the solid line is "easier" and thus highly likely to be endorsed by respondents at a high trait level.

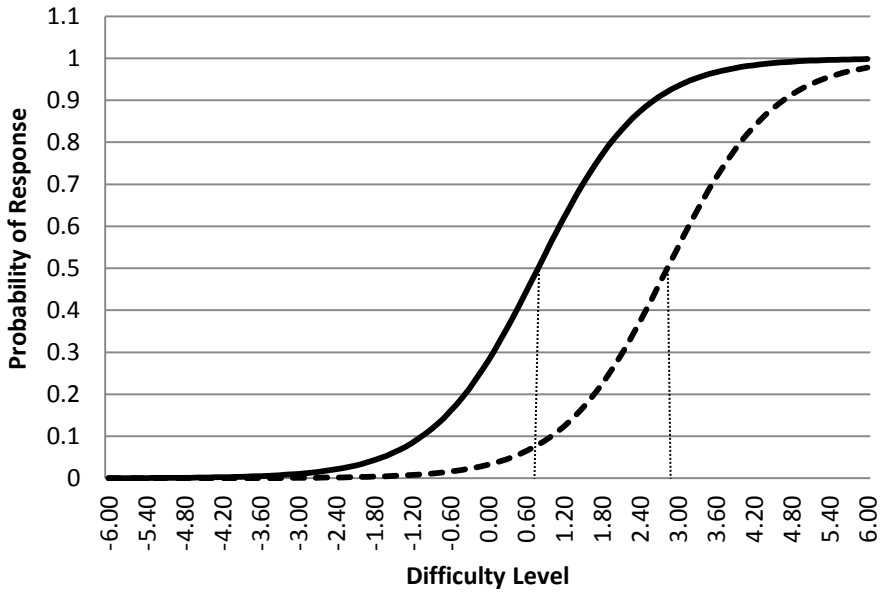

Figure 1. Example of curves with different difficulty values.

Item discrimination parameters represent the extent to which an item is related to the underlying trait, and are analogous to slope parameters. Like item difficulty levels, item discrimination parameters can be plotted. Whereas Figure 1 showed items with the same discrimination parameter values, Figure 2 demonstrates two response probability curves with different discrimination values. As before, the *x*-axis represents latent trait levels and the *y*-axis represents response probabilities. In this plot, the item represented by the dashed line is steeper than the item represented by the solid line, and can therefore be said to have a higher discrimination value. Both items, however, have the same

difficulty value because they cross the .50 probability level at the same level on the ability scale. The item with the higher discrimination value is said to be more strongly related to the latent trait because movement along the latent trait continuum is associated with larger changes in response probabilities for this item than for items with lower discrimination values. At a latent trait level of approximately 1.50, the probability of the item represented by the dashed line being endorsed is about 38%. The probability of the item represented by the solid line being endorsed is about 42%. At a latent trait level of approximately 2.50, the probability of the item represented by the dashed line being endorsed is about 64%. The probability of the item represented by the solid line being endorsed is about 55%. With a one-unit change in the latent trait level, response probabilities for the item represented by the dashed line differ by about 26 percentage points (38% to 64%). That same one-unit change is only associated with about a 13 percentage point difference in response probabilities for the item represented by the solid line. As a result, small differences in latent trait levels are more likely to be associated with differences in responding for the item represented by the dashed line than are the same amount of changes in the latent trait for the item represented by the solid line. Stated another way, the item represented by the dashed line is better able to distinguish between respondents of different underlying trait levels at trait levels where the slope is steep. In figure 2, for instance, the discrimination parameter represented by the solid line does a better job of distinguishing between respondents at latent trait levels of 1.80 and 1.90 than the discrimination parameter represented by the dashed line.

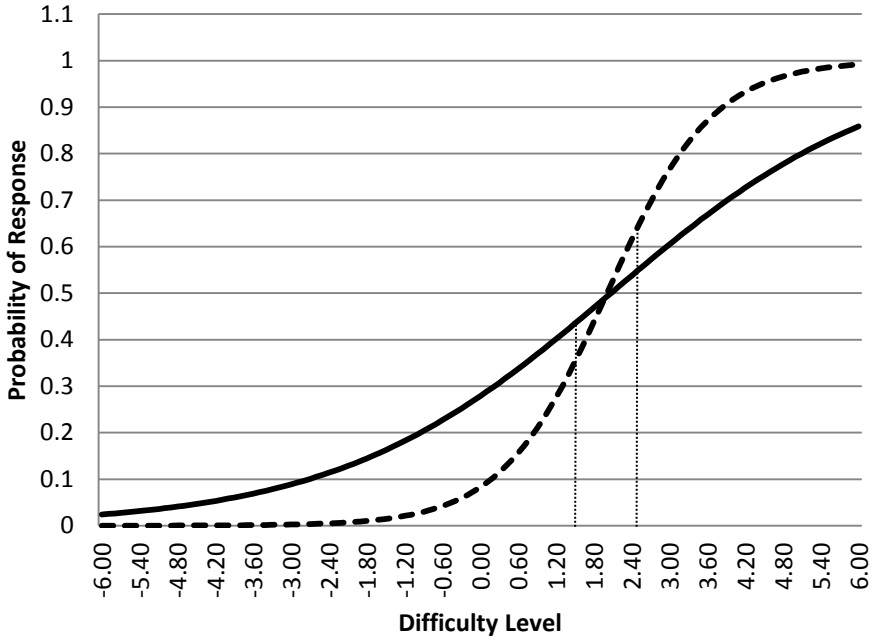

Figure 2. Probability curves with different discrimination values

Item discrimination and difficulty parameters and latent trait levels are also used in GRM. Their interpretation and application change somewhat in GRM because GRM models polytomous (i.e., multiple categories) instead of binary (e.g., endorsing or not endorsing) response patterns.

Graded response model - GRM. In the discussion that follows, the basic assumptions of GRM will be reviewed, followed by an introduction to the mathematical foundations of the model. As a part of the discussion about the mathematical foundations of GRM, four types of plots will be discussed: operating characteristic curves (OCCs), category response curves (CRCs), item information functions (IIFs), and test information functions (TIFs). The discussion of GRM will close with a brief overview of the GRM components that will be relied upon in the analyses proposed for this study.

GRM assumptions. Two-parameter models were originally designed for binary outcomes—(a) endorsing an item or (b) not endorsing an item. GRM allows for the modeling of items with more than two response categories. The following discussion of GRM is based on Samejima (1969). In GRM, the latent trait is still assumed to be continuous and normally distributed. GRM also assumes that the latent trait is monotonically increasing, which means that levels of the latent trait increase without ever decreasing. The response options are assumed to be ordered categories that represent the latent trait. Ordered categories are categories which are arranged in a certain order for the purposes of capturing changes in the trait being measured. For instance, height considered across age ranges is a continuous, but not monotonically increasing, trait. Height is continuous because it could be measured with varying degrees of precision using a ratio-based measure, such as a ruler. Height is not monotonic because it often declines as people reach old age. If ranges of height are divided into a discrete number of categories, the measure is then a set of categories whose order represents an increase (or decrease) in the trait. In the case of height, ordered categories might include the options of *shorter than average*, *average*, and *taller than average*.

GRM mathematical foundation. The definitions of difficulty parameters change somewhat in GRM from the two-parameter logistic models because more than two response categories are being modeled. Instead of modeling whether a participant does or does not endorse an item, GRM models the probability of endorsing or not endorsing each of the response categories within an item. As such, the overall item difficulty parameter is replaced by a set of parameters called threshold parameters. Like difficulty parameters, thresholds mark the point where a respondent has a 50% chance of responding in a particular way. Thresholds differ in that they are defined as the latent trait

level at which a respondent has a 50% probability of responding in a particular category or one below it and a 50% probability of selecting one of the higher categories. Because thresholds represent comparisons between sets of categories, there are one fewer threshold parameters than the total number of response categories. For instance, an item with four response categories would have three threshold parameters. The formulas used to model the variables and parameters of the GRM and the plots and data generated by GRM are discussed in the paragraphs that follow.

Estimating response probabilities (the dependent variable) is a two-step process in GRM. In the first step, probabilities of providing one of the lower versus one of the higher responses for each threshold are calculated, and those probabilities are plotted. The curves associated with this step are the operating characteristic curves (OCCs). In the second step, the probabilities of responding in a particular category are calculated. The curves associated with this step are the category response curves (CRCs). Each of these steps is discussed below.

Operating characteristic curves - OCCs. The formula for the first step of GRM, generating response probabilities for one of the lower versus one of the higher responses, is nearly identical to the formula used in the two-parameter model:

$$P^*(\theta, \beta, \alpha) = \frac{\exp(\alpha(\theta - \beta_j))}{1 + \exp(\alpha(\theta - \beta_j))} \quad [5]$$

In this formula, the dependent variable is the probability (P^*) that a person at a particular latent trait level (θ) will endorse a certain response category or one below it versus one of the higher categories. In this formula, P^* is used to denote probability because this formula refers to estimated probabilities for a randomly selected participant with a true latent trait level of θ. The discrimination parameter (α), which is constant for an item, still

acts as a multiplier of the difference between latent trait levels (θ) and what are now the estimated threshold values (β_j, where j is one of the thresholds) instead of the overall item difficulty level (β).

The probabilities of providing one of the lower versus one of the higher responses for each latent trait level are calculated, and the result is a set of values that can be plotted as a response probability curve, which in this instance is called an OCC. An OCC is plotted for each threshold (β_j). An example of OCCs for a 4-option item can be found in Figure 3. The *y*-axis represents the probabilities of response selection and the x-axis represents trait level. Points on each curve are interpreted as the probability that a participant at a given trait level on the *x*-axis will provide one of the higher response options versus one of the lower response options. As an example, Point A on Figure 3 indicates that a respondent with an underlying trait level of approximately 1.6 will have a 38% chance of providing response options 3 or 4 and, thus, a 62% chance of selecting categories 1 or 2.

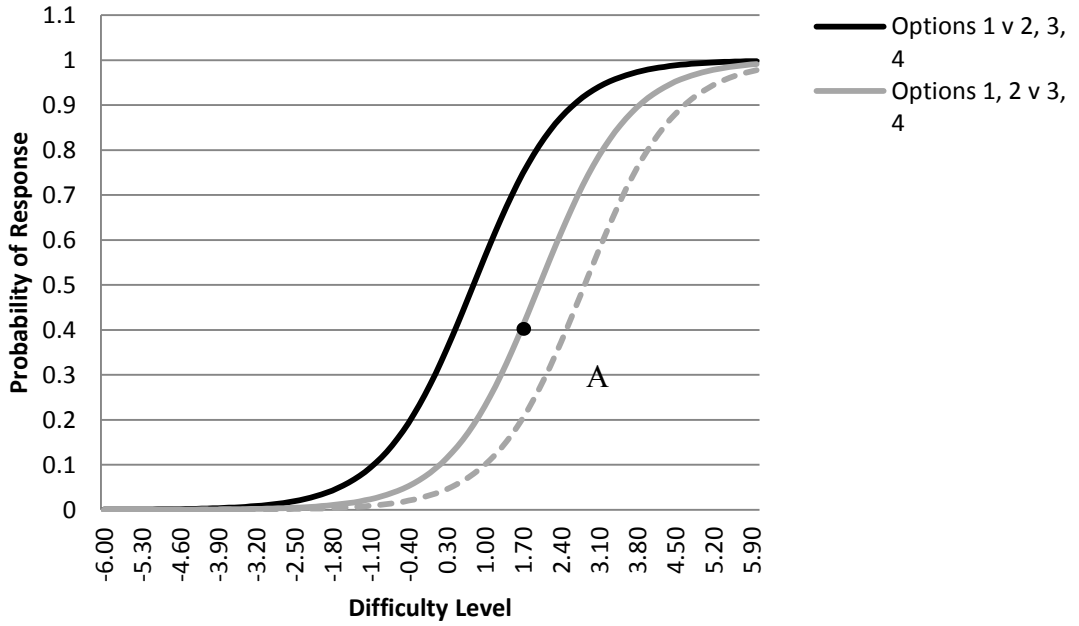

Figure 3. Example of OCCs.

Category response curves - CRCs. The second step of GRM is calculating the response probabilities for each individual category. The probabilities for each individual response category are calculated by subtracting the OCC for the estimated threshold above the response category from the OCC for the estimated threshold below the response category. Because there is no threshold below the first response category and no threshold above the last category, the CRCs for the first and last categories are calculated differently from the CRCs for the central categories. The calculations for each are explained below. An example of CRCs can be found in Figure 4. The plot is modeled on the same data used in the OCC example.

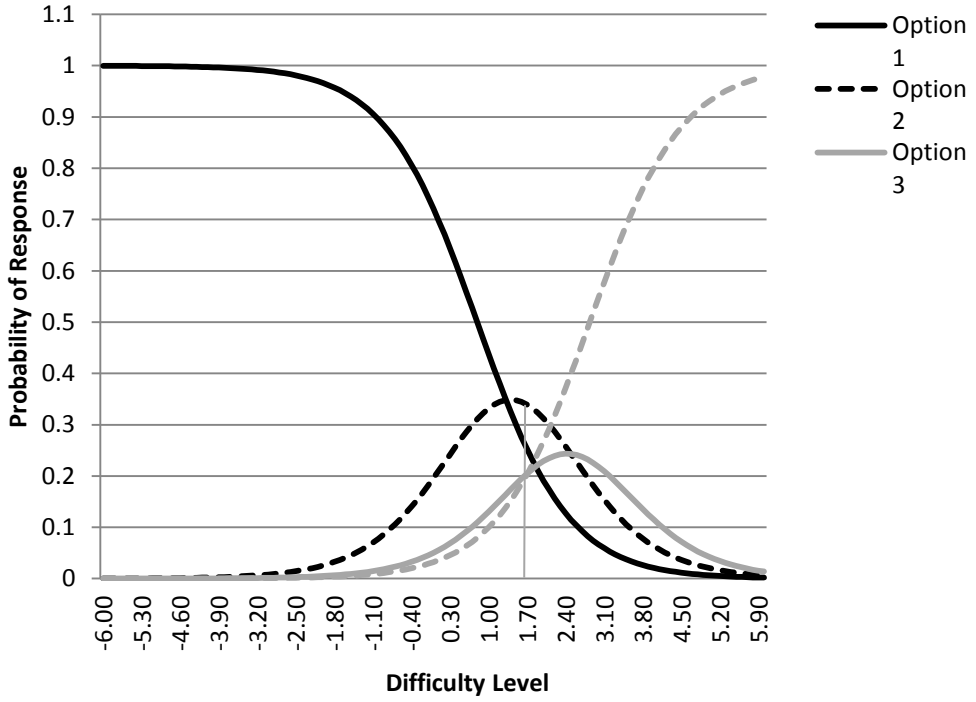

Figure 4. Example of CRCs.

As mentioned above, the CRCs for the central categories are the differences between the OCCs for the estimated thresholds above and below the category for which the CRC is being calculated. For an item with four response options, the options and thresholds can be numbered as follows:

Response	Never		Sometimes		Often		always
options	1		2		3		4
Thresholds		*1*		*2*		*3*	

The CRC formula for category 2 would therefore be:

$$P_2(\theta) = P_1^*(\theta) - P_2^*(\theta) \quad [6]$$

The CRC for category 3 could be calculated the same way, except that the OCC for the third estimated threshold would be subtracted from the OCC for the second estimated threshold. In Figure 4, the two middle curves correspond to the CRCs for response options 2 and 3. According to that figure, at a latent trait level of approximately 1.60, the probability of providing response option 2 is approximately .34. The probability of providing response option 3 is approximately .19.

Because there are no thresholds below the first response category or after the last response category, the CRC for the first category is as follows:

$$P_1(\theta) = 1 - P_1^*(\theta) \quad [7]$$

On Figure 4, this curve is represented as the darker bold line. At a latent trait level of 1.6, the probability of providing response category 1 is approximately .28.

The final CRC is calculated by plotting the OCC for the last estimated threshold. Technically, the procedures still involve subtracting the OCC for the threshold above the last category from the OCC for the threshold below the last category, but since no one

can respond above the final category, the final value to be subtracted is 0. On Figure 4, a latent trait level of 1.6 corresponds to approximately a .19 probability of providing response option 4.

The CRC values at a latent trait for a given response category can be added and compared to category values for an OCC. For instance, the OCC that models the probability of providing response options 3 or 4 versus options 1 and 2 is demonstrated in Figure 3 as a solid grey line. At Point A on that figure, the probability of a participant at a latent trait level of 1.6 providing response option 3 or 4 over response options 1 and 2 is .38. According to Figure 4, the probability of a respondent at a latent trait level of 1.6 providing response option 3 is .19. The probability of that respondent providing response option 4 is also .19. The sum of the CRCs for options 3 and 4 at that trait level is .38, the same value as the probability estimate provided by the OCC.

Item information functions - IIFs. Item information functions (IIFs) plot the amount of psychometric information an item contains at each point along the latent-trait continuum (Embretson & Reise, 2000, p. 183). IIFs are a function of discrimination and threshold parameters. Information is typically considered to be the inverse of the standard error across all levels of the latent trait continuum (θ):

$$Info_{(i|\theta)} = \frac{1}{SE_{\theta}^{2}} \quad [8]$$

where *i* refers to a given item, θ refers to a given difficulty level, and *SE* refers to standard error. Taking the first derivative of formula 8 and expanding it algebraically yields the following (Fitzpatrick, Choi, Chen, Hou, & Dodd, 1994):

$$Info_{(i|\theta)} = \frac{-\alpha_i \exp(\alpha_i(\theta-\beta_{i1}))/(1+\exp(\alpha_i(\theta-\beta_{i1})))^2}{P_{i1}} + \sum_{c=2}^{C-1} \frac{\alpha_i P_{i(c-1)}(1-P_{i(c-1)}) - P_{i(c)}(1-P_{i(c)})}{P_{ic}} +$$

$$\frac{\alpha_i \exp(\alpha_i(\theta-\beta_{i(C-1)}))/(1+\exp(\alpha_i(\theta-\beta_{i(C-1)})))^2}{P_{iC}} \quad [9]$$

In formula 9 *C* represents the total number of categories, *c* represents any given category, *i* represents a particular item, β refers to a threshold, and *P* refers to the probability of responding. Thus, notations such as β_{i1} and $P_{i(c-1)}$ refer to specific thresholds or response probabilities.

Over the range of the latent trait continuum (i.e., across values of θ), the item information function yields information curves, such as those found in Figure 5. These curves are often bell-shaped, but need not be. As Samejima (1969) notes, several factors can affect the shape and maximum height of information curves. Broad threshold ranges can result in rough, non-bell-shaped curves that may also be lower and flatter than the curves for items with a limited threshold range. As more thresholds are added, the information values usually increase, but low discrimination values or wide threshold ranges can decrease the overall height of an information curve. A review of item parameters alone, therefore, is often not sufficient to determine the shape of an item's information curve.

The curves presented in Figure 5 represent common versions of item information curves. The low, broad, curve in the figure represents an item that yields nearly equal information levels for difficulty levels ranging from -1 to 1.5. Thus, the item provides the same information amount of information for people at a trait level of -1 as it does for people at trait level 1.5. The other curve represents an item that has a more restricted range of threshold parameters and a larger discrimination value. The spread of this curve

is smaller, but it has a much higher peak signifying that it gathers much more information at its peak.

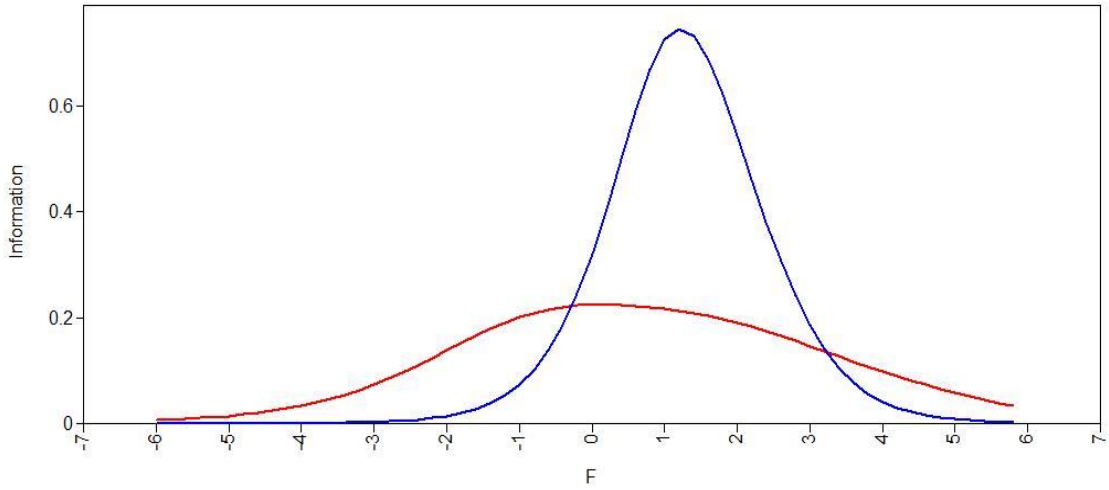

Figure 5. Example of IIFs with different slopes.

Items with more peaked IIFs are preferred because their discrimination values tend to be higher, which in turn means that they are better able to distinguish between participants at neighboring levels of the latent trait. The ideal location of IIF curve peaks depends on the purpose of the measure. A measure that is designed to identify only participants within a particular range of the latent trait should be populated with items that are focused within that latent trait range. For instance, a measure that is intended to identify only the most severely depressed respondents should consist of items that have high IIFs at the upper level of the latent trait continuum. A measure that is intended to distinguish among a range of latent trait levels should consist of items that cover a range of latent trait levels.

Test information functions - TIFs. Test information functions (TIFs) provide summaries of the information provided by groups of items (Hambleton & Jones, 1993). The plots look like IIFs and the interpretation of the curve shapes is similar, but not

identical. For example, if the curves in Figure 5 represented Test 1 and Test 2 instead of Item 1 and Item 2, the low, broad curve could still be interpreted as a test that captures a low level of information about a wide range of trait levels. The steeper curve could be said to perform well at identifying participants with the narrow range of trait levels falling under the curve. TIFs differ from IIFs in that a broader range in a TIF may not be undesirable, as in the case of an intelligence test designed to provide accurate test scores for a range of intelligence levels. If such a test consisted only of difficult items, participants at the lower end of the latent continuum would not be well-measured. Thus, a test that is designed to measure a wide range of latent trait values would necessarily consist of items with information peaks across a wide range of latent trait values. When the information levels of these items were summed to generate a TIF, they would produce a broad TIF. Ideally, that TIF would still be high.

IRT analyses. Of the IRT parameters and functions reviewed in the above section, this study primarily used TIFs because this study is mostly concerned with differences in the information gathered by groups of items, rather than individual items. Other parameters and functions, such as discriminations, difficulties, and item-level curves, were used to supplement the test-level analyses.

Assessment of model fit. The phrase *model fit* refers to how well parameters generated by a statistical model themselves can be used to generate data that approximate the observed sample data. The more closely the generated data set approximates the original data set, the better the model is said to *fit*. Because statistical models do not exactly replicate the original data, there are inherent differences between the observed data and estimated data generated by the model. As model fit declines, error levels in model-generated parameter estimates increase.

Numerous guidelines exist for the assessment of model fit, and different guidelines are appropriate for different models. In the present study, several possible indicators of model fit were available—the chi-square fit statistic, and four fit indices: root mean square error of approximation (RMSEA), comparative fit index (CFI), Tucker-Lewis index (TLI), and weighted root mean square residual (WRMR).

For a well-fitting model, the significance of the chi-square statistic should be greater than .05 (i.e., the statistic should be non-significant). However, the chi-square statistic is a powerful statistic that can yield statistically significant results for even models with very little misfit (Carle, Millsap, & Cole, 2008; Cook, Kallen, & Amtmann, 2009; Hu & Bentler, 1998).

Because the chi-square statistic is so frequently significant, research about the use of fit indices as an alternative to the chi-square statistic has flourished. The RMSEA is one of a class of fit indices referred to as error of approximation indices. It estimates the difference between the examined model and a hypothetical model in which every component in the model is related to every other component. The RMSEA is based on covariance estimates and is sensitive to the number of estimated parameters (p. 449, Cook et al., 2009). Recommended cutoff scores range from .1 or less to .05 or less (Carle, et al., 2008; Yu, 2002).

The CFI and TLI are highly correlated versions of comparative fit indices. They compare the fit of the examined model with a hypothetical, more restricted baseline model in which the model components are not correlated (Hu & Bentler, 1998; Yu, 2002). The CFI has the benefit over the TLI of being less sensitive to sample size. Recommended cutoff scores for both range from .90 or greater to .95 or greater (Carle, et al., 2008; Hu & Bentler, 1998; Yu, 2002).

The WRMR is a relatively newer and less-studied fit index that was introduced by Muthén and Muthén when they introduced M*plus* (Yu, 2002), the statistical software package being used to conduct much of statistical modeling conducted in this study. WRMR is a residual-based fit index that measures the average difference between the sample and estimated population variances and covariances (p. 16, Yu, 2002). According to a posting by Linda Muthén on the M*plus* website, however, the WRMR is not appropriate for use in multiple-group analyses (Muthén, 2013).

Based on the above literature, statistical models in this study were considered to demonstrate sufficient fit if the CFI was .95 or greater and the RMSEA was .05 or less.

Model estimators. Mplus offers a number of estimators for obtaining parameter estimates and standard errors. These estimators vary in their procedures and the information used to generate estimates. Estimators also vary in their suitability for different statistical tests. The weighted least squares with means and variances adjusted (WLSMV) estimator was used. This estimator is appropriate for categorical data and generates weighted least squares parameter estimates using a diagonal weight matrix with standard errors. The chi-square statistic is mean- and variance-adjusted and uses a full weight matrix (p. 533, Muthén & Muthén, 2010). A drawback to using WLSMV in IRT modeling in M*plus* versions prior to version 7 is that the CFA parameters must be transformed into IRT parameters. A SAS software program was used to transform the CFA parameters in IRT parameters, which is provided in Appendix B.

Model parameterization. M*plus* offers two parameterizations – delta and theta. According to Muthén and Muthén (2010), the difference between delta and theta parameterizations is a difference in which possible latent response variable parameters are included in the model. Latent response variables are the hypothetical variables

underlying the observable items to which participants respond; they are the "true" trait underlying participants' imperfect item responses. Because the observed variables (i.e., participants' responses) are imperfect measurements of the "true" latent response variables, the observed variables contain error. Given that there is not enough observed information in a model to estimate all possible parameters for a latent response variable, researchers must choose which latent response variable parameters are modeled and which are fixed to an arbitrary value. In the delta parameterization, scale factors can be modeled. Scale factors are the variances of the latent response variables. In the theta parameterization, residual variances of the latent response variables can be modeled. Delta parameterization is the default because it typically generates better-fitting models when the data are categorical, particularly when the model is based on large sample sizes (Muthén & Muthén, 2010). It was also used in all the CFA models in this study, except for the item independence analysis.

Missing data. The WLSMV estimator treats missing data using pairwise deletion. Thus, all CFA models in this study treated missing data using pairwise deletion. Because the remaining statistical analyses were based on CFA parameters (e.g., *t*-tests of discrimination values), they, too used pairwise deletion.

Analyses Conducted for Research Question 1: Are Retrospective Frequency Self-Report Items Associated with Different Information than Self-Report Items that Do Not Use Retrospective Frequency Reports?

The analyses for Research Question 1 consisted of several steps. Before conducting analyses directly related to the research question, IRT model assumptions and fit were tested. Next, in the first step of the research question analyses, visual examinations of the test information curves were conducted to explore whether the curves

differ in height, location, and steepness. In the second step, statistical tests of differences in test information curves were conducted to determine whether the item sets gathered different amounts of information and to infer whether those differences are associated with differences in threshold locations. In the third and final step, statistical tests of differences between item discrimination parameters were used to supplement and confirm the conclusions drawn in the second step. The procedures performed at each step are discussed further below.

Testing model assumptions. As discussed above, the data for this study are responses to the BDI-Y and CDI and were collected as part of a study on the treatment of childhood depression. Since items from these measures will be analyzed in this study using IRT, this study is assuming that: (a) depression is a continuous latent construct, (b) depression is the same continuous construct across all ages in the sample from which the data were gathered, (c) items on the BDI-Y and the CDI are measuring the same construct, and (d) the items entered into the IRT model are locally independent. The first two assumptions cannot be statistically assessed in this study. They will be discussed with reference to relevant literature in the paragraphs immediately below. The second two assumptions can be tested in this study, and the procedures for testing those assumptions can be found below in the sections on unidimensionality and local independence.

The first assumption, that depression is a continuous latent construct, is supported by the manner in which it is diagnosed. For instance, although minimum criteria for diagnosis exist (APA, 2000), it is also recognized that people can present with symptoms of depression that do not meet diagnostic criteria for a major depressive disorder. Depression can be present at a variety of severity levels and can be in full or partial

remission. Consistent with these criteria, Judd et al. (1998) published (prior to the creation of the DSM-IV-TR criteria) the results of a 12-year study of outpatient depression treatment. They found that people with depression mostly experience depressive symptoms at levels that are below diagnostic criteria but still impairing, and they spend more time with symptoms than without. Critics of the current diagnostic system claim that the current diagnostic guidelines are arbitrary and fail to capture the dimensional and chronic nature of the disorder (e.g., Klein, 2008). Although a diagnostic category exists for major depressive disorder, an array of literature suggests that depression is a condition that can be experienced at a variety of severity levels, not a condition that either does or does not exist.

The second assumption, that depression is the same continuous construct across all ages in the study, is also generally considered to be true as is evidenced in discussions of symptom onset. Depression is thought to occur in children who are as young as 2 years of age (Stark, Sander, Hauser, Simpson, Schnoebelen, Glenn, & Molnar, 2006; Sterba, Egger, & Angold, 2003). While the expression of current DSM-IV-TR (APA, 2000) diagnostic criteria are thought to need some adjustments for children this young, those diagnostic criteria are thought to be applicable to children as early as age 3 (Luby & Belden, 2006). Depression is thought to follow a more severe course if onset is in childhood, such as being associated with higher rates of recurrence, chronicity and impairment in functioning (Hammen, Bistricky, & Ingram, 2010). Despite needing some diagnostic adaptations across developmental levels and presenting different lifetime courses with different ages of onset, depression is treated as a condition that can be expressed at nearly every age.

The final two assumptions, those of unidimensionality and local independence, can be assessed with the empirical data available in this study. Tests for both of these assumptions are explained in further detail below.

Local independence and unidimensionality. Local independence and unidimensionality can be tested simultaneously using confirmatory factor analysis (CFA). The paragraphs that follow discuss each assumption and then discuss the CFA method used for testing them.

The assumption of local independence does not require that the items have no relation to each other whatsoever. Instead, the assumption of local independence requires that any relatedness among items is completely accounted for by the parameters and variables in the IRT model—in this case the thresholds, discriminations, and individuals' theta levels (Embretson & Reise, 2000; Reeve & Mâsse, 2004). The assumption of unidimensionality means that the set of items measures only one latent trait.

A number of statistical procedures have been proposed to test for local independence and unidimensionality (Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001; Embretson & Reise, 2000; Reeve & Mâsse, 2004; Yen, 1984, 1983). This study used a CFA approach because CFA models generated in M*plus* have the benefit of testing simultaneously for local independence and unidimensionality. The absence of either will produce a poor-fitting model.

The items from the CDI and BDI-Y are treated as summative measures of a single latent construct (i.e., depression), and so all items being examined from the CDI and BDI-Y were entered into a single-factor CFA model in M*plus*. Because the individual items in each item pair included in this study are theoretically identical in content and some are also similar in their structure, it is possible that the items may be related in ways

not captured by the IRT model. To compensate for this theoretical non-independence, the errors for the items within the item pairs were covaried in all models conducted for this study. Modeling with correlated error terms has the effect of reducing errors in parameter estimates due to item relatedness not accounted for by the model.

Item independence was also evaluated according to Yen's (1984, 1993) guidelines. Those guidelines state that local item independence holds if item residuals are not correlated. As an approximate guideline, Yen recommends that the absolute value of the correlation of item residuals not be larger than .20. The item residual values can be generated by using a theta parameterization instead of the default delta parameterization used for the rest of the CFA analyses performed in this study.

The current study identified a set of non-independent items using a combination of theoretically presumed item non-independence and the empirical guidelines provided by Yen (1984, 1993). First, a single-factor model that included the theoretically presumed item-error correlations was tested. If item correlations were less than Yen's suggested cutoff values, the correlations were deleted from the model. Additional possible related items were sought in the M*plus* modindices output, an output set that lists possible model alterations along with their potential effects on the chi-square fit statistic. Item-error correlations with large values were included in the model; they were retained if their correlations values met Yen's guidelines. The final model and fit information are discussed further in the *Results* section.

Step 1. Visual analyses of the test information curves for each of the four sets of items. Once a well-fitting, single factor model with conditionally independent items was generated, a test information curve was generated for each of the four item sets: BDI-Y items selected to match CDI items for content and the use of retrospective frequency

structures, CDI items selected to match BDI-Y items for content and frequency structures, BDI-Y items selected to match CDI items for content only, and CDI items selected to match BDI-Y items for content only. These curves were examined for differences in height, location, and steepness. Verbal comparisons of these curve dynamics were generated with the goal of discussing the amount and location of information gathered by the item sets. In general, curves that are more "peaked" were considered to reflect a higher amount of information gathered, and curves with low and broad peaks were considered to show poor discrimination between respondents at different trait levels. If the curves peak at different locations, the thresholds for the items were not said to be centered at the same difficulty levels.[2]

Step 2. Statistical tests of differences between item discrimination parameters. Tests of the discrimination parameters were used to supplement the conclusions from Step 1. If the results of Step 1 indicated that the thresholds may not be contributing to curve differences, then significant discrimination value differences were considered indications that curve differences were due in part to discrimination differences. If Step 1 results indicated that thresholds are affecting curve differences, then significant differences in discriminations were considered indications that these effects were amplified by differences in the discrimination values.

Mplus generates a discrimination parameter for each item entered into the model. Thus, there is one discrimination parameter for each item. Because there are only 5 item pairs in each data set, there are only 5 data points for each item set. The difference scores

---

[2] Matched-sample statistical tests of differences in the information curves were attempted but not successful. These test efforts are described in Appendix C.

for the item pairs were distributed normally, and so the tests were conducted using matched-sample t-tests.

Cohen's *d* (Cohen, 1988) effect sizes were also generated using the following formula:

$$d = \frac{|absolute\ value\ of\ mean\ difference\ scores|}{standard\ deviation\ of\ difference\ scores} \quad [10]$$

It is not known if effect size interpretation guidelines are suitable for interpreting statistics that test for differences in item parameters (i.e., as opposed to differences in scores produced by groups of people). To aid in the interpretation of the significance tests and effect sizes, the results include information about the relation between each item set's discrimination values and the latent factor's construct reliability as defined by Fornell and Larcker (1981). Fornell and Larcker define construct reliability (RC; a measure of internal consistency) as a function of CFA squared standardized item loadings:

$$RC = \frac{(\sum l)^2}{(\sum l)^2 + \sum(1-l^2)} \quad [11]$$

The sum of the squared item loadings $[(\sum l)^2]$ is divided by the total sum of the squared item loadings $[(\sum l)^2]$, and the sum of one minus the squared loadings $[\sum(1 - l^2)]$. The result is a formula that produces increased reliability values as item loadings increase. In this study, the CFA loadings that were used to generate IRT parameters were also used in Fornell and Larcker's formula to generate construct reliability values.

To provide some context for the extent to which construct reliability is associated with changes in item-factor loadings, the relation between loading values and construct reliability is demonstrated in a table that includes construct reliability values across a range of total items in the measure and standardized loading values, available in Table 1. Although there are not extant guidelines about the amount of change in construct

reliability that might be considered meaningful, considering differences in loading values in the context of differences in construct reliability values will provide an initial point of investigation and some indication as to whether the effect sizes yielded by the *t*-tests are meaningful. Further work in a separate project would be needed to develop quantitative interpretation guidelines.

Table 1. Construct reliability values associated with number of items and average standardized loading value

Number of items	Standardized loading value									
	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
5	.21	.36	.47	.56	.63	.68	.73	.77	.80	.83
10	.34	.53	.64	.71	.77	.81	.84	.87	.89	.91
15	.44	.63	.73	.79	.83	.87	.89	.91	.92	.94
20	.51	.69	.78	.83	.87	.90	.92	.93	.94	.95
25	.57	.74	.82	.86	.89	.91	.93	.94	.95	.96
30	.61	.77	.84	.88	.91	.93	.94	.95	.96	.97

Summary of Research Question 1 analyses. The goal of these analyses was to provide information about whether the use of frequency structures is associated with the amount of information gathered by self-report items. The three analyses for Research Question 1 were conducted on two sets of item pairs—one set of pairs matched across the CDI and BDI-Y for content only, and another set matched for content and the use of retrospective frequency structures. The analyses included visual analyses of the item sets' test information curves and matched sample *t*-tests on the discrimination values.

Analyses Conducted for Research Question 2: Is Age Related to the Information Gathered by Self-Report Items?

As discussed in the literature review, children within the concrete operational stage are often treated in measurement research as having similar abilities to self-report. However, a growing body of evidence indicates that children within this range differ in their abilities to provide self-reports and respond to retrospective frequency items. The goal of this question is to investigate whether children within the concrete operational stage differ in their ability to self-report using retrospective frequency items.

The question was explored in two steps. First, measurement invariance across age groups was examined, a process that allowed for comparisons across age groups but not across measures. A benefit of this step is that it allowed for statistical tests of differences between the age groups' parameter estimates. A drawback of this approach is that it does provide fit statistics or indices for individual groups—the fit statistics and indices are provided for the model as a whole. Thus, if the model fits one group very well and another somewhat poorly, it is possible that the fit indices and statistics could indicate overall good model fit despite poor fit in one group. Modeling groups independently can therefore reveal differences between groups' model fit that are not obvious when the groups are entered into the same model. As such, the second step of Research Question 2 was independent modeling of the age groups. The independent modeling allowed for tests of parameter equivalency across measures but not age groups.

The same sets of item pairs examined in Research Question 1 were used in the Research Question 2 analyses.

Step 1: Measurement invariance across age groups. Items are said to be invariant if they are related to the latent construct in identical ways in all participant

groups (Cheung & Rensvold, 1999; Millsap & Yun-Tein, 2004; Vandenberg & Lance, 2000). In this study, analyses of measurement invariance focused on comparisons between the 9- and 12-year-old groups. Because this study is interested in two separate sets of items, the invariance analyses were conducted twice—once for each set of item pairs. The analyses consisted of two major components, each of which are described next.

First component of Step 1: Testing for measurement invariance. Measurement invariance can be explored through a variety of statistical methods. This study used a variation on the confirmatory factor analysis method proposed by Muthén and Muthén (2010). According to this method, the least restrictive model is tested first. The most restrictive model is tested next, and the difference between the model fit is tested using a chi-square test statistic. If the more restrictive model fits the data well (i.e., chi-square is not significant), the measure is considered invariant and no further tests are needed. If the model does not fit the data well (i.e., chi-square is significant), then parameters in the fully constrained model are relaxed until further changes to the model no longer result in statistically significant improvements to model fit. Parameter constraints can be released on a variety of criteria. In this study, constraints were released by measure and parameter: each measure's set of thresholds and then loadings were released in turn. If the model fit significantly improved when those parameters were released, those parameters were considered non-invariant across the age groups.

Some invariance must be assumed in order to provide the modeling program with enough data to perform calculations, a condition otherwise referred to as ensuring that the model is identified. Stated another way, some parameters and values must be artificially assigned. Under Muthén and Muthén's (2010) approach, the least restrictive model is defined as:

1. Thresholds and loadings free
2. Scale factors set to 1 in all groups
3. Factor means set to 0 in all groups

Scale factors are set to 1 and factor means are set to 0 for identification purposes; without such specifications, the model would not be identified. The model is otherwise considered unconstrained because the thresholds and loadings are free. (As a brief reminder, scale factors are the variances of the latent response variables; each item therefore has its own scale factor. Scale factors do not refer to the single factor that is associated with the latent trait.)

In addition to the constraints suggested by Muthén and Muthén (2010), the following constraints were also added:

4. Latent factor variances set to 1 in all groups
5. Covariance values for the matched item residuals constrained to equality across groups

Latent factor variances were set to 1 because this constraint is an assumption of the IRT model. Covariance values for the matched item residuals were constrained to equality so that these values did not become inflated to compensate for the constraints placed on the loadings and thresholds, i.e., to correct for item non-independence.

In the most restrictive model, the thresholds and loadings were constrained to equality across groups. Because constraining the loadings and thresholds added information to the model, only one group needed to have its scale factors set to 1 and factor means set to 0 in order to maintain model identification. Because the restrictions on the latent variances and correlated item errors are a part of the theoretical assumptions and error control procedures, those restrictions remained in the fully constrained model.

Second component of Step 1: IRT effect size indices of measurement invariance. The preceding component demonstrated whether or not measurement invariance exists across age groups. To help supplement the findings from the CFA analyses, three effect size indices were generated using the IRT parameters from the unconstrained model. As will become clearer in the explanations that follow, the effect size indices are based on comparisons between participants' scores using the two groups' parameter sets from the unconstrained model. Because the fully constrained model produces identical parameters for both groups, those parameters cannot be used.

Loosely speaking, the goal of generating the focal group's expected test scores using the reference group's item parameters is to simulate test scores that the focal group might produce if the items were operating as they do in the reference sample. If the two sets of expected scores are the same, then the items function similarly in both samples. Differences in expected scores indicate potential differential item functioning and, therefore, measurement invariance.

Two of the indices display differences between the groups' IRT parameters as differences in expected scores between the two groups (STDS and UETSDS). A third generates a standardized difference between the two groups' expected scores which can then be interpreted using the guidelines for Cohen's *d* (ETSSD). These and other indices are explained in Meade (2010).

All three effect size indices can be explained using some common terminology. All three make reference to *expected test scores*. In this context, the term *test scores* refers to summed scores for sets of items. The indices also all share the commonality of comparing test scores generated with one sample's IRT parameters to test scores generated with another sample's IRT parameters. The item parameters include only the

discrimination and threshold values. The focal group participants' estimated theta levels will remain the same in both sets of expected test score calculations. Specific vocabulary refers to the two samples. The term *focal group* refers to the sample of interest. The term *reference group* refers to the comparison sample. For the sake of convenience, the focal group in this study is the 9-year-old group and the reference sample is the 12-year-old group.

Test scores require that at some point individual item scores are summed, and there are multiple ways to calculate test scores and differences between expected test scores. Each index proposed for use in this study calculates differences between expected test scores in a slightly different way, and therefore has unique properties. The calculation of each index and that index's unique properties are discussed below.

Signed test difference in the sample - STDS. The first index presented is called the signed test difference in the sample (STDS; Meade, 2010). The procedures used to calculate the STDS result in the cancellation of differential functioning across both items and individuals, a statement whose meaning will be explained in more detail as the index calculations are explained.

The STDS is in the metric of the original item and can therefore be interpreted as differences between estimated scores in units of the original response option set. For instance, if an STDS for a 4-option item returned a value of 1.7, it could be said that the difference between the two sets of estimated scores was 1.7 units of difference out of a possible 4 units. Although the STDS is in the metric of the original item, it produces expected scores as though the original ordinal scale was actually an interval scale. Again using a 4-point example with possible responses including the integers 1, 2, 3, and 4, the

STDS will return a value between 1 and 4 that may or may not be one of the whole numbers that originally comprised the scale.

The STDS is computed in several steps. First, the expected item scores for the focal group are generated using the focal group's item parameters:

$$ES_{s(\hat{\theta})i} = \sum_{k=1}^{m} P_{ik}(\hat{\theta})X_{ik} \quad [12]$$

According to the above formula, the expected score (ES) on an item (*i*) for a person (*s*) at a particular level of theta [s($\hat{\theta}$)*i*] is equal to the sum of the probability of responses at that estimated theta level [$P_{ik}(\hat{\theta})$] for each of the available response categories (X_k, which number from $k = 1$ to m). Since the probabilities themselves are a function of the item parameters (see equation 4, above), the estimated scores are a function of individuals' estimated theta levels and the item's parameters.

Next, the expected item scores of that same group are generated using the item parameter estimates for the reference group (but retaining the focal group members' original estimated theta levels). In the case of this study, then, the expected score on an item will be generated for each 9-year-old (i.e., the focal group) twice: once using the 9-year-old group's item parameters, and once using the 12-year-olds' (i.e., the reference group's) item parameters.

In the next step of the STDS, the difference is taken between the two sets of expected test scores. These differences are then averaged across the number of participants in the focal group, resulting in an item-level index called the signed item difference in the sample (SIDS):

$$SIDS_i = \frac{\sum_{s=1}^{N}[ES_{(si|\hat{\theta},\gamma F)} - ES_{(si|\hat{\theta},\gamma R)}]}{N} \quad [13]$$

According to the formula above, the SIDS for an item (*i*) is the average across all focal group participants (*N*) of the difference between each focal group participant's estimated item score using the focal group parameters [$ES_{(si|\hat{\theta},\gamma F)}$] and each focal group participant's estimated item score using the reference group parameters [$ES_{(si|\hat{\theta},\gamma R)}$].

The process of averaging differences in item scores allows for differential item functioning to be canceled across participants. For instance, if there were only four participants, the differences between the two estimated scores for an item could be -1, -2, 2, and 1. In this case, the SIDS value for that item would be 0. Again, it is unlikely that difference scores will be in the form of whole integers. It is more likely that difference scores will fall at smaller intervals between the original scale's range. As an example, one set of expected item scores might be 1.1, 1.7, 2.5, and 0.9. The other set of expected scores might be 1.4, 1.1, 0.28, and 0.9. The result would be difference scores of 0.3, -0.6, 0.3, and 0. The SIDS for this item would be 0. For the sake of simplicity, future examples will use be in the form of whole numbers rather than in decimal form.

In the second step, the SIDS scores for an item set (*i*) are summed, which results in the STDS statistic for the test overall:

$$STDS = \sum SIDS_i \qquad [14]$$

In this step, the STDS allows for positive SIDS scores to be offset by negative SIDS scores, a process which this time cancels differential functioning across items. As a simplified example, a measure may have four items that produce the following four SIDS scores: -1, -2, 2, and 1. Each of those 4 SIDS scores have already cancelled differential functioning across participants. When they are added together, the resulting STDS score would be 0, thereby also cancelling differential functioning across items

In summary, The STDS has the benefit of being sensitive to group differences that are unidirectional across both participants and items, and it cancels differential functioning across participants and items. The STDS provides a useful point of reference for the next index, the UETSDS, described below.

Unsigned expected test score difference in the sample - UETSDS. The UETSDS is the unsigned expected test score difference in the sample, and it allows for cancellation of differential functioning across items but not participants. Like the STDS, it returns a value in the metric of the item's number of response options which is treated intervally rather than ordinally. Also like the STDS, expected scores are generated for the focal group using both the focal group's item parameters and the reference group's item parameters. This time, the expected scores for each item (numbering from $i = 1$ to j) are summed into a person's (s) expected test score (ETS) before any differences between expected scores are taken:

$$ETS_i = \sum_{i=1}^{j} ES_{si} \quad [15]$$

This step allows for cancellation of differences across items to occur. Returning again to a 4 item example, if a test consisted of 4 items, a person's expected item scores using the focal group parameters might be 1, 1, 4, and 4. Using the reference group parameters, the persons' expected item scores might be 4, 4, 1, and 1. Although the pattern of responses differs across the items, both scores result in an expected test score of 10.

The absolute value is then taken of the difference between the focal group's ETS values using the focal group parameters and the ETS values using the reference group parameters. These absolute values are summed and averaged over the number of participants in the focal group, which results in the UETSDS:

$$UETSDS = \frac{\sum_{s=1}^{N}[\left|ES_{(si|\hat{\theta},\gamma F)} - ES_{(si|\hat{\theta},\gamma R)}\right|}{N} \quad [16]$$

According to the formula above, the SIDS for an item (*i*) is the average across all focal group participants (*N*) of the absolute value difference between each focal group participant's estimated item score using the focal group parameters $[ES_{(si|\hat{\theta},\gamma F)}]$ and each focal group participant's estimated item score using the reference group parameters $[ES_{(si|\hat{\theta},\gamma R)}]$.

Because the absolute values are averaged, differences in the signs of ETS values across persons are not allowed to cancel each other (which is allowed in the STDS). For instance, if four people yield raw ETS difference values of -1, -2, 2, and 1, their average will be 1.5 (rather than the 0 it would have been if the absolute values were not taken).

Expected test score standardized difference - ETSSD. The final index to be presented is the expected test score standardized difference (ETSSD). Like the STDS, the ETSSD allows for cancellation of differential functioning across both items and participants. The ETSSD is unique among the indices presented here in that it is a standardized difference between the focal group's expected test scores based on the focal group's IRT parameters and the focal group's expected test scores based on the reference group's parameters:

$$ETSSD = \frac{\overline{ETS}_{(\gamma F)} - \overline{ETS}_{(\gamma R)}}{SD_{TestPooled}} \quad [17]$$

In this procedure, the expected test scores based on each group of parameters are averaged before being divided by the pooled standard deviation of the two groups of ETS values. The resultant ETSSD allows for cancellation of differential functioning across both items and participants. It has the benefit of being interpretable using the guidelines for Cohen's *d*.

Summary of effect size indices. Three effect size indices were calculated: STDS, UETSDS, and ETSSD.

The STDS provides the difference between (a) the focal group's expected test scores when the focal group's parameters are used and (b) the focal group's test scores when the reference group's parameters are used. The STDS averages test scores across participants and items, so any non-uniform differences across participants or items are cancelled out by the averaging process.

The UETDS provides the same comparison as the STDS, except that it makes use of absolute differences across participants. As a result, non-uniform differences are cancelled across items but not participants.

If the STDS and UETSDS are different, then the items are functioning differentially across levels of the latent trait. Much like the t-tests on the information curves, the finding that curve differences appear across latent trait levels only when absolute values are tested strongly imply that thresholds are at least partially accountable for differences in participants' responses. Thus, the comparison between these two indices can be considered an effect size estimate of the threshold measurement invariance tests.

Like the STDS, the ETSSD allows cancellation of non-uniform differences across items and participants. However, it returns a value in a standardized metric and can be interpreted using the same guidelines as Cohen's *d*.

Step 2: Differences in the information gathered by retrospective frequency structures within age groups. The purpose of this component is to examine differences in the information gathered across the matched item subsets (i.e., CDI and BDI-Y

structure items; and the CDI and BDI-Y frequency items), but within the 9- and 12-year-old age groups.

The procedures in this step are largely those used in Research Question 1. For each age group, an IRT model that included all items in the study was generated. As before, a confirmatory factor model was generated in Mplus using the WLSMV estimator and delta parameterization. The same sets of correlated item errors were included in the model, and the factor variances were set to one. The CFA parameters were then transformed into IRT parameters, a procedure which allowed 61 information levels to be generated and then plotted. Differences in raw and absolute information values were reviewed using verbal descriptions of the information curves, and differences in discrimination parameters were tested using matched-sample *t*-tests.

Summary of Research Question 2 analyses. The analyses for Research Question 2 included two steps. The first was an investigation of measurement invariance across age groups through confirmatory factor analysis methods and IRT effect size indices. A separate set of analyses was conducted for each of the two item sets. The second step was an examination of the differences within age groups between information curves for sets of matched item pairs. This step mimicked the procedures used in Research Question 1, except that the procedures were conducted on the 9- and 12-year-old groups separately.

Results

Descriptive Statistics

Response frequencies for the BDI and CDI items that were used in the analyses for this study are presented in Tables 2 and 3. Items are presented in their cross-measure matched pairs. CDI categories were reverse coded as necessary to match the order of the

BDI-Y response categories. The BDI-Y categories are numbered from lowest to highest. All items had at least one missing data point. The BDI-Y items had more missing data than the CDI items.

Table 2 displays the response frequencies for the items matched for content and the use of frequency structures. According to that table, every CDI item had a response rate of at least 50% in one response category, which was always the first response category. In contrast, only two of the BDI-Y items (18 and 17) had a 50% or greater response rate in one response category, and in both cases it was the second response category. The remaining items' response rates did not reach 50% until response rates for the first two categories were totaled. Based on this disbursed BDI-Y response pattern, it would not be surprising if the associated information curve for this item set was found to be fairly flat and wide—the lack of a clustered response pattern means that the item probably does not discriminate well among respondents and probably has a wide range of threshold values. In contrast, the CDI information curve for this item set is expected to have a more defined curve shape because its items have more of a clustered response pattern.

Table 3 displays the response frequencies for the set of items matched for content only. In this item set, all the CDI items again had at least a 50% response rate in the lowest response category. This time, all but one BDI-Y item (8) also had at a response rate of at least 50% in the lowest response category. Because both measures' items response frequencies are clustered in one category and are clustered in the lowest category, the information curves for these measures are expected to look similar to each other.

Table 2. Response frequencies – items matched for content and frequency

CDI				BDI-Y			
Item		Percentage	n	Item		Percentage	n
1	1	78.17	1296	18	1	16.94	279
	2	19.54	324		2	61.45	1012
	3	2.29	38		3	16.27	268
		100.00	*1658*		4	5.34	88
						100.00	*1647*
10	1	77.32	1282	17	1	21.37	353
	2	16.16	268		2	56.84	939
	3	6.51	108		3	15.25	252
		99.99	*1658*		4	6.54	108
						100.00	*1652*
12	1	82.07	1359	16	1	34.75	573
	2	16.67	276		2	47.48	783
	3	1.27	21		3	14.55	240
		100.01	*1656*		4	3.21	53
						99.99	*1649*
16	1	57.66	952	5	1	28.84	475
	2	30.04	496		2	48.76	803
	3	12.30	203		3	13.66	225
		100.00	*1651*		4	8.74	144
						100.00	*1647*
20	1	56.61	938	8	1	40.84	673
	2	38.02	630		2	42.42	699
	3	5.37	89		3	11.65	192
		100.00	*1657*		4	5.10	84
						100.01	*1648*

Note: Totals are presented in italics. Some percentages are inaccurate due to rounding error.

Table 3. Response frequencies – Items Matched for Content Only

CDI				BDI-Y			
Item		Percentage	n	Item		Percentage	N
2	1	56.25	932	20	1	59.41	985
	2	38.93	645		2	29.92	496
	3	4.83	80		3	6.76	112
		100.01	*1657*		4	3.92	65
						100.01	*1658*
7	1	76.51	1267	15	1	65.05	1072
	2	17.93	297		2	25.18	415
	3	5.56	92		3	6.55	108
		100.00	*1656*		4	3.22	53
						100.00	*1648*
8	1	73.67	1220	7	1	42.97	709
	2	22.16	367		2	45.03	743
	3	4.17	69		3	7.94	131
		100.00	*1656*		4	4.06	67
						100.00	*1650*
9	1	69.02	1145	4	1	74.49	1232
	2	28.99	481		2	19.17	317
	3	1.99	33		3	4.41	73
		100.00	*1659*		4	1.93	32
						100.00	*1654*
25	1	81.07	1345	6	1	62.85	1037
	2	16.27	270		2	28.00	462
	3	2.65	44		3	6.06	100
		99.99	*1659*		4	3.09	51
						100.00	*1650*

Note: Totals are presented in italics. Some percentages are inaccurate due to rounding error.

Test of IRT Assumptions and Final Baseline Model

As outlined in the *Method* section, unidimensionality was tested using confirmatory factor analysis. The initial model tested was based on data from all participants (i.e., those aged 9 through 12 years). Also as discussed in the *Method* section, the model included one latent factor whose variance was set to 1, and the errors of the matched item pairs were allowed to covary to reduce parameter misestimation due to item non-independence. The statistical results from that model indicated poor fit: $\chi^2(160, N = 1659) = 1306.71$, $p < .01$. Of the fit indices, the RMSEA also indicated poor fit: RMSEA = .07, CFI = .96.

Based on Yen's (1984, 1993) recommendations, two of the included residual covariance pairs did not show evidence of non-independence: BDI-Y18/CDI1 (retrospective frequency pair about feeling sad) and BDI-Y20/CDI2 (content-only pair about thinking life will be bad). Those correlations were therefore removed from the model.

The *modindices* output was then examined for potential causes of model misfit. The *modindices* output lists possible model modifications and their projected effects of chi-square fit statistics. Based on that output and evidence that they met Yen's (1984, 1993) guidelines when entered into the model, two pairs of error correlations were added: BDI-Y17/BDI-Y18 (items about feeling sad and feeling like crying) and CDI1/CDI10 (the CDI items about feeling sad and feeling like crying).Although the fit statistic showed poor fit, the fit indices did show good model fit given the criteria chosen in this study: χ^2 $(160, N = 1659) = 821.31$, $p < .01$; RMSEA = .05; CFI = .95. Based on the fit guidelines discussed above, this model was retained for use in the remaining analyses. The final baseline model that used for all subsequent IRT analyses was:

- One latent factor, variance set to 1 (because of model assumptions)
- Covaried item residuals for all item pairs being examined in this study, with the exception of BDI-Y18/CDI1 and BDI-Y20/CDI2
- Covaried item residuals for the two additional item pairs of BDI-Y17/BDI-Y18 and CDI1/CDI10

Research Question 1.1: Visual Analyses of Information Curves

Model parameter estimates. The Mplus model parameter estimates were transformed into IRT parameter estimates using a SAS program, which is included in Appendix B. The SAS program was also used to generate information values. The BDI-Y and CDI CFA and IRT parameter estimates are listed in Tables 4 and 5. In each table, the items matched for content and frequency are listed first, followed by the items matched for content only. The items are listed in the tables so that item pairs are presented in the same order, e.g., the first item in the CDI table is the pair to the first item in the BDI-Y table. The items were also given letters indicating the item pair to which they belong.

Table 4. CDI CFA and IRT parameter estimates

Item number	Item pair	CFA Estimates			IRT Estimates		
		Loading	Tau 1	Tau 2	Discrimination	Threshold 1	Threshold 2
Items matched for content and frequency							
CDI 1	A	0.75	0.78	2.00	1.15	1.03	2.65
CDI 10	B	0.71	0.75	1.51	1.01	1.05	2.13
CDI 12	C	0.51	0.92	2.24	0.60	1.79	4.37
CDI16	D	0.58	0.19	1.16	0.70	0.34	2.02
CDI 20	E	0.77	0.17	1.61	1.20	0.22	2.09
Items matched for content only							
CDI 2	F	0.67	0.16	1.66	0.90	0.23	2.48
CDI 7	G	0.81	0.72	1.59	1.40	0.89	1.96
CDI 8	H	0.60	0.63	1.73	0.75	1.06	2.89
CDI 9	I	0.69	0.50	2.06	0.96	0.72	2.97
CDI 25	J	0.70	0.88	1.94	0.97	1.27	2.78

Table 5. BDI-Y CFA and IRT parameter estimates

Item number	Item pair	CFA Estimates				IRT Estimates			
		Loading	Tau 1	Tau 2	Tau 3	Discrim	Threshold 1	Threshold 2	Threshold 3
Items matched for content and frequency									
BDI-Y 18	A	0.71	-0.96	0.79	1.61	1.01	-1.35	1.11	2.27
BDI-Y 17	B	0.68	-0.79	0.78	1.51	0.93	-1.17	1.15	2.22
BDI-Y 16	C	0.52	-0.39	0.92	1.85	0.61	-0.75	1.77	3.55
BDI-Y 5	D	0.44	-0.56	0.76	1.36	0.49	-1.26	1.72	3.07
BDI-Y 8	E	0.68	-0.23	0.96	1.64	0.94	-0.34	1.41	2.40
Items matched for content only									
BDI-Y 20	F	0.80	0.24	1.24	1.76	1.32	0.30	1.56	2.11
BDI-Y 15	G	0.84	0.39	1.30	1.85	1.55	0.46	1.54	2.20
BDI-Y 7	H	0.69	-0.18	1.18	1.74	0.96	-0.26	1.70	2.52
BDI-Y 4	I	0.80	0.66	1.53	2.07	1.32	0.83	1.92	2.59
BDI-Y 6	J	0.73	0.33	1.33	1.87	1.07	0.45	1.82	1.56

Note: Discrim *refers to the discrimination parameter estimate.*

Table 6 presents mean parameter estimates for Tables 4 and 5. The mean BDI-Y discrimination parameter estimates are higher than the mean CDI estimates in the set of items matched for content and the use of frequency structures, while the CDI discrimination estimates are higher than the BDI-Y estimates in the set of items matched for content only. In both sets, the BDI-Y items cover a larger range of the latent trait, meaning that it collects information across more levels of the latent trait continuum than the CDI.

Table 6. Mean IRT parameter estimates for BDI-Y and CDI items in both item sets

	BDI-Y	CDI
Content and frequency items		
Discriminations	0.80	0.93
Threshold ranges	3.39	1.75
Content only items		
Discrimination	1.24	1.00
Threshold ranges	1.84	1.78

Information curves. Information values for each item were also generated using the SAS program included in Appendix B. Information values were generated at 0.10 intervals from -3 to 3 on the latent trait scale, generating a total of 61 information values for each item. The information values at each 0.10 interval point along the latent trait level were summed for each item set. The plots in Figures 6 and 7 are the plots of those summed information levels.

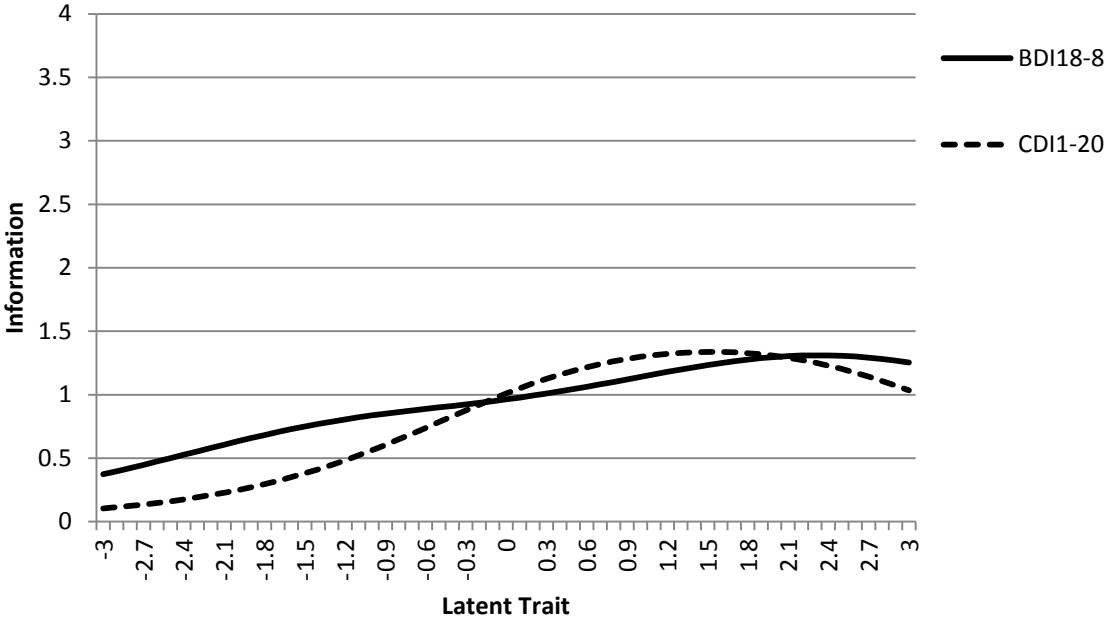

Figure 6. Information plot of items matched for **content and frequency**.

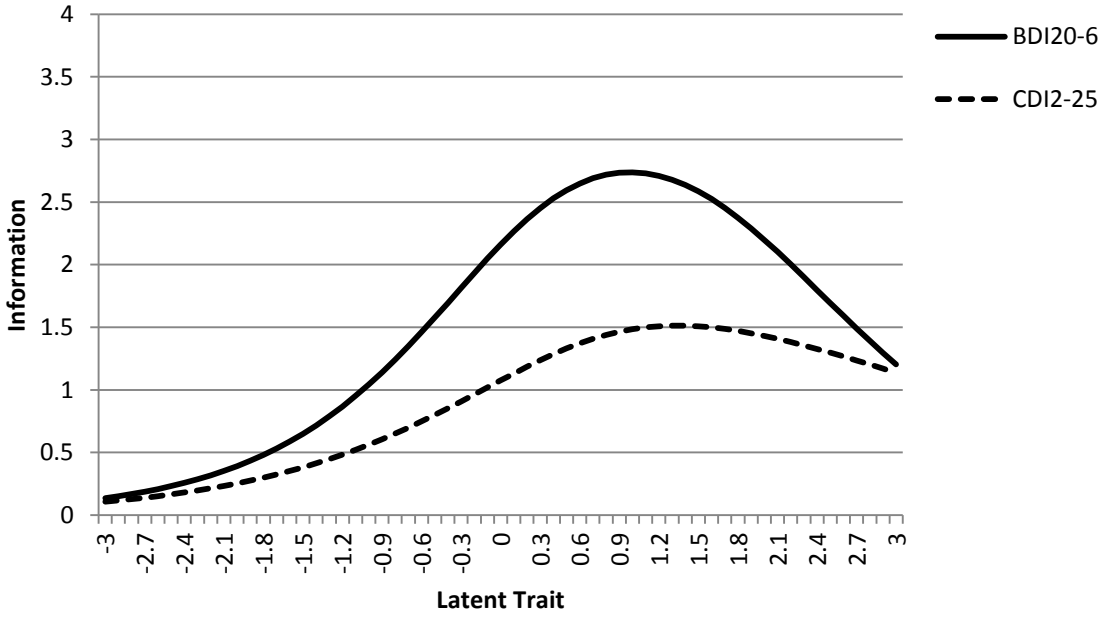

Figure 7. Information plot of items matched for **content only**.

Descriptive statistics for the information scores and the information difference scores are presented in Tables 7 and 8. The information from these tables is used in conjunction with Figures 6 and 7 to discuss the patterns of information gathered by the item sets. Those discussions are presented in the three sections that follow the tables.

Table 7. Descriptive statistics for information scores by item set and measure across 61 points on the latent trait continuum

	Minimum value	Maximum value	Theta level of max info*	Median	Mean	Standard deviation
Content and frequency items						
BDI-Y	0.29	1.31	2.3	0.96	0.95	0.29
CDI	0.10	1.34	1.6	1.01	0.83	0.45
Content only items						
BDI-Y	0.13	2.74	1.0	1.60	1.53	0.90
CDI	0.11	1.51	1.3	1.08	0.91	0.51

**Note. Refers to the level of theta at which the information curve peaks.*

Table 8. Descriptive statistics for information difference scores across 61 points on the latent trait continuum

	Minimum value	Maximum value	Median	Mean	Standard deviation
Frequency & content items	-0.16	0.39	0.12	0.12	0.20
Frequency & content – absolute values	0.01	0.39	0.16	0.19	0.12
Content only item	0.03	1.28	0.59	0.63	0.45

Visual analysis of curves for items matched for content and frequency.

According to Figure 6, the CDI information curve appears to peak at approximately the same information level as the BDI-Y curve, and to peak at a lower level of the latent trait

continuum. The curves appear very similar to each other, however, and the curves alone do not make it clear which measure gathers more information. Further, neither curve seems very steep, implying that neither set is associated with a large discrimination parameter estimate. The data in Table 7 confirm the visual analyses, with the CDI and BDI-Y showing similar maximum information values (although the CDI maximum information value is slightly higher), and the CDI showing a maximum information value at a lower level of the latent train continuum than the BDI-Y. Despite having a slightly higher maximum information value, the CDI's mean information value is lower than the BDI-Y's mean information value, indicating that the BDI-Y gathers more information across all levels of the latent trait continuum while the CDI gathers more information at a singular trait level.

Visual analysis of curves for items matched for content only. According to Figure 7, the BDI-Y information curve appears to gather quite a bit more information than the CDI information curve and to peak at a slightly lower level of the latent trait continuum. The BDI-Y curve also shows a well-defined peak while the CDI again shows a somewhat broad, flat peak, implying that the BDI-Y information curve is associated with a larger discrimination parameter estimate than the CDI information curve. A review of the information values in Table 7 confirms these visual interpretations, with the BDI-Y showing a higher maximum and mean information value than the CDI, and lower level of theta at which the information curve peaks.

Comparisons between Figures 6 and 7. The CDI items show similar levels of maximum information gathered in both plots. In contrast, the BDI-Y items show very different levels of information across the two plots. Consistent with this observation, Table 8 shows that the mean information difference score is larger for the set of items

matched for content only than the set of items matched for content and frequency. Table 8 also shows negative information difference scores in the set of items matched for content and frequency. Negative information difference scores can deflate mean difference scores, and so absolute difference scores were also calculated. When the absolute difference scores for the set of items matched for content and frequency are considered, the difference scores for that set of items increase somewhat, but are still smaller than the difference scores for the set of items matched for content only. The level of the information curve at which the information curves peak is higher for both measures in the set of items matched for content and frequency than in the set of items matched for content only.

Research Question 1.2: Statistical Tests of Differences between Item Discrimination Parameters

The item discrimination parameter values in both item sets were analyzed using matched-sample tests to determine whether the measures' discrimination values differed statistically. Each item set consisted of 5 discrimination value pairs—one pair for each item pair in the item sets. Given the small number of discrimination parameter pairs, statistical power was expected to be low. Construct reliability coefficients were also calculated to assist in the interpretation of differences in discrimination values.

Matched-sample statistical tests. Because of the few numbers of item pairs in each test, the difference scores for the pairs were first screened for normality using a Shapiro-Wilk test. Results were non-significant in both cases: for the set matched for content and frequency, $W(N = 5) = .99$, $p = .93$; for the set matched for content only, $W(N = 61) = .92$, $p = .52$.

Because the Shapiro-Wilk test results were non-significant, matched-sample *t*-tests were used to compare the discrimination values. In the item set matched for content and frequency, the differences between the discrimination values was significant, $t(4) = -2.82$, $p = .05$. Cohen's *d* was calculated to be 1.26, which is a large effect size (Cohen, 1988). In the item set matched for content only, differences between the discrimination values were again significant, $t(4) = 4.17$, $p = .01$. Cohen's *d* was again large, 1.87. A review of the item parameters (Tables 4 - 6) showed that the discrimination parameters in the item set matched for content and frequency were larger for the CDI; in the set matched for content only, the BDI-Y discrimination parameters were larger.

Construct reliability calculations. Construct reliabilities were calculated using squared standardized CFA item loadings as described in Equation 8 (Fornell & Larcker, 1981). The resulting reliability estimates are presented in Table 9. The difference between the measures' reliability values for the set matched for content and frequency was .05. The difference in reliability estimates for the set matched for content was .06.

Table 9. Construct reliability values – items matched for content and frequency

	Content & Frequency	Content Only
BDI	.75	.88
CDI	.80	.82

Research Question 1 Results Summary

In the set of items matched for content and frequency, the BDI-Y response frequencies were distributed across the first two response options, while the CDI

response frequencies were clustered at the lowest response option. The BDI-Y therefore had lower discrimination parameter estimates and a wider range of threshold parameter estimates than the CDI. The BDI-Y information curve was relatively flat and peaked at a higher level on the latent trait continuum than the CDI information curve. Although the BDI-Y information curve reached a slightly lower maximum information value than the CDI information curve, the mean BDI-Y information value was higher than the mean CDI information value. In contrast to the BDI-Y curve's flat shape, the CDI information curve was well-defined.

In the set of items matched for content only, the response frequencies for the BDI-Y and CDI were clustered in the lowest response option. The two measures were associated with similar discrimination and threshold parameter estimates, although the BDI-Y discrimination parameter estimates were somewhat higher and the threshold estimate range was somewhat larger. The information curves for both measures were well-defined and peaked at similar levels of the latent trait continuum. Although the BDI-Y information curve reached higher maximum and mean information values than the CDI information curve, it is not clear if that is because of the BDI-Y parameter estimates or because the BDI-Y has more thresholds.

When comparisons were made across item sets, the CDI information curves reached similar maximum information levels and peaked at similar levels of the latent trait continuum in both item sets. In contrast, the BDI-Y information values and curve peaks varied across the item sets. Consistent with these observations, reviews of information difference scores and item parameter estimates showed that the information difference scores were higher in the set of items matched for content only than in the set of items matched for content and frequency, while the threshold values at which the

information curves peaked were somewhat higher in the set matched for content and frequency. The CDI discrimination and threshold parameter estimates were similar across the item sets, while the BDI-Y parameter estimates varied quite a bit.

The matched-sample statistical tests showed that the measures' discrimination parameters in both item sets were significantly different with large effect sizes. Despite large *t*-test effect sizes, the differences in construct reliability associated with the tests of the discrimination values were relatively small, implying that the statistically significant findings may have limited practical applications.

Research Question 2

Response frequencies for the 9- and 12-year-old groups are presented in Tables 10 and 11. Items are presented in their cross-measure matched pairs. The BDI-Y categories are numbered from lowest frequency to highest. CDI categories were reverse coded as necessary to match the order of the BDI-Y response categories. All items had at least one missing data point. The BDI-Y items had more missing data than the CDI items.

Table 10 displays the response frequencies for the items matched for content and the use of frequency structures. According to that table, the 12-year-old group's CDI response frequency distributions were somewhat more distributed across the first two response options than the 9-year-old group's response frequency distributions, implying that the 12-year-old group is likely to have a larger mean threshold parameter estimate range than the 9-year-old group. In the set of BDI-Y items, the 9-year-old group's response frequency distributions were more evenly spread across all 4 response options while the 12-year-old group typically had higher frequencies for the first response, implying that the 9-year-old group is likely to have a higher mean threshold estimate range.

Table 11 displays the response frequencies for the set of items matched for content only. In this item set, a pattern of differences between the age groups' frequency distributions for the CDI was not clear; the 9-year-old group seemed to have more clustered response patterns for some items (7, 8, 9), the 12-year-old group had a more clustered response pattern for a fourth item (2), and the age groups' frequency distributions were nearly identical for the fifth item (25). Predictions about which group will have a larger mean threshold parameter estimate range are therefore difficult to make. On the BDI-Y, with one exception (item 7), the 9-year-old group's responses were more evenly distributed across the response options than the 12-year-old group's responses. The 9-year-old group's mean BDI-Y threshold estimate range can therefore be expected to be larger than the 12-year-old group's mean threshold estimate range.

Table 10. Response frequency percentages– items matched for content and frequency

CDI				BDI-Y			
Item		9 y/o	12 y/o	Item		9 y/o	12 y/o
1	1	77.59	75.41	18	1	12.97	16.94
	2	20.05	22.40		2	63.32	61.20
	3	2.36	2.19		3	17.69	16.67
		100	*100*		4	6.60	4.92
						100.58	*99.73*
10	1	78.54	71.86	17	1	14.86	24.59
	2	13.68	22.95		2	61.32	52.19
	3	7.78	5.19		3	14.86	17.21
		100	*100*		4	8.73	5.64
						99.77	*99.63*
12	1	80.90	80.87	16	1	35.85	36.07
	2	17.45	17.76		2	42.92	46.99
	3	1.65	1.09		3	16.27	13.11
		100	*99.72*		4	4.25	3.83
						99.29	*100*
16	1	51.65	60.66	5	1	26.18	29.51
	2	30.90	29.23		2	46.70	51.64
	3	17.45	9.84		3	13.68	10.66
		100	*99.73*		4	12.74	7.10
						99.30	*98.91*
20	1	53.30	57.10	8	1	35.38	41.26
	2	39.15	39.62		2	44.81	41.26
	3	7.31	3.28		3	14.15	11.75
		99.76	*100*		4	5.19	4.65
						99.53	*98.92*

Note: Totals are presented in italics. Some percentages are inaccurate due to rounding error.

Table 11. Response frequency percentages – items matched for content only

CDI				BDI-Y			
Item		9 y/o	12 y/o	Item		9 y/o	12 y/o
2	1	53.77	54.10	20	1	53.07	57.65
	2	39.62	41.53		2	34.43	30.60
	3	6.60	4.37		3	9.20	7.65
		99.99	*100*		4	3.30	4.10
						100	*100*
7	1	76.18	71.04	15	1	59.91	62.02
	2	17.22	24.04		2	28.07	28.42
	3	6.60	4.37		3	6.84	7.10
		100	*99.45*		4	3.77	2.19
						98.59	*99.73*
8	1	72.88	67.76	7	1	37.74	44.26
	2	21.93	27.32		2	48.82	43.17
	3	4.95	4.64		3	9.43	8.20
		99.76	*99.72*		4	3.77	4.10
						99.76	*99.73*
9	1	65.33	63.66	4	1	76.89	65.85
	2	32.78	34.70		2	18.16	25.14
	3	1.89	1.64		3	3.07	6.83
		100	*100*		4	1.42	1.64
						99.54	*99.46*
25	1	79.72	79.78	6	1	59.20	60.11
	2	17.45	17.49		2	31.13	28.69
	3	2.83	2.73		3	6.84	7.65
		100.05	*100*		4	2.36	3.01
						99.53	*99.46*

Note: Totals are presented in italics. Some percentages are inaccurate due to rounding error.

Research Question 2, Step 1: Measurement Invariance Analyses and Invariance Effect Size Indices

The baseline CFA model used in Research Question 1 was used as the baseline model for Research Question 2. Two of the four age groups were included in the analyses—the 9-year-old group ($N = 424$) and the 12-year-old group ($N = 366$).

Overall measurement invariance tests. The baseline unconstrained model fit well according to the fit indices (RMSEA = .05, CFI = .98) although the chi-square statistic was significant [$\chi^2(330, N = 790) = 680.55, p < .01$]. The difference in model fit between the least and most constrained model was significant, $\chi^2(49, N = 790) = 121.57$, $p < .01$, indicating the presence of non-invariance between the 9- and 12-year-old groups. To identify possible sources of the non-invariance, partial measurement invariance tests were conducted. For each measure in each item set, the loading and threshold constraints across the age groups were released one set at a time while the other parameters were constrained. In total, eight partial measurement invariance tests were conducted (i.e., BDI-Y content and frequency items' loadings, BDI-Y content and frequency items' thresholds, CDI content and frequency items' loadings, CDI content and frequency items' thresholds; the same pattern was repeated for the items matched for content only). Effect size indices (i.e., STDS, UETSDS, ETSSD) were then calculated for each item set's overall measurement invariance results (i.e., as opposed to the partial measurement invariance results). Results are presented below by item set: results for items matched for frequency and content are presented first, followed by the results for the items matched for content only.

Items matched for frequency and content. The 9- and 12-year-olds' IRT parameter estimates (from the unconstrained model) are presented in Tables 12 and 13

below. Table 14, which provides the mean parameter estimates from Tables 12 and 13, shows that for the BDI-Y, the 9-year-old group's discrimination estimates are lower than the 12-year-old group's while their threshold estimates cover a wider range of the latent trait continuum. As expected, Figure 8 shows that although the information curves for both age groups are fairly flat, the curve for the 9-year-old group is lower and flatter than the curve for the 12-year-old group. For the CDI, Table 14 shows that the 12-year-old group has higher discrimination parameter estimates and a larger threshold estimate range than the 9-year-old group. Figure 9 is consistent with those estimates, displaying a higher and more peaked information curve for the 12-year-old group than the 9-year-old group.

Table 12. Unconstrained BDI-Y IRT parameter estimates for items matched for content and frequency

9-year-old group					12-year-old group				
Item	Discrim	Threshold 1	Threshold 2	Threshold 3	Item	Discrim	Threshold 1	Threshold 2	Threshold 3
18	0.85	-1.73	1.06	2.31	18	1.27	-1.22	1.00	2.10
17	0.86	-1.60	1.11	2.09	17	0.98	-0.98	1.07	2.29
16	0.49	-0.81	1.85	3.89	16	0.70	-0.62	1.67	3.09
5	0.43	-1.62	1.60	2.90	5	0.43	-1.35	2.35	3.74
8	0.71	-0.64	1.50	2.82	8	1.07	-0.29	1.33	2.29

Note: Discrim refers to the discrimination parameter estimate value.

Table 13. Unconstrained CDI IRT parameter estimates for items matched for content and frequency

9-year-old group				12-year-old group			
Item	Discrim	Threshold 1	Threshold 2	Item	Discrim	Threshold 1	Threshold 2
1	0.81	1.21	3.12	1	1.28	0.87	2.56
10	0.94	1.16	2.08	10	1.08	0.79	2.21
12	0.53	1.87	4.66	12	0.76	0.15	3.79
16	0.68	0.07	1.67	16	0.64	0.51	2.40
20	1.00	0.12	2.05	20	1.46	0.22	2.23

Note: Discrim refers to the discrimination parameter estimate value.

Table 14. Unconstrained model's mean parameter estimates, items matched for content and frequency

	BDI-Y		CDI	
	9 y/o	12 y/o	9 y/o	12 y/o
Discriminations	0.67	0.89	0.79	1.04
Threshold ranges	4.08	3.03	1.83	2.13

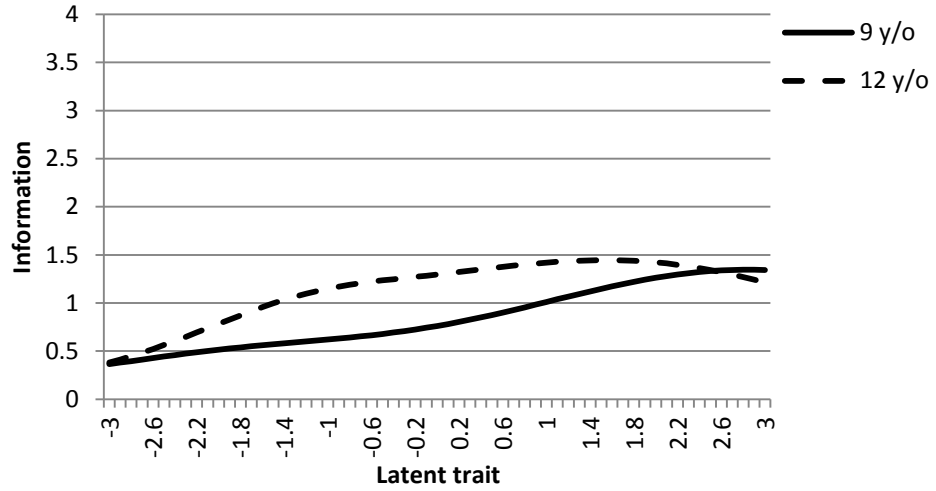

Figure 8. Unconstrained model's **BDI-Y content and frequency** items' information plots for 9 and 12 year-old groups

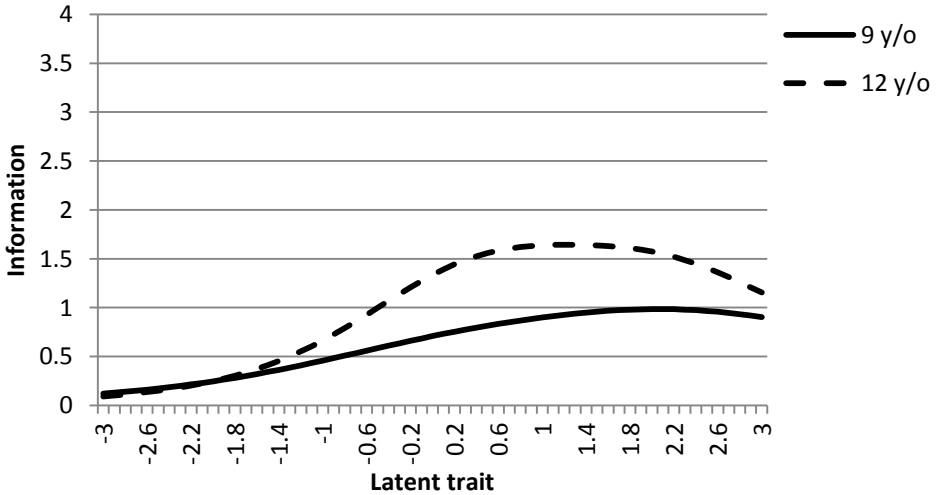

Figure 9. Unconstrained model's **CDI content and frequency** items' information plots for 9 and 12 year-old groups

Partial measurement invariance analyses. The results from the partial measurement invariance analyses for the set of items matched for frequency and content are presented in Table 15. According to those analyses, the loadings for both measures were invariant across the two age groups, and the thresholds for both measures were found to be non-invariant.

According to the information values presented in Table 16, the measures' threshold non-invariance is associated with the 9-year-old group's information curves peaking at a higher level of the latent trait continuum in both measures. Table 16 also shows that the 12-year-old group achieves a higher maximum information value than the 9-year-old group for both measures. Table 17 shows that the mean information difference score is slightly higher for the CDI than the BDI-Y, meaning that the differences between the 9-year-old and 12-year-old information curves is greater for the CDI than the BDI-Y.

The negative minimum information scores in Table 17 are an indication that the information curves cross each other, which could potentially deflate the information difference score mean. Although the curves did cross over, they did so only at the tails of the curves, making the impact of the negative information difference scores minimal.

Table 15. Change in chi-square for partial measurement invariance tests of content and frequency items

	X^2	df	N	*P*
BDI-Y loadings	10.62	5	790	.06
BDI-Y thresholds	48.37	15	790	< .01
CDI loadings	8.41	5	790	.14
CDI thresholds	27.63	10	790	< .01

Table 16. Information score descriptive statistics for items matched for content and frequency (across 61 points on the latent trait continuum) by measure and age group

	Minimum value	Maximum value	Theta level of max info*	Median	Mean	Standard deviation
BDI-Y						
9-year-olds	0.36	1.35	2.8	0.84	0.84	0.32
12-year-olds	0.38	1.44	1.5	1.27	1.15	0.31
CDI						
9-year-olds	0.12	0.98	2.1	0.71	0.63	0.31
12-year-olds	0.09	1.64	1.2	1.21	1.02	0.58

**Note. This column notes the theta level at which the maximum information value occurs.*

Table 17. Between-age-group information difference score descriptive statistics for items matched for content and frequency (across 61 points on the latent trait continuum)

	Minimum value	Maximum value	Median	Mean	Standard deviation
BDI-Y	-0.13	0.56	0.35	0.31	0.21
CDI	-0.03	0.75	0.41	0.38	0.29

Effect size indices. The measurement invariance effect size indices for the two item sets were calculated using the unconstrained IRT parameter estimates presented in Tables 12 and 13. The 9-year-old group was treated as the focal group, which means that their expected scores were calculated with both the IRT parameter estimates generated for their group and with the IRT parameter estimates generated for the 12-year-old group. The differences between those two sets of expected scores served as the foundation for the effect size indices presented in Table 18.

The smaller BDI-Y STDS (0.54) as compared to the UETSDS (0.56) score indicates that the differences between the 9-year-old group's expected scores when calculated with the 9-year-old group's parameter estimates and the 9-year-old group's

expected scores when calculated with the 12-year-old group's parameter estimates are non-uniform. This finding is consistent with the information score plots in Figure 10, and supports the measurement invariance findings that threshold differences between the groups play a role in the different levels of information gathered by the two age groups across the latent trait. According to the STDS, the 9-year-old group's scores would be an average of .54 points per item higher if their scores were estimated using the 12-year-old group's parameter estimates. The ETSSD (0.06), which is interpreted using Cohen's *d* (1988) guidelines, shows that the effect size associated with those score differences is quite small.

The STDS (-0.06) and UETSDS (0.15) values for the CDI show that the CDI's expected score differences are also non-uniform, as is reflected in Figure 11. Again, this finding supports the measurement invariance findings that threshold differences between the age groups play a role in the different amount of information gathered across the groups. According to the STDS value, the 9-year-old group's scores would be an average of .06 points per item lower if their scores were estimated using the 12-year-old group's parameter estimates. The ETSSD (-0.04), which is interpreted using Cohen's *d* (1988) guidelines, shows that the effect size associated with those score differences is quite small.

Table 18. Measurement invariance effect size indices – items matched for content and frequency

	STDS	UETSDS	ETSSD
BDI-Y	0.54	0.56	0.06
CDI	-0.06	0.15	-0.04

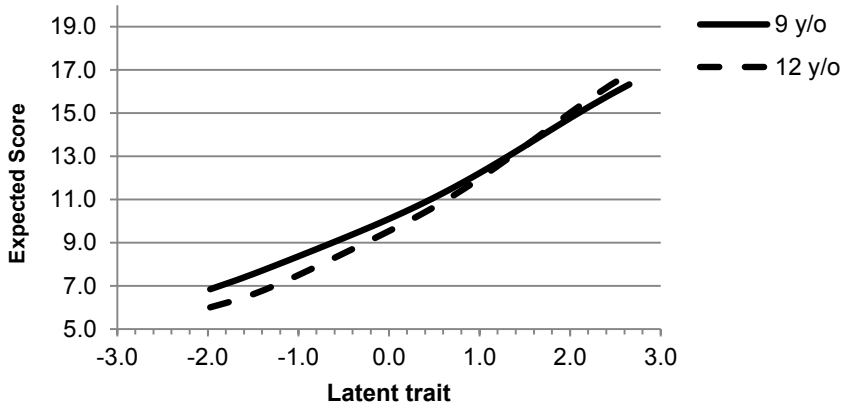

Figure 10. 9 year-old group's expected **BDI-Y** scores for the items matched for **content and frequency** using the 9 year-old and 12 year-old groups' IRT parameters

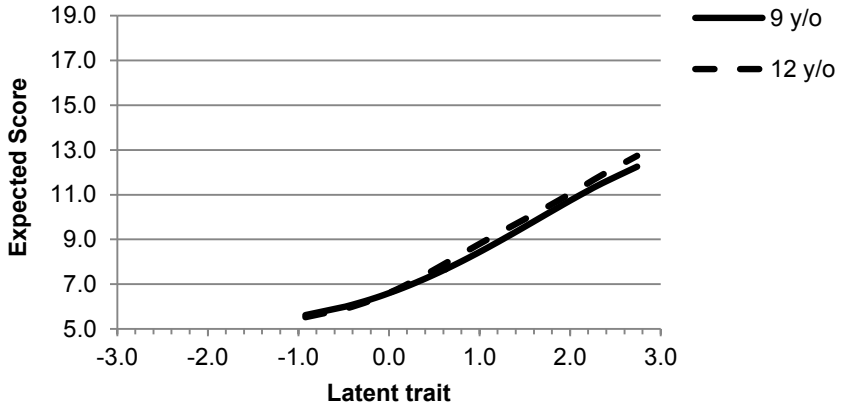

Figure 11. 9 year-old group's expected **CDI** scores for the items matched for **content and frequency** using the 9 year-old and 12 year-old groups' IRT parameters

Items matched for content only. The 9- and 12-year-olds' IRT parameter estimates from the unconstrained model are presented in Tables 19 and 20 below. Table 21, which provides the mean parameter estimates for Tables 19 and 20, shows that for both measures, the 12 -year-old group's discrimination parameter estimates are larger than the 9-year-old group's estimates, and the 9-year-old group's threshold parameter estimates cover a wider range of the latent trait continuum than the 12-year-old group's

estimates. Consistent with these estimate patterns, Figures 12 and 13 show that for both measures, the 12-year-old group's information curves are more peaked and reach a higher maximum information level than the 9-year-old group's information curves.

Table 19. Unconstrained BDI-Y IRT parameter estimates for items matched for content only

9-year-old group					12-year-old group				
Item	Discrim	Threshold 1	Threshold 2	Threshold 3	Item	Discrim	Threshold 1	Threshold 2	Threshold 3
20	0.98	0.11	1.65	2.63	20	1.59	0.23	1.40	2.06
15	1.27	0.35	1.58	2.25	15	1.87	0.35	1.50	2.29
7	0.77	-0.51	1.83	2.91	7	0.96	-0.20	1.68	2.52
4	1.06	1.03	2.22	3.01	4	1.24	0.54	1.76	2.74
6	0.91	0.36	1.97	2.95	6	1.40	0.33	1.53	2.31

Note: Discrim *refers to the discrimination estimates.*

Table 20. Unconstrained CDI IRT parameter estimates for items matched for content only

9-year-old group				12-year-old group			
Item	Discrim	Threshold 1	Threshold 2	Item	Discrim	Threshold 1	Threshold 2
2	0.67	0.17	2.17	2	1.00	0.15	2.42
7	1.21	0.92	1.95	7	1.44	0.69	2.08
8	0.50	1.34	3.70	8	0.77	0.76	2.75
9	0.69	0.70	3.68	9	1.04	0.48	2.96
25	0.81	1.32	3.03	25	1.14	1.11	2.55

Note: Discrim *refers to the discrimination estimate.*

Table 21. Unconstrained model's mean parameter estimates, items matched for content only

	BDI-Y		CDI	
	9 y/o	12 y/o	9 y/o	12 y/o
Discriminations	1.00	1.41	0.79	1.08
Threshold ranges	2.90	2.13	2.02	1.91

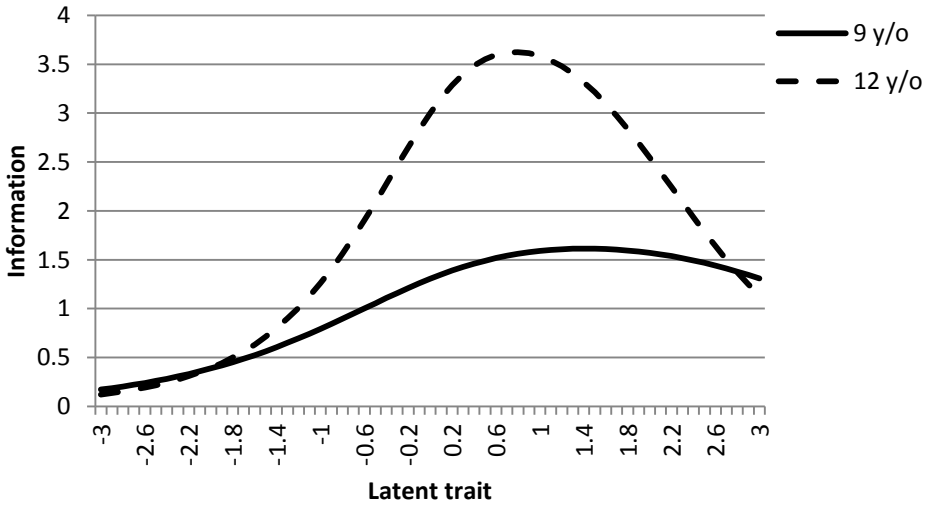

Figure 12. Unconstrained model's **BDI-Y content only** items' information plots for 9 and 12 year-old groups

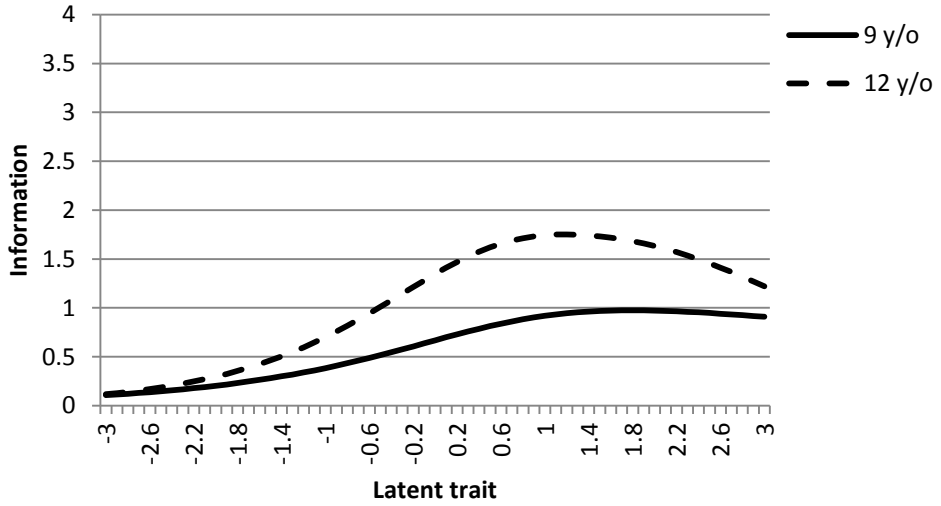

Figure 13. Unconstrained model's **CDI content only** items' information plots for 9 and 12 year-old groups

Partial measurement invariance analyses. Partial measurement analyses were conducted to determine possible sources of the overall finding of measurement non-invariance. According to the results of those analyses, which are presented in Table 22, only the CDI loadings were invariant; the CDI thresholds, BDI-Y loadings, and BDI-Y thresholds were non-invariant across the groups.

According to the information values presented in Table 23, the finding of threshold non-invariance in both measures is associated with the 9-year-old group's information curves peaking at a higher level of the latent trait than the 12-year-old group's information curves in both measures. The finding of BDI-Y loading non-invariance is associated with higher maximum and mean information values in the 12-year-old group than the 9-year-old group. Although the 12-year-old group also had higher maximum and mean information values than the 9-year-old group for the CDI, the CDI loading estimates were found to be invariant across age groups. The information difference scores presented in Table 24 show that the BDI-Y's maximum and mean information difference scores are higher than those for the CDI, further demonstrating that the differences between the 9- and 12-year-old group's loading estimates were greater for the BDI-Y than the CDI. As in the set of items matched for content and frequency, the negative minimum information score differences in Table 24 were associated only with minor curve cross-over at the curve tails, and were therefore not analyzed further.

Table 22. Change in chi-square for partial measurement invariance tests of items matched for content only

	X^2	df	N	p
BDI-Y loadings	16.30	5	790	< .01
BDI-Y thresholds	37.94	15	790	< .01
CDI loadings	10.13	5	790	.07
CDI thresholds	19.04	10	790	.04

Table 23. Information score descriptive statistics for items matched for content only (across 61 points on the latent trait continuum) by measure and age group

	Minimum value	Maximum value	Theta level of max info*	Median	Mean	Standard deviation
BDI-Y						
9-year-olds	0.17	1.61	1.4	1.30	1.07	0.51
12-year-olds	0.12	3.62	0.8	1.90	1.90	1.23
CDI						
9-year-olds	0.11	0.98	1.8	0.67	0.61	0.33
12-year-olds	0.12	1.75	1.2	1.27	1.07	0.59

**Note. This column notes the theta level at which the maximum information value occurs.*

Table 24. Information difference score descriptive statistics for items matched for content only (across 61 points on the latent trait continuum) by measure

	Minimum value	Maximum value	Median	Mean	Standard deviation
BDI-Y	-0.17	2.08	0.67	0.83	0.80
CDI	0.01	0.83	0.49	0.46	0.28

Effect size indices. The effect size indices for the measurement invariance analyses are presented in Table 25. Because the 9-year-old group's UETSDS (0.41) for the BDI-Y is larger than the associated STDS 0(.07), the difference between the 9-year-old group's expected scores when calculated with the 9-year-old group's parameter estimates is non-uniform when compared to the 9-year-old group's expected scores when calculated with the 12-year-old group's parameter estimates. This finding is consistent with the plot of expected scores presented in Figure 14 and supports the measurement invariance analyses' findings of threshold non-invariance. The STDS value indicates that the 9-year-old group's expected scores would be an average of 0.07 points higher if their

scores were calculated using the 12-year-old group's parameter estimates. Using Cohen's *d* (1988) guidelines to interpret the ETSSD (0.03), the differences between two score sets are small.

The pattern of CDI effect size indices findings is similar. The smaller CDI STDS (0.04) value and larger UETSDS (0.22) value again show that the difference between the 9-year-old groups' expected scores is non-uniform across expected score sets. This pattern of non-uniform differences is displayed in Figure 15 and is consistent with the measurement invariance analyses' findings of non-invariance for the CDI thresholds. According to the STDS, the expected scores for the 9-year-old group would be an average of .04 points per item higher if their scores were calculated using the 12-year-old group's parameter estimates. The associated ETSSD (0.03) value, which is interpreted using Cohen's *d* (1988) guidelines, indicates that the degree of these differences is quite small.

Table 25. Measurement invariance effect size indices – items matched for content only

	STDS	UETSDS	ETSSD
BDI-Y	0.07	0.41	0.03
CDI	0.04	0.22	0.03

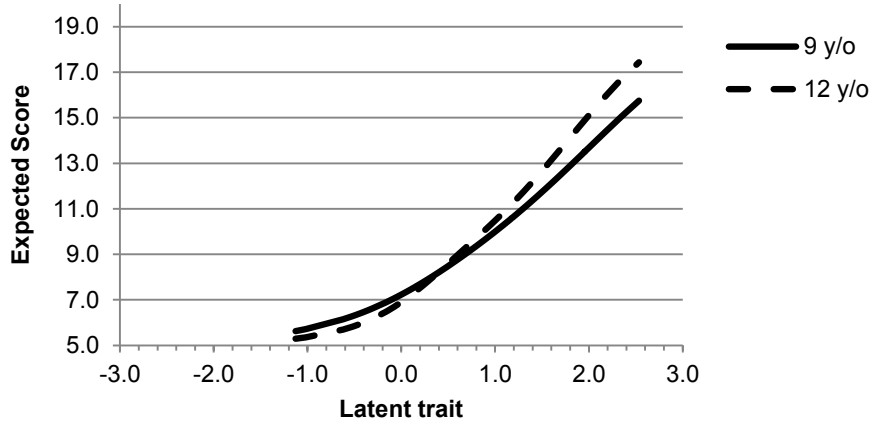

Figure 14. 9 year-old group's expected **BDI-Y** scores for the items matched for **content only** using the 9 year-old and 12 year-old groups' IRT parameters

Figure 15. 9 year-old group's expected **CDI** scores for the items matched for **content only** using the 9 year-old and 12 year-old groups' IRT parameters

Research Question 2, Step 2: Differences in the Information Gathered by Retrospective Frequency Structures within Age Groups

For both age groups, the following procedures were conducted:

- Generation of CFA model for each age group
- Generation of IRT parameters
- Generation of information curves for matched item sets

- Matched-sample statistical tests of discrimination values, supplemented by construct reliability calculations

The results are presented by step—each age group's results are presented for each step.

CFA models. The IRT models generated for each group used the same structure as that used in Research Question 1 (i.e., the pattern of covaried item errors included and excluded in Research Question 1 were retained). Although chi-square was significant, the fit indices showed that the model fit the 12-year-old group well: $\chi^2(160, N=366) = 246.74$, $p < .01$; RMSEA = .04, CFI = .99. A somewhat high RMSEA value provided evidence that the model did not fit the 9-year-old group well: $\chi^2(160, N = 424) = 409.41$, $p < .01$; RMSEA = .06, CFI = .95.

To improve the fit of the 9-year-old group's model, the modindices output was examined for potential model changes. The largest possible change was shown to be the covariation of the item errors for BDI-Y items 4 and 15 (items related to wishing to be dead, and hating oneself, respectively). After the inclusion of this covariation, RMSEA decreased somewhat, but was still approximately .06 (i.e., .055). Using the modindices output for that model, the next largest possible change was added, a covariation of the item errors for CDI items 8 and 25 (items related to guilt and feeling unloved). With this inclusion, the fit indices provided evidence of acceptable model fit although chi-square remained significant: $\chi^2(158, N=424) = 341.74$, $p < .01$; RMSEA = .05, CFI = .97. This final model was used in the remaining analyses for this sub-question.

To maintain consistency between the 9- and 12-year-old groups' models, the two covariations added to the 9-year-olds group's model were also added to the 12-year-old group's model. Although the chi-square statistic was still significant [$\chi^2(158, N=424) =$

240.49, $p < .01$], the fit indices supported a conclusion of good model fit, RMSEA = .04, CFI = .99.

IRT parameters. The Mplus model parameter estimates were transformed into IRT parameter estimates using the same SAS program and procedures used for Research Question 1. The IRT parameter estimates for the 9- and 12-year-old groups are displayed in Tables 26 and 27, and the mean parameter estimates are presented in Table 28.

Table 28 shows that within age groups, the pattern of parameter estimate differences across measures is the same as was found in Research Question 1 and is the same in both age groups: in the set of items matched for content and frequency, the CDI discrimination parameter estimate is higher than the BDI-Y estimate, and the BDI-Y threshold range covers more of the latent trait continuum than the CDI threshold range; in the set of items matched for content only, the BDI-Y discrimination parameter estimates were larger than the CDI estimates, and the BDI-Y threshold estimates covered a slightly wider range of the latent trait continuum than the CDI estimate.

When the estimates for the two age groups were compared to each other, the 12-year-old group's discrimination parameter estimates were higher than the 9-year-old group's estimates in both item sets for both measures. With one exception, the 9-year-old group's threshold estimates covered a wider range of the latent trait continuum than the 12-year-old group's estimates for both measures in both item sets.

Table 26. 9-year-old group's IRT parameter estimates

BDI-Y					CDI			
Item	Discrim	Threshold 1	Threshold 2	Threshold 3	Item	Discrim	Threshold 1	Threshold 2
Items matched for content and frequency								
18	0.85	-1.73	1.06	2.31	1	0.80	1.21	3.18
17	0.83	-1.63	1.12	2.12	10	0.90	1.18	2.12
16	0.49	-0.80	1.85	3.88	12	0.53	1.87	4.57
5	0.44	-1.57	1.55	2.82	16	0.70	0.07	1.63
8	0.70	-0.65	1.50	2.84	20	1.01	0.12	2.05
Items matched for content only								
20	1.01	0.11	1.62	2.59	2	0.68	0.17	3.53
15	1.03	0.38	1.73	2.47	7	1.18	0.93	1.97
7	0.80	-0.50	1.79	2.86	8	0.48	1.42	3.81
4	0.86	1.14	2.60	3.36	9	0.70	0.69	3.64
6	0.92	0.36	1.96	2.93	25	0.77	1.37	3.13

Note: Discrim *refers to the discrimination parameter estimates.*

Table 27. 12-year-old group's IRT parameter estimates

BDI-Y					CDI			
Item	Discrim	Threshold 1	Threshold 2	Threshold 3	Item	Discrim	Threshold 1	Threshold 2
Items matched for content and frequency								
18	1.30	-1.21	0.99	2.09	1	1.31	0.87	2.54
17	1.01	-0.96	1.05	2.25	10	1.12	0.78	2.18
16	0.71	-0.62	1.66	3.07	12	0.77	1.44	3.76
5	0.42	-1.37	2.37	3.77	16	0.63	0.51	2.41
8	1.10	-0.28	1.31	2.27	20	1.49	0.22	2.22
Items matched for content only								
20	1.57	0.23	1.41	2.06	2	0.99	0.15	2.42
15	1.81	0.35	1.51	2.30	7	1.46	0.69	2.07
7	0.94	-0.21	1.69	2.54	8	0.76	0.77	2.78
4	1.17	0.55	1.81	2.81	9	1.03	0.49	2.97
6	1.42	0.32	1.52	2.30	25	1.17	1.10	2.53

Note: Discrim *refers to the discrimination parameter estimates.*

Table 28. Mean BDI-Y and CDI IRT parameter estimates for 9- and 12-year-old groups, both item sets

	9 year-old group		12 year-old group	
	BDI-Y	CDI	BDI-Y	CDI
Content and frequency items				
Discriminations	0.66	0.79	0.91	1.06
Threshold ranges	4.07	1.82	3.58	1.86
Content only items				
Discrimination	0.92	0.76	1.38	1.08
Threshold ranges	2.54	2.45	2.15	1.91

Visual analyses of information curves. The IRT parameters from Tables 26 and 27 were used to generate information curves for each age group and both item sets. Those curves are presented in Figures 16 through 19. The visual analyses were supplemented with information and information difference score descriptive statistics presented in Tables 29 and 30.

Table 29. Information value descriptive statistics (across 61 points on the latent trait continuum), both age groups, measures, and item sets

	Minimum Value	Maximum value	Theta level of max info*	Median	Mean	Standard deviation
Content and frequency items						
9-year-old BDI-Y	0.36	1.35	2.90	0.74	0.83	0.33
9-year-old CDI	0.12	0.97	2.10	0.70	0.63	0.30
12-year-old BDI-Y	0.38	1.49	1.40	1.31	1.18	0.33
12-year-old CDI	0.09	1.71	1.20	1.26	1.05	0.60
Content only items						
9-year-old BDI-Y	0.18	1.36	2.30	1.05	0.92	0.41
9-year-old CDI	0.11	0.92	1.90	0.64	0.59	0.30
12-year-old BDI-Y	0.12	3.47	0.80	1.86	1.84	1.17
12-year-old CDI	0.12	1.78	1.20	1.27	1.07	0.60

**Note. Refers to the level of theta at which the information curve peaks.*

Table 30. Information difference score descriptive statistics (across 61 points on the latent trait continuum) by item set and age group

	Minimum value	Maximum value	Median	Mean	Standard deviation
Content and frequency items					
9-year-olds	0.04	0.44	0.21	0.20	0.11
12-year-olds	-0.24	0.63	0.02	0.13	0.33
12-year-olds, absolute values	< 0.01	0.63	0.23	0.29	0.19
Content only items					
9-year-olds	0.07	0.45	0.39	0.33	0.12
12-year-olds	-0.07	1.76	0.63	0.76	0.64

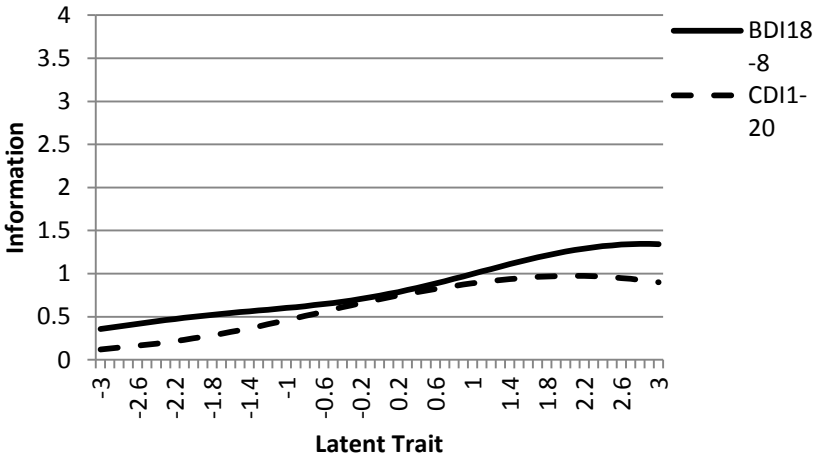

Figure 16. Information plot of **content and frequency** items in 9-year-old group.

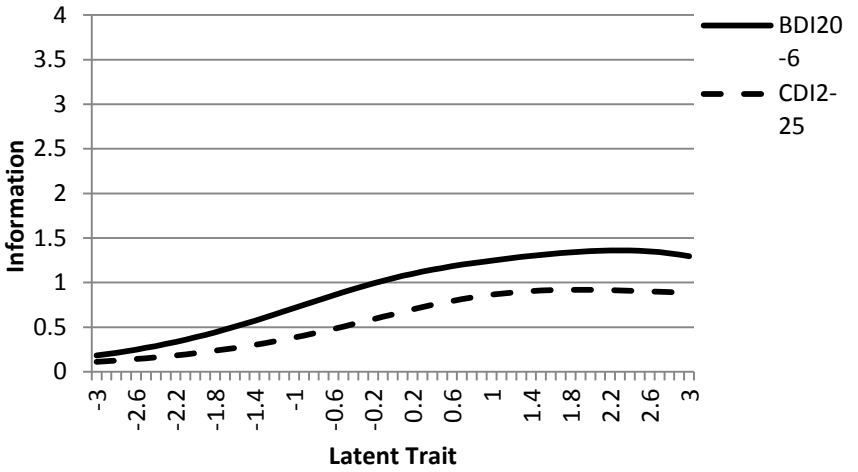

Figure 17. Information plot of **content only** items in 9-year-old group.

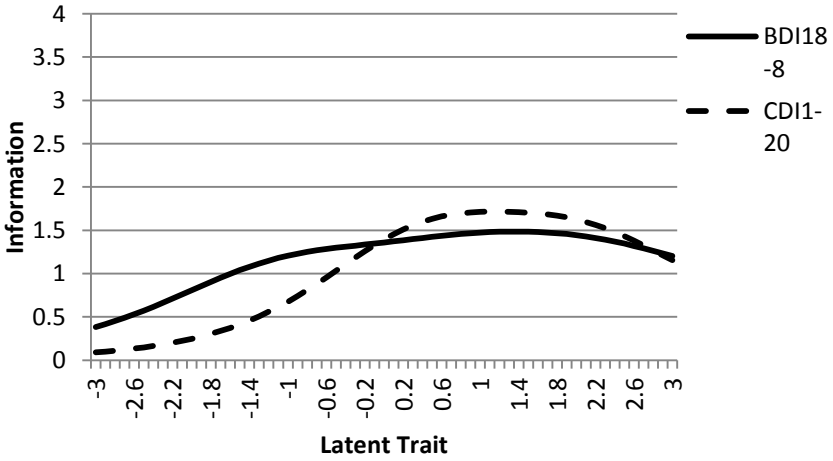

Figure 18. Information plot of **content and frequency** items in the 12-year-old group.

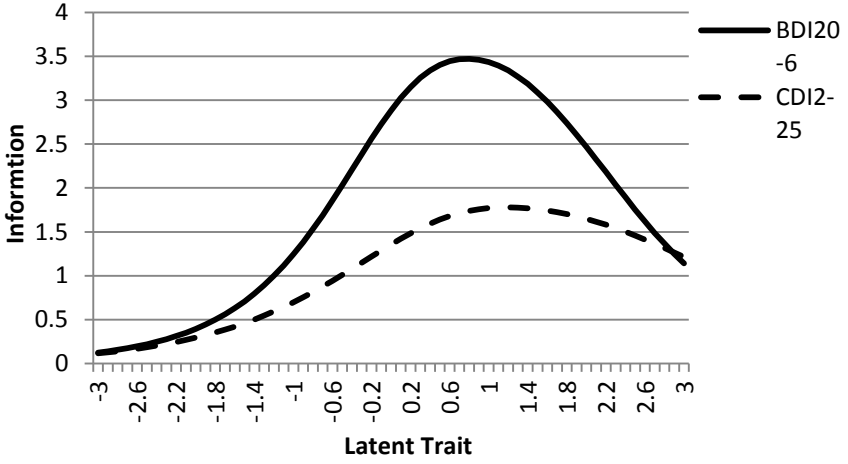

Figure 19. Information plot of **content only** items in the 12-year-old group.

As expected, the 9-year-old group's BDI-Y and CDI information curves for the set of items matched for content and frequency (Figure 16) are fairly flat, and they are flatter and lower than the 12-year-old group's information curves (Figure 18). The 12-year-old group's information curves show that the BDI-Y information curve is also relatively flat, whereas the CDI information curve is more defined and appears to reach a higher maximum information level. These visual analyses are supported by data from Table 29, which shows that the 9-year-old groups maximum information values for both measures are lower than the maximum information values for the 12-year-old group. Table 29 also shows that the 9-year-old group's information curves for both measures peak at a higher level of the latent trait continuum than the 12-year-old group's curves. The larger maximum information difference score for the 12-year-old group in Table 30 shows that the curves for the 12-year-old group differ more from each other than the curves for the 9-year-old group. This conclusion is also supported by the difference between the 12-year-old group's absolute mean information difference scores and the 9-year-old group's mean difference scores.

In the set of items matched for content only, the 9-year-old group's information curves (Figure 17) are again fairly flat, and they are again flatter and lower than the 12-year-old group's information curves (Figure 19). In contrast, the 12-year-old group's information curves both show distinct peaks, with the BDI-Y reaching a higher maximum information value than the CDI. Table 29 again shows that the 12-year-old group's curves reach higher maximum information values than the 9-year-old group's curves and that the 9-year-old group's curves peak at a higher point on the latent trait continuum. The larger maximum and mean information difference scores for the 12-year-old group in

Table 30 also show that the 12-year-old group's information curves again differ more than do the 9-year-old group's information curves.

Looking at the curves across item sets shows that the CDI's maximum information levels remain fairly constant within age groups, although the CDI's information levels are higher in the 12-year-old group than the 9-year-old group. This finding is supported by the CDI maximum information values from Table 29, which are similar within the age groups and across the item sets. The BDI-Y curves' maximum information levels vary across item sets, although the change appears greater in the 12-year-old group than the 9-year-old group. Consistent with these visual analyses, Table 29 shows that the difference between the 9-year-old group's maximum BDI-Y information values across the item sets is 0.09, and the difference in the 12-year-old group's values is 0.66.

Matched-sample statistical tests of discrimination values. In this set of analyses, the differences between the BDI-Y and CDI discrimination parameters within the two item sets were tested in both age groups, for a total of 4 tests, as displayed in Table 31.

As in Research Question 1, the item discrimination parameter values in both item sets and both age groups were analyzed using matched-sample tests. In all cases, Shapiro-Wilk tests of the between-measure difference scores' normality were non-significant. [In the 9-year-old group, for content and frequency-matched items, $W(61) = .91$, $p = .48$ and for content only-matched items, $W(61) = .85$, $p = .22$. In the 12 year-old group, for content and frequency-matched items, $W(61) = .93$, $p = .57$; and for content only-matched items, $W(61) = .91$, $p = .44$.] Because the difference scores were normally distributed,

matched-sample *t*-tests were used. The results of those tests are presented in Table 31. Although not all tests were significant, effect sizes were large. The effect sizes for the 12 year-old group were larger than 9-year-old group's effect sizes, and the tests in the 9-year-old group were both highly similar to each other and not statistically significant. In the 12-year-old group, the test values for the set of items matched for content only were larger than the values for the set matched for content and frequency.

Table 31. Matched-sample *t*-test of discrimination values

	9 year-olds				12 year-olds			
	t	df	*p*	*d*	*t*	df	*p*	*d*
Content & frequency	-1.83	4	.14	0.81	2.33	4	.08	1.05
Content only	1.87	4	.14	0.84	3.83	4	.02	1.71

The construct reliability values associated with the two age groups' model parameters are presented in Table 32. Overall, the reliability values for the 9-year-old group are smaller than the reliability values for the 12-year-old group. In both age groups, the reliability values for the item set matched for content only are higher than the reliability values for the item set matched for content and frequency. Differences between the measures' reliability values range from .05 (both item sets in the 12-year-old group) to .10 (content-only items in 9-year-old group).

Table 32. Construct reliability values, both age groups and item sets

	9-year-olds		12-year-olds	
	BDI-Y	CDI	BDI-Y	CDI
Content & frequency	.68	.75	.78	.83
Content only	.83	.73	.90	.85

Research Question 2 Results Summary

Research Question 2 was conducted in 2 parts. In the first part, measurement invariance analyses tested for statistical differences between the 9- and 12-year-old groups' model parameter estimates. In the second part, independent group modeling was conducted to test for possible differences in model fit that were not evident when the groups were modeled simultaneously.

Prior to conducting the analyses for Parts 1 and 2, a review of the response frequency distributions was conducted. That review showed that the 9-year-old group's item response frequency distributions were more dispersed for the BDI-Y in both item sets. For the CDI, the response frequency distributions seemed broader for the 12-year-old group than the 9-year-old group in the set of items matched for content and frequency. The age groups' response frequency distributions for the CDI seemed relatively equally dispersed in the set of item matched for content only. A review of the mean parameter estimates for Parts 1 and 2 revealed findings that were generally consistent with these response distribution patterns, and showed that the 9-year-old group typically had wider threshold estimate ranges and lower discrimination parameter estimates than the 12-year-old group. The information value descriptive statistics in Parts 1 and 2 also tended to show that the 9-year-old group's information curves peaked at higher levels of the latent trait continuum, indicating that they had to be more depressed to endorse the items.

In Research Question 2, Part 1, measurement invariance tests comparing the 9-year-old group's responses to the 12-year-old group's responses showed that measurement non-invariance was present in both item sets. In the set of items matched for content and frequency, partial measurement invariance analyses showed that the

thresholds for the BDI-Y and CDI were non-invariant across age groups. Reviews of the groups' mean parameter values showed that the 9-year-old group's BDI-Y threshold range was larger than the 12-year-old group's range, whereas the 12-year-old group's CDI threshold range was larger than the 9-year-old group's range. However, the 9-year-old group's information curves peaked at higher levels of the latent trait for both measures. The 12-year-old group's discrimination parameter estimates for both measures were higher than the 9-year-old group's estimates, and the 12-year-old group's mean and maximum information values were also higher. Nevertheless, the information difference scores between the age groups were similar for both measures (although slightly larger for the 12-year-old group). The effect size indices associated with the non-invariance statistics indicated that differences in threshold estimates contributed to between-group information differences, but that the overall magnitude of the non-invariance was small.

In the set of items matched for content only, the partial measurement invariance analyses revealed that the CDI loadings were invariant, and the CDI thresholds, BDI-Y thresholds, and BDI-Y loadings were non-invariant. A review of the mean item parameter estimates showed that the 9-year-old group's mean threshold estimate ranges were larger than the 12-year-old group's in both measures, and the 9-year-old group's information curves again peaked at higher levels of the latent trait continuum than the 12-year-old group's. The finding of BDI-Y loading non-invariance was supported by a higher mean BDI-Y discrimination parameter estimate in the 12-year-old group than the 9-year-old group. Smaller information differences scores for the CDI than the BDI-Y provide further evidence that the differences in the CDI discrimination parameter estimates across the age groups was not significant. Again, the effect size indices

associated with the non-invariance findings provided evidence that differences in the threshold values contributed to between-group differences in information gathered and that the magnitude of the non-invariance differences was small.

In Part 2 of Research Question 2, the visual analyses of the information curves showed that the 9-year-old group's curves in both item sets were similar to each other; the CDI and BDI-Y curves showed little difference in shape or amount of information gathered. In contrast, the 12-year-old-group's curves showed quite a bit of difference between information gathered across the two measures; the curve differences were most noticeable in the set of items matched for content only, where the BDI-Y gathered more information than the CDI.

Information score data supported the visual analyses, showing that the 12-year-old group's information scores were higher than the 9-year-old group's information scores in both item sets, and their maximum information values occurred at lower levels of the latent trait continuum, much like the findings from Research Question 2, Part 1. Also as in Part 1, the 9-year-old group's threshold estimates tended to cover a wider range of the latent trait continuum and the CDI showed similar levels of maximum information across item sets (i.e., and within age groups). Unlike Part 1, in the set of items matched for content and frequency, the CDI discrimination parameter estimate was not consistently higher than the BDI-Y estimate—in these analyses, the 12-year-old discrimination parameter estimate was higher for the BDI-Y than it was for the CDI.

The matched-sample *t*-tests of discrimination values in Part 2 indicated that discrimination values for the BDI-Y and CDI differed significantly in both item sets for the 12-year-old group but not the 9-year-old group, although effect sizes for both age

groups and both item sets were large. The differences in the construct reliabilities associated with the differences in discrimination parameters were relatively small.

Discussion

The present study had two main aims: (a) to compare the information gathered by self-report items that use retrospective frequency structures to the information gathered by self-report items that do not use frequency structures and (b) to determine whether the age of the participant is related to the amount of information gathered by retrospective frequency self-report items.

The next two sections provide brief summaries of the research questions' findings along with interpretations of those findings. Those sections are followed by a review of the study's limitations and strengths. The discussion concludes with a consideration of implications and possible future directions.

Interpretation of Findings – Research Question 1

Research Question 1 was conducted in two steps. In the first step, information curves were generated for the measures in each item set, and the curves were visually analyzed for differences in shape and height. In the second step, differences between the measures' discrimination parameters within each item set were statistically tested with matched-sample tests.

In the first step, the visual analyses of the information curves failed to produce strong evidence for the effects of retrospective frequency structures on the amount of information gathered by items, although initial reviews of the plots seemed to indicate otherwise. At first glance, the fact that the BDI-Y curve was higher than the CDI curve in the set of items matched for content only seemed to imply that the presence of

retrospective frequency structures affects information levels. The nearly equal BDI-Y and CDI information levels in the set of items matched for content and frequency seemed to further support this conclusion. A closer examination of the curves across the measures revealed otherwise. In contrast to this study's hypotheses, the CDI information levels remained nearly the same in both item sets, meaning that the CDI information levels were almost constant whether or not retrospective frequency structures were used. Meanwhile, the BDI-Y items' information levels varied quite a bit across the item sets despite always using retrospective frequency structures.

Further explorations of the item sets' threshold ranges, frequency distributions, and item wording revealed two possible alternative explanations for the Research Question 1 findings: first, the BDI-Y's thresholds may have affected information levels; second, the BDI-Y information levels may have varied because of item wording. These two explanations are discussed further in the sections that follow and are supplemented with a brief discussion of effect size findings.

Threshold effects. The mean threshold parameter estimate ranges revealed that the CDI and BDI-Y threshold estimate ranges for the set of items matched for content only were highly similar. Similar threshold ranges indicate that response frequencies in the measures were similarly distributed; with one exception (BDI-Y 7, which measures whether children think bad things will happen), the response frequencies in the measures were clustered in the lowest response option. Because item information is affected by both threshold ranges and the number of thresholds, the similar threshold ranges across the measures could mean that the higher information levels in the BDI-Y are at least partially due to the BDI-Y's greater number of thresholds, rather than the BDI-Y's use of

retrospective frequency structures. Although the threshold ranges for the BDI-Y and CDI in the set of items matched for content only were similar, the threshold ranges for the BDI-Y in the set of items matched for content and frequency appeared to be larger than the threshold ranges for the CDI. Because an increased threshold range can reduce maximum information levels gathered, it is possible that the BDI-Y's increased threshold range decreased its information curve height.

Item wording. One possible explanation for the BDI-Y's threshold ranges in the set of items matched for content and frequency could lie in the items' wording. In this set of items, the constructs being measured are constructs that many people have probably experienced at some point—feeling sad, crying, wanting to be alone, having trouble sleeping, and feeling lonely. Research on emotional recall shows that the more common an experience is, the more difficult it is to remember; the more difficult it is to remember, the more likely people are to provide frequency estimates that are affected by a variety of past and current confounds, such as current mood and current beliefs about past events (Courage & Cowan, 2009; Flavell, Miller, & Miller, 2002; Kihlstrom, Eich, Sandbrand, & Tobias, 2000; Levine & Pizarro, 2004; Levine & Safer, 2002; Robinson & Clore, 2002a; Robinson & Clore, 2002b; Sudman, Bradburn, & Schwarz, 1996). For common events, people usually generate retrospective frequency estimates for a specific period of time by counting how often such things typically happen and then applying that general estimate to the specific period of time being measured (Robinson & Clore, 2002b). For instance, in the case of the item about trouble sleeping, if a child believes she has trouble sleeping once every weekend, she is more likely to provide a frequency estimate related to that belief rather than her actual recent experience of sleeping. In short, because the

items in the set matched for content and frequency are about things that commonly happen to people, it is possible that the respondents in this study simply guessed about the frequency with which they have these experiences and thereby were less careful in their answering. The result was a more dispersed pattern of responding for experiences that occur frequently.

Another possibility is that the response options for the BDI-Y do not fit the item stems well. For all items, the lowest possible BDI-Y response option is *never*, a response which is probably difficult to supply in response to an item that measures a common experience, such as crying or having trouble sleeping. The CDI response set, on the other hand, provides less severe low-frequency response options: in place of *never,* the lowest frequency response options consist of wordings like *once in a while* which are probably easier to provide for these items. Children may have therefore provided a wider range of responses to the BDI-Y items because the frequency options do not fit the item stems well.

Compared to the set of items matched for frequency and content, the items in the set matched for content only measure symptoms that are more severe and could therefore be easier to answer with an option of *never*. In this set of items, the symptoms being measured are: thinking that life will be bad, hating oneself, thinking that the respondent causes bad things to happen, wishing to be dead, and feeling unloved. With one exception—BDI-Y 7—the response frequencies for the items in both measures were clustered in the lowest available response option. Because these symptoms seem less likely to occur than the set of symptoms measured in the set of items matched for content only, they could therefore be easier to answer with the BDI-Y's lowest response option of

never. The CDI's lowest available response options are as mildly worded as the response options in the set matched for frequency and content, and consist of wordings like *somebody likes me* and *OK*. Such mild responses are probably easy to provide in response to these items, again making the CDI's lowest response option the most frequent response.

The one exception to the pattern of response frequency distributions in the set of items matched for content only is BDI-Y 7, which measures the belief that the respondent causes bad things to happen. The remaining items are about feelings (hating oneself, wishing to be dead, feeling unloved) and beliefs about future experiences (thinking that life will be bad). Compared to the rest of the items, BDI-Y 7 is present-focused and more concrete. Research on children's responses to emotional self-report items shows that children seem better able to report on behaviors than emotions (Chambers & Johnston, 2002; Cremeens et al., 2007; Jacobson et al., 2012). It is possible that the respondents in this study viewed BDI-Y 7 as more behavioral than the other items in this set. Like the participants in the Chambers and Johnston (2002) study, the better the participants better understood the behavioral item than the emotional items, and so were less likely to provide extreme responses. This explanation may also help to explain the pattern of more dispersed response frequencies in the set matched for content and frequency, several items of which measured behavioral symptoms like crying, being alone, and trouble sleeping.

Effect sizes. The effect sizes for the statistical analyses of discrimination parameter differences between the measures (i.e., the Cohen's *d* effect sizes for the matched-sample *t*-tests) were relatively large. However, the differences between the

measures' construct reliabilities were relatively small, indicating that the practical effects of the discrimination parameter estimate differences (i.e., on construct reliability) may not always be large. To some extent, this finding is not surprising—if the measures did not have relatively strong internal consistency reliabilities, they would not be suitable for clinical use. A potential conclusion is that the practical implications of the statistical analyses—at least in terms of internal consistency—may be limited.

Summary. Combined, the Research Question 1 findings seem to have provided evidence for effects of item wording on threshold ranges and thereby on item information levels. Findings for the effects of retrospective frequency structures on item information were inconclusive. Statistical tests of discrimination parameters, although significant, may not always translate into practically meaningful differences in construct reliability differences, a finding which is not entirely unsurprising in light of the assumption that measures used in clinical settings already possess high reliability levels.

Interpretation of Findings – Research Question 2

The goal of Research Question 2 was to determine whether the ages of the participants were related to the amount of information gathered by retrospective frequency self-report items. Analyses were conducted on 9-year-olds' and 12-year-olds' responding and were conducted in two steps. The first step included both age groups in the same model for the purpose of conducting measurement invariance tests. The benefit of this model is that it allowed for statistical tests of differences in the groups' parameter estimates. A drawback to this approach is that model fit information is provided only for the model as a whole; differences between the groups' model fit can be tested statistically, but exact model fit statistics and indices cannot be generated for each group.

To compensate for this limitation, the age groups were modeled independently in the second part of the research question.

The findings from Research Question 2 are provided in the sections that follow. Interpretations of each part's findings are presented, followed by possible interpretations for the question as a whole. The section finishes with a brief summary.

Part 1 findings. Part 1 analyses were tests of measurement invariance analyses across the 9- and 12-year-old groups. These findings revealed that measurement non-invariance was present in both item sets and that the most frequent source of measurement non-invariance was the 9-year-old group's threshold estimates. In both item sets, the 9-year-old group's threshold estimate ranges were typically larger and centered at a higher level of the latent trait continuum than the 12-year-old group's threshold estimates. Consistent with these findings, the 9 year-olds' response frequencies were typically more dispersed across the available response options than the 12 year-olds' response frequencies. Taken together, these findings and observations indicate that the 9-year-old participants provided less consistent response patterns and had to be more depressed than the 12-year-old participants to endorse the higher response categories.

Effect sizes associated with differences in predicted scores (i.e., STDS, UETDS, ETSSD) were small. A variety of effect size estimates were presented, with the most consistent finding being that effect sizes interpreted using Cohen's *d* (1988) guidelines were small. The relatively small effect sizes imply that although the age groups present significantly different parameter estimates, the practical effects of these parameter differences may be small.

Part 2 findings. In part 2, data for the two age groups were modeled independently. The Part 2 analyses were conducted using the same methods as the Research Question 1 analyses: differences in information curves were compared visually, and differences between the measures' discrimination parameters were analyzed with matched-sample statistical tests.

When the 9- and 12-year-old groups' data were modeled independently, the most striking finding was that the model used throughout the earlier analyses did not fit the 9-year-old group's data well. When the groups were modeled together in earlier analyses, the model required two changes (i.e., two additional item-error covariations) to achieve sufficient model fit. When the age groups were modeled independently, the 9-year-old group's model required another two changes (i.e., another two item-error covariations) to fit the 9-year-old group's data. (To maintain consistency across models, these changes were also added to the 12-year-old group's model.) So many needed adaptations to the model indicate that the model may not be well-suited to the 9-year-old group.

Another finding in this part of the analyses was the pattern of greater differences between the 12-year-old group's information curves across measures compared to differences between the 9-year-old group's information curves. In the 12-year-old group's plots, the differences between the BDI-Y and CDI curves were noticeable, particularly in the set of items matched for content only. In the 9-year-old group's plots, the measures' curve shapes and heights were similar to each other in both item sets. The implication of these findings is that the 12-year-old group was more sensitive to differences in item structure than the 9-year-old group.

As in Research Question 1, the effect sizes associated with differences between the measures' discrimination parameters were large. The estimated differences between the construct reliability values was still somewhat small, although in two cases it was twice the size of the differences found in Research Question 1.

Possible alternative explanations. The findings from Research Question 2 could be associated with a number of possible explanations. Two possibilities include (a) potential differences between the age groups' factor structures and (b) developmental differences between the 9-and-12 year-olds' ability to self-report. Each of these alternatives is discussed below.

Factor structure. The finding from Part 2 of Research Question 2 indicating that the model did not fit the 9-year-old group well strongly implies that the 9-year-old group had a different factor structure than the 12-year-old group. The findings from Part 1 also suggest possible factor structure differences through the findings of significant differences between the age groups' thresholds. Although there is evidence to suggest that rates of depression increase as girls approach puberty (Ge, Conger, & Elder, 2001; Patton, et al., 2008; Twenge & Nolen-Hoeksema, 2002), little has been done to explore whether these differences in presentation across age are due to measures' inconsistent factor structures across ages. Instead, depression is generally presumed to present relatively similarly across age groups despite evidence that some facets of depression become more prevalent as children age (Stark et al., 2006). For instance, very young children are more likely to present with agitation or anhedonia as the primary mood symptom, rather than flat affect (Luby et al., 2003). As another example, prepubertal children are less likely than older children to endorse items reflecting cognitive

symptoms, such as feelings of hopelessness and self-blame (Weiss & Garber, 2003; Yourbik, Birmaher, Axelson, Williamson, & Neal, 2004).

Because depression is presumed to present similarly in all age groups, depression assessment measures, such as the ones sampled in this study, commonly have children of all ages respond to the same items. Differences in rates of symptom reporting are then accounted for by creating different norming groups according to age and sometimes gender. For the CDI (Kovacs, 1992), norming groups are provided for gender and age group; the age groups provided are 7 to 12 years-old in one group and 13 to 17 years-old in the other. For the BDI-Y (Beck et al., 2001), norming groups are again provided by gender and age group; in this measure the age groups are 7 to 10 years old, 11 to 14 years old, and 15 years and older. In both measures, the older age groups require higher scores to achieve clinically significant levels of depression.

Although the norming groups are assumed to correct for differences in symptom presentation, it is not clear whether this correction is sufficient. It is possible that children from various age groups manifest depressive symptoms so differently that some items are completely irrelevant for some respondents (thereby resulting in symptom underreporting) or are associated with constructs other than depression (thereby resulting in symptom overreporting). Few studies have examined this question, and their methods have been sufficiently diverse that comparing findings across the studies is difficult (e.g., Boylan, Miller, Vaillancourt, & Szatmari, 2011; Fonseca-Pedrero, et al., 2010; van Beek, Hessen, Hutteman, Verhulp, & van Leuven, 2012; Verhoeven, Sawyer, & Spence, 2013). These studies have examined item stability over time within the same sample, examined item stability across age groups, studied different age groups, used different assessment

measures, and used different statistical procedures. Of the four studies cited here, one concluded that depressive symptoms were unstable across childhood as the same sample aged (e.g., 8-14 years-old; Boylan et al., 2011), the second concluded that symptoms are measurement non-invariant across childhood ages (van Beek et al., 2012), a third concluded that symptoms are stable within the same sample as adolescents age (Fonseca-Pedrero et al., 2010), and a fourth found that factor structures remained the same across adolescence although the strength of the item loadings for some items changed across age groups (Verhoeven et al., 2013). Taken together, the results of these and the current study have yet to clarify whether norming groups are sufficient to account for children's different presentations of depressive symptoms across age groups, or whether children's depression measures should include different items for different age groups.

Developmental differences in self-reporting abilities. Children of various ages differ in their ability to self-report. In general, older children are thought to be better self-reporters on internal states than younger children for two broad reasons: first, older children seem to better understand the concepts involved in the items; and second, older children seem better able to make use of differences in items' structures. In this study, both reasons could have accounted for the 12-year-old group having higher information levels than the 9-year-old group.

The hypothesis that older children seem to better understand the concepts in self-report items is in part founded on child development research suggesting that older children have a more established sense of self (Stone & Lamanek, 1990) and a better sense of time (Friedman, 2007; Hudson & Mayhew, 2009). The hypothesis is also founded on studies that have shown that children of all ages—and younger children in

particular—have trouble understanding the time concepts in many measures (Breton et al., 1995; Fallon & Schwab-Stone, 1994; Levine & Humberman, 2002; Lucas et al., 1999; Schwab-Stone et al., 1994). This body of research also shows that younger children also have more trouble understanding the emotional concepts in self-report items than the behavioral concepts (Chambers & Johnston, 2002; Cremeens et al., 2007; Jacobson et al., 2012). Like the findings in this study, extant research has also shown that younger children's responses to items are less sensitive to item structure, with the result that they often provide highly similar answers even when item structures are significantly changed (Borgers et al., 2003).

Summary. The findings from Research Question 2 strongly suggest that the items studied are non-invariant across age groups and gather more information in the 12-year-old group than the 9-year-old group. Because of the many elements which could have affected the information levels across the age group, it is not clear whether the use of retrospective frequency structures affected information curve levels.

Study Limitations

Internal validity. At least six elements may have limited the internal validity of the conclusions from this study: number of item pairs, differences in the measures' item structures other than the use of retrospective frequency structures, the sample sizes used, measure administration procedures, available statistical modeling guidelines, and the assumption that depression is the same continuous latent construct in all age groups. These topics are discussed in the paragraphs that follow.

First, in both sets of analyses, the number of item pairs available for analysis was limited. Findings could therefore have been due to peculiarities about this set of items

that might have been compensated for were the item pool larger (which are discussed in paragraphs below). In the cases of the statistical tests of discrimination parameters, some of the tests showed large effect sizes even though the tests were not statistically significant. The limited item pool meant that some of the statistical tests were under-powered.

A second consideration is the number of differences between the structures of the measures' items. The items vary across the measures in their amount of wording, number of response options, physical placement of response options, and strength of response option wording. These differences are in addition to the examined difference between the measures' use of retrospective frequency structures. With so many differences between the measures' item structures, it is possible that some factor other than retrospective frequency structures could have affected the information levels found in this study. As one example and as already discussed above, the BDI-Y's larger number of thresholds could be causing its higher information levels in the set of items matched for content only. Another possibility is the placement of the response options. The CDI response options are ordered vertically whereas the BDI-Y's options are ordered horizontally. Previous research has shown that children prefer vertically oriented scales (Rebok et al., 2001; van Laerhoven et al., 2004), implying that the CDI items might have been easier for the children to answer. As another example, the response options on the BDI-Y were ordered so that the most negative response option (*never*) was first. Had they been ordered so that the most endorsing response option (*always*) had been first, children might have provided higher or different scores (Betts & Hartley, 2012). As a final example, differences in the response option wording could have caused differences in

children's responses to the measures. The response options for the CDI items are all full sentences. The response options for the BDI-Y items consist of the same four frequency words for every item. Because children have trouble with vague and non-concrete wording (Borgers & Hox, 2001; Holaday & Turner-Henson, 1989; Otter, Mellenbergh, & de Glopper, 1995; Rebok et al., 2001; Scott, 1997), the children may have found the items on one set easier to understand than items on the other. Given the CDI items' relatively constant information levels across item sets, it could be that they found the fully-labeled CDI response options equally easy (or difficult) to understand in both item sets. As discussed above, the varying levels of the BDI-Y items' information levels across item sets could indicate that the children found some of the BDI-Y response options difficult to understand when paired with certain item stems, such as the common behavioral experiences with the lowest response option of *never*.

A third consideration is the sample size used. Typically, a sample size of 500 in each group modeled is recommended for stable model estimation (e.g., model and statistical precision; Embretson & Reise, 2000; Tabachnik & Fidell, 2001). In this study, however, the sample sizes for the two groups in Research Question 2 were 424 and 366. It is possible that a larger sample size would have produced results that differ from the current results due to an increase in the model's statistical precision; stated differently, larger sample sizes could have allowed for the estimation of a more accurate model and different parameter estimates. What the pattern of results would be with larger sample sizes cannot be determined in advance.

Fourth, the measure administration could have affected responding. The measures were administered to children in groups as large as 100. Under such conditions, it is

possible that children rushed, were embarrassed to provide accurate responses for fear that classmates might be watching them, or did not ask questions about items they did not understand. Data are not available to determine whether there were differences between the administration patterns in the 9-year-old and 12-year-old groups. Because the results of this study seem to show that the 9-year-old group was less sensitive to differences in measure structure than the 12-year-old group, the 9-year-olds' response patterns might have been quite different if administration settings were more individualized or they had more adult support to help them understand the items.

The fifth element considered here is the set of guidelines available for the modeling used in this study. As briefly mentioned in the *Method* section above, recommendations about the number of fit indices to use and appropriate cut off scores varies across published guidelines. As such, new studies have a wide range of fit criteria from which to choose. A change in the selection of fit criteria within this study could change this study's conclusions. For instance, the fit criteria for this study could have been relaxed by using lower cut-off values for the fit indices. Under such criteria, it might have been possible to conclude that the IRT model in Part 2 of Research Question 2 did fit both the 9-year-old and 12-year-old groups well (i.e., instead of the current interpretation, which states that the model does not fit the 9-year-old group well). It might also have been possible to make the fit criteria more stringent, for instance by requiring that the chi-square statistic be non-significant. In this case, the baseline model would not have been said to fit well, and further adjustments would have been necessary in order to proceed with the study.

The sixth and final element considered in this discussion is the assumption that all age groups included in this study could be equally well-represented by the same continuous latent construct, which is an assumption of all IRT models. The items in this study were drawn from two measures that assess for depressive symptomatology. The methods to test for unidimensionality followed extant guidelines about the assumptions for the Graded Response Model (Samejima, 1969), procedures for using confirmatory factor analysis to test for unidimensional model fit (Muthén & Muthén, 2002), and guidelines for interpreting model fit indices and statistics (Carle et al., 2008; Cook et al., 2009; Hu & Bentler, 1998; Yu, 2002). Based on these procedures, a unidimensional model could be said to fit the entire sample well. When the 9-year-old and 12-year-old groups were modeled independently, however, the model no longer fit the 9-year-old group well, implying that the 9-year-olds have a different factor structure from the 12-year-olds. The extent to which the results from Research Question 1 apply to the 9-year-old group must therefore be questioned because violations of the unidimensionality assumption mean that non-random elements other than those identified by the model are affecting the model (De Vellis, 2003; Embretson & Reise, 2000; Fletcher & Hattie, 2004; Hambleton & Jones, 1993; Reeve and Mâsse, 2004; Reise & Henson, 2003; Samejima, 1969) As such, it is possible that the parameter estimates do not reflect the younger groups' (particularly the 9-year-old group's) data well and error levels in the younger groups are high.

Unfortunately, the results from this study do not clearly identify the non-random elements that could be affecting the 9-year-olds' responses. This study does point to two possibilities, both of which have already been introduced. One of those possibilities is

that 9- and 12-year-olds' depressive symptomatology may not be the same (e.g., Boylan, Miller, Vaillancourt, & Szatmari, 2011; Fonseca-Pedrero, et al., 2010; van Beek, Hessen, Hutteman, Verhulp, & van Leuven, 2012; Verhoeven, Sawyer, & Spence, 2013). In other words, not all the items examined in this study may measure depression in 9-year-old children. A second possibility could be differences in factor structure related to some as-yet-unidentified component of the items' structure. For instance, at least two studies (Benson & Hocevar, 1985; Marsh, 1986) found that negatively-worded items form different factor structures than positively-worded items when both are entered into a factor analysis model. A third study (Borgers et al., 2003) found that fully-labelled response options produced a different factor structure from partially-labelled response options. Although it is not clear in this study what feature or features about the items could be generating differences in factor structure, the use of negative wording is one possibility. A number of the items on the CDI were reverse-worded from the matching BDI-Y items—2 items in the set matched for content and frequency, and 4 in the set matched for content only. Another possibility is the presence of both behaviorally-worded and emotionally-worded items. Literature already reviewed in this study has shown that children do not understand emotional items as well as they understand behavioral items (Chambers & Johnston, 2002; Cremeens et al., 2007; Jacobson et al., 2012). Because the literature reviewed earlier also suggests that younger children are less sensitive to differences in item structure (Borgers et al., 2003), hypotheses about factor structure differences in the 9-year-old group due to differences in item structure are difficult to support without further research.

External validity and generalizability. This study used a limited set of items and a sample that that is non-representative of the general population. Combined with a set of statistical procedures whose implementation practices vary across studies, a number of limitations to external validity and generalizability may be present.

A difficulty with available statistical procedural guidelines is the absence of clear guidelines regarding the identification of well-fitting models and judgments about differences in information curves (as described in Appendix A). In this study, as in many others, decisions about model fit and meaningful differences in curves were made on the basis of divergent guidelines. Studies often use disparate methods to determine model fit and meanings of modeling outcomes (e.g., Beherend, Thompson, Meade, Newton, & Grayson, 2008; Capara, Steca, Zelli, & Capanna, 2005; Cooper & Gomez, 2008; Edelen, McCaffrey, Marshall, & Jaycox, 2008; Eden, McCaffrey, Marshall, & Jaycox, 2008; Fletcher & Hattie, 2004; Gomez, 2008; Hafsteinsson, Donovan, & Breland, 2007; Hall, Reise, & Haviland, 2007; Olino et al., 2012; Purpura & Lonigan, 2009). Comparing findings across studies is therefore difficult.

The sample used in this study may also limit generalizability. This study's sample included only girls, 9 to 12 years of age, and included a sample in which 40% of respondents self-reported their ethnicity as Hispanic. Although IRT parameter estimates are supposed to be sample-independent, the sample independence only applies to samples drawn from similar groups (Embretson & Reise, 2000). Thus, the results generated by a sample with similar demographics would match the results found in this study, although modeling based on a sample with very different demographics might find different results.

The type of items analyzed must also be considered. This study examined items that are intended to measure symptoms of an internalizing disorder. Furthermore, the items in this study contained items that seemed to assess a mixture of behavioral, emotional, and cognitive symptoms. As reviewed earlier, children do not report as well on internalizing symptoms as they do on behaviors. It is possible that if the items were all clearly about behaviors, the effects of the presence and absence of frequency structures may have been easier to identify. Thus, findings from this study may not generalize to items that do not examine internalizing symptoms such as measures that examine children's academic behaviors, eating behaviors, or physical activity. The item sets on these measures could be easier for children to understand with the result that when they are paired with the hypothetically difficult retrospective frequency response sets, information levels would be lower than information levels for identical item stems paired with non-frequency response options.

A final limitation is that the items studied here were subsets of larger measures and not measures in their entirety. Thus, comparisons can be made between these items and other, similar items, but comparisons cannot be made across measures—the CDI as a whole, for example, cannot be said to be more stable in the amount of information its items gather than the BDI-Y. For instance, it is possible that if all the items from the two measures were modeled, the BDI-Y items in the set of items matched for content and frequency could be the only items whose information levels were low; the remainder may consistently gather more information than the CDI items, as in the set matched for content only. It could also be true that the CDI items modeled in this study are unique in

their relatively constant levels of information; the rest of the items could gather a widely diverse amount of information.

Study Strengths

This study possesses several strengths that make it a unique foundation for future research. The large Hispanic population in this study's sample and the finding that the proposed IRT model fit both the Hispanic and non-Hispanic groups well adds to the limited research on self-reporting of depressive symptoms in Hispanic populations of elementary- and middle-school-aged children (e.g., Marmorstein, 2012; Molina, Gomez, & Pastrana, 2009; Stapleton, Sander, & Stark, 2007; Twenge & Nolen-Hoeksema, 2002). Far more work has been done with adolescent-aged Hispanic populations (e.g., Cespedes & Huey, 2008; Chao & Otsuki-Clutter, 2011; Crockett, Randall, Shen, Russell, & Driscoll, 2005; Lorenzo-Blanco, Unger, Baezconde-Garbanati, Ritt-Olson, & Soto, 2012; McLaughlin, Hilt, & Nolen-Hoeksema, 2007; Stefanek, Strohmeier, Fandrem, & Spiel, 2012). Although there are studies showing that Hispanic populations self-report higher rates of depression than non-Hispanic White populations (Twenge & Nolen-Hoeksema, 2002), researchers are still not sure of the reasons for this disparity (Cespedes & Huey, 2008). Consistent with extant research, this study found that the Hispanic group reported a higher rate of depressive symptoms than the non-Hispanic group. When the groups' data were modeled with a fully constrained confirmatory factor analysis, the model fit the group well, indicating that the differences in the groups' scores are potentially due to causes other than differences between the groups' factor structures.

This study has the additional benefit of using items that are commonly used in clinical practice. In other words, they were not items that were generated in a research lab

setting and which may or may not resemble items that are used in applied settings. Using items that are used in clinical practice has the benefit of allowing for easier generalization of the results to applied settings. These items also allow for examinations of the effects of multiple differences in item facets on items' effectiveness, whereas most of the research reviewed in this paper shows item facets in lab studies are often adjusted individually.

In addition to using items commonly used in clinical practice, this study was able identify both a test set of items and a control set of items from the same two measures. Finding existing items that are highly matched for content and basic wording is a significant challenge. This study was fortunate to have access to not just a test set of matched items but a control set as well.

To date, this study is one of the few that has examined differences in self-reporting ability within the concrete operational stage. In most studies, children within this age range are commonly treated as equally capable self-reporters (Amato & Ochiltree, 1987; Borgers et al., 2000; Fallon & Schwab-Stone, 1994; Woolley, Bowen, & Bowen, 2004), which this study has shown not to be the case.

Implications

This study is the first to examine the effects of retrospective frequency structures on item information levels across age groups within Piaget's concrete operational stage. Although conclusions about the effects of those structures are tenuous at best, this study highlighted four other important findings. The first two are related to the assessment of depression in children and the second two are related to the general study of self-report measures for children. The four implications are discussed in separate paragraphs below and are followed by a paragraph that considers clinical implications.

The first two implications are tied to each other and concern depression self-report measures for children. First, this study cast doubt on whether the factor structure underlying the items in this study was the same for the 9-year-old group as it was for the 12-year-old group. Although the items were not the depression measures in their entirety, they were a subset of items intended to measure depression and could therefore have been reasonably assumed to have represented a single latent depression factor. Given the longstanding theory that depression presents similarly across all age groups (e.g., Stark et al., 2006), the unconstrained statistical model examined in Research Question 2, Part 2 of this study (i.e., the model that tested only for the number of latent factors) should have demonstrated similar levels of model fit in both age groups. The failure of the model to fit the 9-year-old group well strongly implies that the factor structure of the 9-year-old group is different than the factor structure of the 12-year-old group. Extant research on child depression suggests that the factor structure differences could be due to different presentations of depression across childhood, with behavioral symptoms being more relevant to younger children and cognitive symptoms more relevant to older children. Younger children may also have trouble understanding the emotional concepts in some items, and therefore rely more on the extreme ends of response option sets for those items.

Because this study provides some evidence for differences in the 9- and 12-year-old groups' factor structures, the second implication is that the use of norming groups to correct for age groups' rates of responding on self-report measures is worthy of further investigation. It may be the case that a measure which contains a balance of both cognitive and behavioral symptoms is capable of fully representing the symptoms present

in both age groups. For such a measure, higher cutoff scores for the older age group may adequately correct for the tendency of 9-year-olds to endorse fewer symptoms overall. How many items would be needed, whether current measures do contain such a balance, and whether norming groups are a sufficient correction is unclear and was not addressed by this study.

The third implication shifts this discussion to considerations about the general study of self-report items. Differences between the amount of information gathered by items commonly used in clinical practice are unlikely to be due to a single component of the items' structures. Instead, differences in information levels are more likely due to some combination of differences in items' structures, such as the use of retrospective frequency structures *and* item wording *and* response option wording. In contrast, clinical studies have often examined differences between items' effectiveness by changing one item component at a time. As a result, the effects of multiple differences in item structure are not yet clear; for instance, it is possible that the information-lowering effects of one difference (e.g., the use of frequency structures) could be outweighed by the presence of another (e.g., clearer item wording) or even reversed by the presence of another (e.g., more response options). While this study does not clearly define the effects of multiple differences in item structure, what it did discover was an instance in which a set of items (e.g., the BDI-Y items) should have gathered less information than another set of items (e.g., the CDI items) because of the presence of retrospective frequency structures, but instead gathered as much or more information.

The fourth implication is the addition of this study to the small body of literature indicating that younger children are less sensitive to differences in item structure than

older children. Importantly, this study highlighted the fact that the differences between the children's ages do not have to be great for children's sensitivity to item structure to manifest. According to the results from this study, even children who are well within the bounds of Piaget's concrete operational stage can manifest different levels of sensitivity to item structure. Thus, measure designers who wish to create a measure for children of varying ages must take extra care to be sure that the important features of the measures are easily recognizable and understood by all ages targeted

The above four implications yield some clinical recommendations for measurement with children who are approximately 9 years old. First, these children will probably provide more accurate responses to items that assess emotional content if the items use simple emotional concepts. For instance, instead of asking about *frustration* and *anxiety*, more effective wording options might include *mad* and *scared*. Internalizing symptoms may also be better measured in this age group through the use of items that measure behavior. For instance, a 9-year-old may more accurately answer an item about crying frequency than an item about sadness frequency. As a corollary to the implication that these children do not always accurately answer items measuring internalizing symptoms, their failure to do so should not be construed as evidence that they are not experiencing those symptoms. This study's results also implied that 9-year-old children may not be sensitive to subtle differences in item structure, which means that key aspects of items should be highlighted and item wording should be clear. As an example, if a measure is designed to assess for symptoms over the previous two weeks, such directions would probably be more effective if they were visually obvious by using techniques such as large or bold font and graphics. Response option wording should also carefully match

the constructs being measured. For instance, instead of using a low-extreme anchor of *never*, some of the BDI-Y items in this study may have displayed lower error in the 9-year-old age group if given less extreme anchor wording (e.g., *almost never* instead of *never* or *almost always* instead of *always*).

Future Directions

The findings from this study lend themselves to multiple suggestions for future research. The suggestions included here are presented in two groups: the first group consists of a set of two sets of replication studies: the first is related to content and the second is related to sampling. The second group of suggestions is more distantly related research questions pertaining to methodological and statistical concerns.

The first group of potential replication studies is related to the topics examined in this project. One set encompasses possible replications of the current study with corrections for some of this study's shortcomings. These replications are studies that include a larger set of items, examine only behavioral or only emotional items, compare strongly and mildly worded items, and use items with the same number of response options. In this study, all of these measure features were confounded. Although this confounding was a strength in that it allowed for an unusual study of items implemented in clinical practice, this confounding limited conclusions because it made the isolation of unique effects difficult. Future studies that replicate this project's use of clinical items but use a larger set of such items would likely reduce the error in the study results. For example, the two sets of items in this study coincidentally differed in the extent to which they measured unusual emotional and cognitive experiences. A larger sample of items might have mitigated the effects of these content differences. Studies that compare children's

responses to behavioral items to their responses to emotional items will allow for further clarification of the extent to which children of various age groups accurately and reliably respond to these two item groups. For instance, one group of children of various ages could be asked to respond to items that assess only behavioral symptoms of depression, and another could be asked to respond to items that assess only cognitive and emotional symptoms of depression. As another variation, a within-groups design could compare a single group's responses to both. Between- and within-groups designs could additionally be used to compare children's responses to strongly and mildly worded response options. Children's responses to self-report items containing retrospective frequency appraisals could also be examined over time and compared to the responses of several known adults and peers to assess for consistency and accuracy. The children's responses could also be compared to clinical judgment to assess for diagnostic accuracy. Across all these variations, maintaining a constant number of response options for the items would help to isolate the effects of the item feature being studied.

A second set of replication studies are related to the ages included in this project, and encompass studies across broader age ranges, studies across more of the age groups between those examined in this study (i.e., 10- and 11-year-olds), and examinations of the effects of reading or education levels on responding patterns. This study only examined differences across 9- and 12-year-olds' self-reports. Results from this study showed that differences existed across the 9- and 12-year-old groups' response patterns, but identical analyses across the 10- and 11-year-old groups might have yielded further insight into the age at which response patterns change. Other studies could also expand the lower age range of children included. According to Piaget's theories, 9-year-olds are

in the mid-range of the concrete operational stage and 12-year-olds can vary in their belonging to the concrete and formal operational age groups. Including children as young as 7-years-old in future analyses may reveal further differences in the responses of children of varying ages to depression self-report items across Piaget's concrete operational stage. For instance, studies of symptomology across 7- and 12-year-old depressed clients might help reveal not only differences across age groups' self-reporting abilities but also differences across age groups' symptomatology. Finally, this study assumed that children within age groups possessed similar reading and educational levels. Depending on the time of a measure's administration, the typical 9-year-old can be in either third, fourth, or fifth grade. Reading and reasoning abilities could therefore vary quite a bit as a function of not only age but also school grade or academic ability (e.g., standardized testing scores, school grades, achievement test scores). Age may be a convenient norm because it is simple to assess, but it is possible that reading or academic achievement assessments are a more accurate measure of self-reporting ability.

The second group of potential future directions is related to statistical and methodological concerns. This paragraph discusses future research related to statistical concerns and the next paragraph discusses future research directions related to methodological concerns. One limitation that pervaded the study was that it was forced to rely on subjective comparisons of information curve levels because guidelines methods for testing differences in information curve levels for IRT studies have not been defined. Future research could include efforts to clarify methods for examining differences between information curves by summarizing and testing existing methods. Statistical tests for two other comparisons did exist – tests between measures' discrimination

parameters and measurement invariance tests between the two age groups' model parameter estimates. For both sets of tests, the effect size procedures did not consistently provide convincing evidence that the statistically significant results would result in practically significant differences. Future research could work toward clarifying the relation between statistical and practical significance. Research efforts to do so might include studies in which statistically significant findings are compared to benchmarks of practical significance like diagnostic rates. To improve generalizability and cross-study comparisons, research efforts could also refine guidelines about acceptable model fit by summarizing and testing existing guidelines. Further research about the effects of threshold ranges and numbers on information levels is also needed. Although it is known that more threshold ranges typically increase information levels, if both threshold numbers and ranges vary, their effect on information levels remains unclear and obfuscates results interpretation.

Methodological concerns include those related to further investigations of the relative effects of various differences across measures' item structure as well as studies of the synergistic effects of multiple item structure differences. One approach to pursing these efforts might involve the creation of multiple variants of an existing measure, with each variant possessing only one difference between the variant's item structure and the original's item structure. Differences between children's responses to the multiple variants and the original measure would then be compared. As an example, such a study could use one of the measures included in this project, like the BDI-Y. One variant might eliminate retrospective frequency structures, another might use less extreme response option wording, and another might increase the amount of response option wording.

Children's responses to each of these variants could be compared to their responses to the original measure to determine which of the variants produced responses most like responses to the original measure (and thereby provided evidence of lesser effects on children's responses) and which of the variants produced responses most unlike the responses to the original measure (and thereby produced evidence of greater effects on children's responses). This method of comparing responses across multiple measure variants could be expanded to include studies of the synergistic effects of multiple item feature differences. Instead of changing only one item structure component, a study of synergistic effects could involve variants that changed multiple features, such as one variant that altered strength of response option wording and number of response options, and another variant that altered the strength of response option wording and the use of retrospective frequency structures. Like this example, variants could be created in sets such that the variants contained at least one commonality. Examinations could thereby be conducted in nested levels across individual variants and groups of variants.

References

Aaker, J., Drolet, A., & Griffin, D. (2008). Recalling mixed emotions. *Journal of Consumer Research, 35*, 268-278.

Amato, P., & Ochiltree, G. (1987). Interviewing children about their families: A note on data quality. *Journal of Marriage and the family, 49*, 669-675.

American Psychiatric Association (2000). *Diagnostic and statistical manual of mental disorders (4th ed., text revision)*. Arlington, VA: American Psychiatric Association.

Ayidiya, S., & McClendon, M. (1990). Response effects in mail surveys. *Public Opinion Quarterly, 54*, 229-247.

Baxter, S. D., Thompson, W., Litaker, M., Frye, F., & Guinn, C. (2002). Low accuracy and low consistency of fourth-graders' school breakfast and school lunch recalls. *Journal of the American Dietetic Association, 102*, 386-395.

Beck, J., Beck, A., & Jolly, J. (2001). *Beck Youth Inventories of Emotional and Social Impairment*. The Psychological Corporation.

Behrend, T., Thompson, L.F., Meade, A.W., Newton, D.A., & Grayson, M.S. (2008). Using IRT to study gender differences in medical students' specialization decisions. *Journal of Career Development, 35,* 60-83.

Benson, J., & Hocevar, D. (1985). The impact of item phrasing on the validity of attitude scales for elementary school children. *Journal of Educational Measurement, 22*, 231-240.

Ben-Zeev, D., Young, M., & Madsen, J. (2009). Retrospective recall of affect in clinically depressed individuals and controls. *Cognition and Emotion, 23*, 1021-1040.

Betts, L., & Hartley, J. (2012). The effects of changes in the order of verbal labels and numerical values on children's scores on attitude and rating scales. *British Educational Research Journal, 38,* 319-331.

Borgers, N., de Leeuw, D., & Hox, J. (2000). Children as respondents in survey research: Cognitive development and response quality. *Bulletin de Methodologie Sociologique, 66*, 60-75.

Borgers, N., & Hox, J. (2001). Item nonresponse in questionnaire research with children. *Journal of Official Statistics, 17*, 321-335.

Borgers, N., Hox, J., & Sikkel, D. (2003). Response quality in survey research with children and adolescents: The effect of labeled response options and vague quantifiers. *International Journal of Public Opinion Research, 15*, 83-94.

Borgers, N., Hox, J., & Sikkel, D. (2004). Response effects in surveys on children and adolescents: The effect of number of response options, negative wording, and neutral mid-point. *Quality & Quantity, 38*, 17-33.

Boylan, K., Miller, J., Vaillancourt, T., & Szatmari, P. (2011). Confirmatory factor structure of anxiety and depression : Evidence of item variance across childhood. *International Journal of Methods in Psychiatric Research, 20*, 194-202.

Brainerd, C. (1978). *Piaget's theory of intelligence*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Breton, J.-J. Bergeron, L., Valla, J.-P., Lepine, S., Honde, L., & Gauden, N. (1995). Do children aged 9-11 years understand the DISC Version 2.25 questions? *Journal of the American Academy of Child and Adolescent Psychiatry, 34*, 946-954.

Brown, N., Williams, R., Barker, E., & Galambos, N. (2007). Estimating frequencies of emotions and actions: A web-based diary study. *Applied Cognitive Psychology, 21*, 259-276.

Capara, G.V., Steca, P., Zeli, A., & Capanna, C. (2005). A new scale for measuring adults' prosocialness. *European Journal of Psychological Assessment, 21(2)*, 77-89.

Carle, A., Millsap, R., & Cole, D. (2008). Measurement bias across gender on the Children's Depression Inventory: Evidence for invariance from two latent variable models. *Educational and Psychological Measurement, 68*, 281-303.

Céspedes, Y., & Huey, S., Jr. (2008). Depression in Latino adolescents: A cultural discrepancy perspective. *Cultural Diversity and Ethnic Minority Psychology, 14*, 168-172. DOI: 10.1037/1099-9809.14.2.168

Chambers, C., & Johnston, C. (2002). Developmental differences in children's use of rating scales. *Journal of Pediatric Psychology, 27*, 27-36.

Chao, R., & Otsuki-Clutter, M. (2011). Racial and ethnic differences: Sociocultural and contextual explanations. *Journal of Research on Adolescence, 21*, 47-60. DOI: 10.1111/j.1532-7795.2010.00714.x

Chernyshenko, O., Stark, S., Chan, K.-Y., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Researcher, 36*, 523-562.

Cheung, G., & Rensvold, R. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management, 25*, 1-27.

Christianson, S.-A., & Safer, M. (1996). Emotional events and emotions in autobiographical memories. in D. Ruben (Ed.), *Remembering our past* (pp. 218-243). Oxford: Oxford University Press.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd edition). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cooper, A., & Gomez, R. (2008). The development of a short form of the Sensitivity and Punishment and Sensitivity to Reward Questionnaire. *Journal of Individual Differences, 29(2),* 90-94.

Courage, M., & Cowan, N. (2009). What's new in research on the development of memory in infants & children? In M. Courage & N. Cowan (Eds.), *Development of memory in infancy and childhood* (pp. 1-10). New York: Psychology Press.

Craighead, W. E., Smucker, M., Craighead, L. W., Ilardi, S. (1998). Factor analysis of the Children's Depression Inventory in a community sample. *Psychological Assessment, 10*, 156-165.

Cremeens, J., Eiser, C., & Blades, M. (2007). Brief report : Assessing the impact of rating scale type, types of items, and age on the measurement of school-age children's self-reported quality of life. *Journal of Pediatric Psychology, 32*, 32-138. doi:10.1093/jpepsy/jsj119

Crockett, L., Randall, B., Shen, Y.-L., Russell, S., & Driscoll, A. (2005). Measurement equivalence of the Center for Epidemiological Studies Depression Scale for

Latino and Anglo adolescents : A national study. *Journal of Consulting and Clinical Psychology, 73*, 47-58. DOI: 10.1037/0022-006X.73.1.47

Dawes, R., Faust, D., & Meehl, P. (1989). Clinical versus actuarial judgment. *Science, 243*, 1668-1674.

de Leeuw, E., Borgers, N., & Smits, A. (2004). Pretesting questionnaires for children and adolescents. In S. Presser, J. Rothgeb, M. Couper, J. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires* (pp.400-429). Hoboken, NJ: John Wiley & Sons, Inc.

de Tovar, C., von Baeyer, C., Wood, C., Alibeu, J.-P., Houfani, M., & Arvieux, C. (2010). Postoperative self-report of pain in children: Interscale agreement, response to analgesic, and preference for a faces scale and a visual analogue scale. *Pain Research and Management, 15*, 163-168.

DeVellis, R. (2003). *Scale development: Theory and applications (2nd ed.)*. Thousand Oaks, CA: Sage Publications.

Dillman, D. (2000). *Mail and internet surveys: The tailored design method (2nd ed.)*. New York: John Wiley & Sons, Inc.

Edelen, M.O., McCaffrey, D.F., Marshall, G.N., & Jaycox, L.H. (2009). Measurement of teen dating violence attitudes: An item response theory evaluation of differential item functioning according to gender. *Journal of Interpersonal Violence, 24*, 1243-1263.

Eden, M.O., McCaffrey, D.F., Marshall, G.N., & Jaycox, L.H. (2008). Measurement of teen dating violence attitudes: An item response theory evaluation to differential

item functioning according to gender. *Journal of Interpersonal Violence, 24*, 1243-1263.

Embretson, S., & Reise, S. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Fallon, Jr., T., & Schwab-Stone, M. (1994). Determinants of reliability in psychiatric surveys of children aged 6-12. *Journal of Child Psychology and Psychiatry, 55*, 1391-1408.

Faul, F., Erdfelder, E., Lang, A.-G. & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39, 175-191.

Fisher, Lucas, L., Lucas, C., Sarsfield, & Shaffer (2006). *Interviewer Manual*, Columbia DISC Development Group. downloaded 6/29/13 from http://www.cdc.gov/nchs/data/nhanes/limited_access/interviewer_manual.pdf.

Fitzpatrick, S.J., Cois, S. S., Chen, S.-K., Hou, L., & Dodd, B. G. (1994). IRTINFO: A SAS macro program to compute item and test information. *Applied Psychological Measurement, 18*¸390.

Fivush, R. (1998). Children' recollections of traumatic and nontraumatic events. *Development and Psychopathology, 10*, 699-716.

Flavell, J., Miller, P., & Miller, S. (2002). *Cognitive development, 4th Edition*. Upper Saddle River, NJ: Prentice Hall.

Fletcher, R., & Hattie, J. (2004). An examination of the psychometric properties of the physical self-description questionnaire using a polytomous item response model. *Psychology of Sport and Exercise, 5*, 423-446.

Fonseca-Pedrero, E., Wells, C., Paino, M., Lemos-Giraldez, S., Villazon-Garcia, U., Sierra, S.,…& Muniz, J. (2010). Measurement invariance of the Reynolds Depression Adolescent Scale across gender and age. *International Journal of Testing, 10*, 133-148.

Fornell, C., & Larcker, D.F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*, 39-50.

Friedman, W. (2007). The development of temporal metamemory. *Child Development, 78*, 1472-1491.

Garson, G. D. (2011). Factor analysis, from *Statnotes: Topics in Multivariate Analysis.* Retrieved 05/08/2011 from http://faculty.chass.ncsu.edu/garson/pa765/statnote.htm.

Ge, X., Conger, R., Elder, G. H., Jr., 2001. Pubertal transition, stressful life events, and the emergence of gender differences in adolescent depressive symptoms. *Developmental Psychology, 37*, 404-417.

Gomez, R. (2008). Item response theory analyses of the Parent and Teacher Ratings of the DIM-IV ADHD Rating Scale. *Journal of Abnormal Child Psychology, 36*, 865-885.

Grills-Taquechel, A., & Ollendick, T. (2008). Diagnostic interviewing. In M. Hersen & A. Gross (Eds.), *Handbook of clinical psychology, Volume 2: Children and adolescents* (pp. 458-479). Hoboken, NJ: John Wiley & Sons.

Hafsteinsson, L.G., Donovan, J.J., Breland, B.T. (2007). An item response theory examination of two popular goal orientation measures. *Educational and Psychological Measurement, 67*, 719-739.

Hall, T., Reise, S., & Haviland, M. (2007). An item response theory analysis of the Spiritual Assessment Inventory. *The International Journal for the Psychology of Religion, 17*, 157-178.

Hambleton, R., & Jones, R. (1993). comparsion of classical test theory and item response theory and their applications to test development. *Instructional Topics in Educational Measurement, Module 16*, 253-262.

Hammen, C., Bistricky, S., & Ingram, R. (2010). Vulnerability to depression in adulthood. In R. Ingram, & J. Price (Eds.), *Vulnerability to psychopathology: Risk across the lifespan (2nd ed.,* pp. 248-281). New York, NY: The Guilford Press.

Hastie, R. (1987). Information processing theory for the survey researcher. In H.-J. Hippler, N. Schwarz, & S. Sudman (Eds.), *Social information processing and survey methodology* (pp. 42-70). New York: Springer-Verlag.

Holaday, B., & Turner-Henson, A. (1989). Response effects in surveys with school-age children. *Nursing Research, 38*, 248-250.

Hu, L-t., & Bentler, P. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3*, 424-453.

Hudson, J., & Mayhew, E. (2009). The development of memory for recurring events. In M. Courage & N. Cowan (Eds.), *The development of memory in infancy and childhood* (pp. 69-91). New York: Psychology Press.

Jacobson, C.J., Farrell, J., Kashikar-Zuck, S., Seid, M., Verkamp, E., DeWitt, M. (2012). Disclosure and self-report of emotional, social, and physical health in children and adolescents with chronic pain – a qualitative study of PROMIS pediatric measures. *Journal of Pediatric Psychology, 38*, 82-93.

Judd, L., Akiskal, H., Maser, J., Zeller, P., Endicott, J., Coryell, M.,…Keller, M. (1998). Major depressive disorder: A prospective study of residual subthreshold depressive symptoms as predictor of rapid relapse [Special issue: George Winokur]. *Journal of Affective Disorders, 50*, 97-108.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist, 64*, 515-526.

Kihlstrom, J., Eich, E., Sandbrand, D., & Tobias, B. (2000). Emotion and memory: Implications for self-report. In A. Stone, J. Turkkan, C. Bachrach, J. Jobe, H. Kurtzman, & V. Cain (Eds.), *The science of self-report: Implications for research and practice* (pp. 81-103). Mahwah, NJ: Lawrence Erlbaum Associates.

Kovacs, M. (1992). The Children's Depression Inventory manual. North Tonawanda, NJ: Multihealth Systems.

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

Krosnick, J., & Fabrigar, L. (1997). Designing rating scales for effective measurement in surveys. In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, & D. Trewin (Eds.), *Survey measurement and process quality* (pp. 141-164). Hoboken, NJ: John Wiley & Sons.

Kiddie-Sads-Present and Lifetime Version (K-SADS-PL). Version 1.0 of October 1996. [http://www.wpic.pitt.edu/KSADS/ksads-pl.pdf].

Killip, S., Mahfoud, Z., & Pearce, K. (2004). What is an intracluster correlation coefficient? Crucial concepts for primary care researchers. *Annals of Family Medicine, 2*, 204-208.

Klein, D., (2008). Classification of depressive disorders in the DSM-V: Proposal for a two-dimension system. *Journal of Abnormal Psychology, 117*, 552-560.

La Greca, A. (1990). Issues and perspectives on the child assessment process. In A. La Greca (Ed.), *Through the eyes of the child: Obtaining self-reports from children and adolescents* (pp. 3-17). Boston, MA: Allyn and Bacon.

Levine, L., & Bluck, S. (2004). Painting with broad strokes: Happiness and the malleability of event memory. *Cognition and Emotion, 18*, 559-574.

Levine, L, & Pizarro, D. (2004). Emotion and memory research: A grumpy overview. *Social Cognition, 22*, 530-554.

Levine, L., Safer, M., & Lench, H. (2006). Remembering and misremembering emotions. In L. Sanna & E. Change (Eds.), *Judgments over time: The interplay of thoughts, feelings, and behaviors* (pp.271-290). New York: Oxford University Press.

Levine, R., & Huberman, M. (2002). *What types of survey items can elicit valid responses from fourth and eighth grade students?* U.S. Department of Education, National Center for Education Statistics.

Levine, R., Huberman, M., & Buckner, K. (2002). *The measurement of instructional background indicators: Cogntivie laboratory investigations of the responses of fourth and eighth grade students and teachers to questionnaire items.* U.S.

Department of Education Office of Educational Research and Improvement, National Center for Education Statistics.

Lorenzo-Blanco, E., Unger, J., Baezconde-Garbanati, L., Ritt-Olson, A., Soto, D. (2012). Acculturation, enculturation, and symptoms of depression in Hispanic youth: The roles of gender, Hispanic cultural values, and family functioning. *Journal of Youth and Adolescence, 41,* 1350-1365. DOI 10.1007/s10964-012-9774-7

Luby, J., & Belden, A. (2006). Mood disorders: Phenomenology and a developmental emotion reactivity model. In J. Luby (Ed.), *Handbook of preschool mental health: Development, disorders, and treatment* (pp. 209-230). New York: The Guilford Press.

Luby, J., Hefflinger, A. M., Mrakotsky, C., Hessler, M. J., Wallis, J. M., Spitznagel, E. L. (2003). The clinical picture of depression in preschool children. *Journal of the American Academy of Child and Adolescent Psychiatry, 42*, 340-348.

Lucas, C., Fisher, P., Piacentini, J., Zhang, H., Jensen, P., Shaffer, D., Dulcan, M., Schwab-Stone, M., Regier, D., & Canino, G. (1999). Features of interview questions associated with attenuation of symptom reports. *Journal of Abnormal Child Psychology, 6*, 429-437.

Marmorstein, N. (2012). Associations between dispositions to rash action and internalizing and externalizing symptoms in children. *Journal of Clinical Child & Adolescent Psychology, 42*, 131-138. DOI: 10.1080/15374416.2012.734021

Marsh, H. (1986). Negative item bias in ratings scales for preadolescent children: A cognitive-developmental phenomenon. *Developmental Psychology, 22*, 37-49.

Mash, E., & Hunsley, J. (2007). Assessment of child and family disturbance: A developmental-systems approach. In E. Mash & R. Barkley, (Eds.), *Assessment of childhood disorders, 4th Edition* (pp. 3-50). New York: The Guilford Press.

Matthey, S., & Petrovski, P. (2002). The Children's Depression Inventory: Error in cutoff scores for screening purposes. *Psychological Assessment 14*, 146-149.

McLaughlin, K., Hilt, L., Nolen-Hoeksema, S. (2007). Racial/ethnic differences in internalizing and externalizing symptoms in adolescents. *Journal of Abnormal Child Psychology, 35,* 801-816.

Meade, A. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *Journal of Applied Psychology, 95*, 728-743.

Millsap, R., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research 39*, 479-515.

Molina, c., Gomez, J., & Pastrana, M. (2009). Psychometric properties of the Spanish-language Child Depression Inventory with Hispanic children who are secondary victims of domestic violence. *Adolescence, 44*, 133-148.

Muthén, B. (2004). Mplus technical appendices. Los Angeles, CA: Muthén & Muthén.

Muthén, L. (2013, January 4). WRMR and multiple group models [Web log message]. Retrieved from http://www.statmodel.com/discussion/messages/11/11403.html?1357431506

Muthén, L., & Muthén, B. (2002). How to use a Monte Carlo simulation to decide on sample size and determine power. *Structural Equation Models 9*, 599-620.

Muthén, L., & Muthén, B. (2010). Mplus user's guide (6th ed.). Los Angeles, CA: Muthén & Muthén.

Olino, T., Yu, L., Klein, D., Rohde, P., Seeley, J., Pilkonis, P., & Lewinsohn, P. (2012). Measuring depression using item response theory: an examination of three measures of depressive symptomatology. *International Journal of Methods in Psychiatric Research, 21*, 76-85.

Otter, M., Mellenbergh, G., & de Glopper, K. (1995). The relation between information-processing variables and test-retest stability for questionnaire items. *Journal of Educational Measurement, 32*, 199-216.

Parkinson, B., Briner, R., Reynolds, S., & Totterdell, P. (1995). Time frames for mood: Relations between momentary and generalized ratings of affect. *Personality and Social Psychology Bulletin, 21*, 331-339.

Patton, G., Olsson, C., Bond, L., Toumbourou, J., Carlin, J., Hemphill, S., & Catalano, R. (2008). Predicting female depression across puberty: A two-nation longitudinal study. *Journal of the American Academy of Child and Adolescent Psychiatry, 47*, 1424-1432.

Paz-Alonso, P., Larson, R., Catelli, P., Alley, D., & Goodman, G. (2009). Memory development: Emotion, stress and trauma. In M. Courage & N. Cowan (Eds.), *The development of memory in infancy and childhood* (pp. 197-239). New York: Psychology Press.

Perez, R., Ascaso, L., Massons, J., & Chaparro, N. (1998). Characteristics of the subject and interview influencing the test-retest reliability of the Diagnostic Interview for Children and Adolescents-Revised. *Journal of Child and Adolescent Psychiatry, 39*, 963-972.

Purpura, D., & Lonigan, C. (2009). Conner's Teacher Rating Scale for Preschool Children: A revised, brief, age-specific measure. *Journal of Child and Adolescent Psychology, 38,* 263-272.

Raajimakers, A., van Hoof, A., t' Hart, H., Verbogt, T., & Vollebergh, W. (2000). Adolescents' midpoint responses on Likert-type scale items: Neutral or missing values? *International Journal of Public Opinion Research, 12*, 208-216.

Rauch, W., Schweizer, K., & Moosbrugger, H. (2008). An IRT analysis of the Personal Optimism Scale. *European Journal of Psychological Assessment, 24*, 49-56.

Rebok, G., Riley, A., Forrest, C., Starfield, B., Green, B., Robertson, J., & Tambor, E. (2001). Elementary school-aged children's reports of their health: A cognitive interviewing study. *Quality of Life Research, 10*, 59-70.

Reeve, B., & Mâsse, L. (2004). Item response theory modeling for questionnaire evaluation. In S. Presser, J. Rothgeb, M. Couper, J. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires* (pp.247-273). Hoboken, NJ: John Wiley & Sons, Inc.

Reise, S., & Henson, J. (2003). A discussion of modern versus traditional psychometrics as applied to personality assessment scales. *Journal of Personality Assessment, 81*(2), 93-103.

Robinson, M., & Clore, G. (2002a). Belief and feeling: Evidence for an accessibility model of emotional self-report. *Psychological Bulletin, 128*(6), 934-960.

Robinson, M., & Clore, G. (2002b). Episodic and semantic knowledge in emotional self-report: Evidence for two judgment processes. *Journal of Personality and Social Psychology, 83*, 198-215.

Rubie-Davies, S., & Hattie, J. (2012). The dangers of extreme positive responses in Likert scales administered to young children. *The International Journal of Educational and Psychological Assessment, 11*, 75-89.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34*(4, pt. 2).

SAS 9.2 [Apparatus and software]. (2008). Cary, NC: SAS Institute Inc.

Schimmack, U. (2002). Frequency judgments of emotions: The cognitive basis of personality assessment. In P. Sedelmeier & T. Betsch, *Etc.: Frequency processing and cognition* (pp. 189-204). Oxford: Oxford University Press.

Schwab-Stone, M., Fallon, T., Briggs, M., & Crowther, B. (1994). Reliability of diagnostic reporting for children aged 6-11 years: A test-retest study of the Diagnostic Interview Schedule for Children – Revised. *American Journal of Psychiatry, 151*, 1048-1054.

Schwarz, N. (1997). Questionnaire design: The rocky road from concepts to answers. In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, & D. Trewin (Eds.), *Survey measurement and process quality* (pp. 29-45). Hoboken, NJ: John Wiley & Sons.

Schwarz, N. (2007). Cognitive aspects of survey methodology. *Applied Cognitive Psychology, 21*, 277-287.

Scott, J. (1997). Children as respondents: Methods for improving data quality. In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, & D. Trewin (Eds.), *Survey measurement and process quality* (pp. 331-350). Hoboken, NJ: John Wiley & Sons.

Shaffer, D., Fisher, P, Lucas, C., Dulcan, M., & Schwab-Stone, M. (2000). NIMH Diagnostic Interview for Children Version IV (NIMH DISC-IV): Description, differences from previous versions, and reliability of some common diagnoses. *Journal of the American Academy of Child and Adolescent Psychiatry, 39*, 28-38.

Skowronski, J., Gibbons, J., Vogl, R., & Walker, W. R. (2004). The effect of social disclosure on the intensity of affect provoked by autobiographical memories. *Self and Identity, 3*, 285-309.

Stapleton, L., Sander, B., & Stark, K. (2007). Psychometric properties of the Beck Depression Inventory for Youth in a sample of girls. *Psychological Assessment, 19*, 230-235.

Stark, K., Sander, J., Hauser, M., Simpson, J., Schnoebelen, S., Glenn, R., & Molnar, J. (2006). Depressive disorders during childhood and adolescents. In E. Mash & R. Barkley (Eds.), *Treatment of childhood disorders (3rd edition;* pp. 336-407). New York: The Guilford Press.

Stefanek, E., Strohmeier, D., Fandrem, H., & Spiel, C. (2012). Depressive symptoms in native and immigrant adolescents: The role of critical life events and daily hassles. *Anxiety, Stress, & Coping, 25*, 201-217.

Stone, W., & Lemanek, K. (1990). Developmental issues in children's self-reports. In A. La Greca (Ed.), *Trhough the eyes of the child: Obtaining self-reports from children and adolescents* (pp.18-56). Boston, MA: Allyn and Bacon.

Sudman, S., Bradburn, N., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology*. San Francisco, CA: Jossey-Bass Publishers.

Stark, K., Sander, J., Hauser, M., Simpson, J., Schnoebelen, S., Glenn, R., & Molnar, J. (2006). Depressive disorders during childhood and adolescence. In E. Mash & R. Barkley (Eds.), *Treatment of childhood disorders, Third edition* (pp 336-407). New York: The Guilford Press.

Sterba, S., Egger, H., & Angold, A. (2007). Diagnostic specificity and nonspecificity in the dimensions of preschool psychopathology. *Journal of Child Psychology and Psychiatry, 48*, 1005-1013.

Tabachnik, B., & Fidell, L. (2000). *Using multivariate statistics (4th ed.)*. Boston, MA: Allyn & Bacon.

Talarico, J., LaBar, K., & Rubin, D. (2004). Emotional intensity predicts autobiographical memory experience. *Memory & Cognition, 32*, 1118-1132.

Thomas, D., & Diener, E. (1990). Memory accuracy in the recall of emotions. *Journal of Personality and Social Psychology, 59*, 291-297.

Tourangeau, R., & Rasinski, K. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin, 103*, 299-314.

Tourangeau, R., Rips, E. J., & Rasinski, K. (2000). *The psychology of survey response.* Cambridge: Cambridge Press.

Twenge, J., & Nolen-Hoeksema, S. (2002). Age, gender, race, socioeconomic status, and birth cohort differences on the Children's Depression Inventory: A meta-analysis. *Journal of Abnormal Psychology, 111*, 578-588.

U.S. Department of Commerce, United States Census Bureau (2012). Hispanic heritage month, September 15 – October 15. *Profile America, Facts for Features*, August 6, 2012.

van Beek, Y., Hessen, D., Hutteman, R., Verhulp, E., & van Leuven, M. (2012). Age and gender differences in depression across adolescence: Real or 'bias'? *Journal of Child Psychology and Psychiatry, 53*, 973-985.

Vandenberg, R., & Lance, C. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4-69.

van Laerhoven, H., van der Zaag-Loonen, H.J., Derkx, B.H.F. (2004). A comparison of LIkert scale and visual analogue scales as response options in children's questionnaires. *Acta Paediatrica, 93*, 830-835.

Verhoeven, M., Sawyer, M., & Spence, S. (2013). The factorial invariance of the CES-D during adolescence: Are symptom profiles for depression stable across gender and time? *Journal of Adolescence, 36*, 181-190.

Wanke, M., & Schwarz, N. (1997). Reducing question order effects: The operation of buffer items. In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, & D. Trewin (Eds.), *Survey measurement and process quality* (pp. 115-140). Hoboken, NJ: John Wiley & Sons.

Weems, G. (2004). Impact of the number of response categories on frequency scales. *Research in the Schools, 11*, 41-49.

Weems, G., Onwuegbuzie, A., & Collins, K. (2006). The role of reading comprehension in responses to positively and negatively worded items on rating scales. *Evaluation and Research in Education, 19*, 3-20.

Weiss, B., & Garber, J. (2003). Developmental differences in the phenomenology of depression. *Development and Psychopathology, 15*, 403-430.

Wilson, T., Meyers, J., & Gilbert, D. (2003). "How happy was I, anyway?" A retrospective impact bias. *Social Cognition, 21*, 421-446.

Winkielman, P., Knauper, B., & Schwarz, N. (1998). Looking back at anger: Reference periods change the interpretation of emotion frequency questions. *Journal of Personality and Social Psychology, 75*, 719-728.

Wirtz, D., Kruger, J., Scollon, C., & Diener, E. (2003). What to do on spring break? The role of predicted, on-line, and remembered experience in future choice. *Psychological Science, 14*, 520-524.

Woolley, M., Bowen, G., & Bowen, N. (2004). Cognitive pretesting and developmental validity of child self-report instruments: Theory and applications. *Research on Social Work Practice, 14*, 191-200.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125-145.

Yen, W. (1993). Scaling performance assessments: Strategies for managing local item dependence [Special issue: Performance assessment]. *Journal of Educational Measurement, 30*, 187-213.

Yorbik, Oz., Birmaher, B., Axelson, D., Williamson, D., & Neal, R. (2004). Clinical characteristics of depressive symptoms in children and adolescents with major depressive disorders. *Journal of Clinical Psychiatry, 65*, 1654-1659.

Yu, C.-Y. (2002). Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes. (Doctoral dissertation). Retrieved from http://www.statmodel.com/download/Yudissertation.pdf.

Appendix A

Analyses of Hispanic and Non-Hispanic Participants' Mean Scores and Data Factor Structures

This appendix details the results of the *t*-test and factor structure analyses conducted on the Hispanic and non-Hispanic groups' data.

A total of 1000 students reported that they were non-Hispanic, 655 reported being Hispanic, and data were missing for four students. Those four students were excluded from the following analyses.

Mean differences in BDI-Y and CDI scores. Two-tailed, independent samples *t*-tests of the Hispanic and not-Hispanic groups' total scores were performed for both the BDI-Y (20 items) and the CDI (27 items).

A Levene's test for equality of variances was significant for the BDI-Y, $F(1, 1653) = 15.64$, $p < .01$. The *t*-test results using modified degrees of freedom for the unequal variances was also significant, although the effect size was small, $t(1616.81) = 2.27$, $p = .02$, $d = 0.11$ (Cohen, 1988). The mean for the Hispanic group was 23.54. The mean for the non-Hispanic group was 28.17.

A Levene's test for equality of variances was not significant for the CDI, $F(1, 1653) = 0.09$, $p = .76$. The *t*-test was also not significant, $t(1653) = -.92$, $df = 1653$, $p = .36$, $d = 0.05$. The mean for the Hispanic group was 14.16. The mean for the non-Hispanic group was 13.42.

Confirmatory factor analyses. The Hispanic and non-Hispanic groups were included as separate groups in a single-factor confirmatory factor analysis to test whether both groups' data could be modeled with a single factor. The modeling procedures were

the same as those used in the measurement invariance analyses in the first component of Research Question 2. Those methods are explained more fully in that portion of the *Method* section.

As a very brief summary of the methods used, the weighted least squares with adjusted means and variances estimator (WLSMV) was again used because it is appropriate for categorical data (Muthén & Muthén, 2010). The model also included the item error covariances included in the Research Question 2 measurement invariance analyses (i.e., those in the first component of the questions. The two groups' loadings and thresholds were constrained to equality across groups, scale factors were set to 1 in one group and free in the other, factor means were set to 0 in one group and free in the other, factor variances were set to 1 in all groups, and the errors for conceptually related items were correlated.

Results from this fully constrained confirmatory factor analysis model showed good model fit according to the fit indices, although chi-square was significant: $\chi^2(379, N = 1655) = 866.56$, $p < .01$; RMSEA = .04; CFI = .98. Because of the model's good fit, a lack of evidence for different factor structures can be presumed, and the two groups can be treated as a single sample in subsequent analyses.

Appendix B

SAS Programming for IRT Parameter Generation

This syntax requires two data sets. The first is the set of each item's CFA factor loading and thresholds. Each item should be on its own line, loading listed first followed by the thresholds. In the example below, it is the first data set referred to. The second data set is the set of each participant's responses and their factor scores (which can be saved as part of the M*plus* OUTPUT command). Each participant should have their own line of data, responses first followed by their factor score.

```
%LET NI=5;                ** number of items**
%LET NC=4;                **number of item categories**
%LET J=3;                 ** number of thresholds**
%LET TT=15;               **(number of items)*(number of thresholds)**
%LET NS=366;              **number of participants**
%LET ESTIM=2;             **1 = FIML; 2 = WLS/WLSM/WLSMV**
%LET THETARANGE=3;        **range of the latent trait to be estimated**

DATA READ_EST;            **data set input directions for parameters**
INFILE 'C:\data\RQ2.1 c only bdi parameters 12.txt';
input load tau1-tau&J;
RUN;

DATA READ_SCORES;         **data set input directions for responses and fscores**
missing M;
INFILE 'C:\data\RQ2.1 cdi c only fscores 12.txt' dsd dlm='09'x missover;
INPUT Q1-Q&NI THETA;
ARRAY Q{&NI} Q1-Q&NI;
TOTAL=SUM (OF Q1-Q&NI);
SORT=1;
DO i=1 to &NI;
/*Q{i} = Q{i} + 1;*/
END;
RUN;

PROC PRINT; RUN;

OPTIONS PAGENO=1 NOCENTER NODATE;
```

```
DATA EST_ONE_REC;
SET READ_EST;
ARRAY TAU{&J} TAU1-TAU&J;
ARRAY L{&NI};
ARRAY T{&J, &NI};
RETAIN CNTR 1;
RETAIN L1-L&NI T1-T&TT;
DO I=1 TO &NI; IF CNTR = i THEN L{i}=LOAD; END;
DO k=1 TO &J;
DO I=1 TO &NI; IF CNTR = i THEN T{k,i}=TAU{k}; END;
END;
CNTR+1;
IF CNTR=&NI+1;
SORT=1;
RUN;



DATA TRANSFORMED_ESTIMATES;
SET EST_ONE_REC;
ARRAY LAMBDA{&NI} L1-L&NI;
ARRAY T{&J,&NI} T1-T&TT;
ARRAY DISCRIM{&NI};
ARRAY THR{&J,&NI};
FILE PRINT;
TITLE1 'REPORT #1: IRT ITEM DISCRIMINATION AND THRESHOLD
        ESTIMATES';
TITLE2 ' ';
TITLE3 ' ';
PUT @1 'Item' @8 'Discrimination' @25 'Thresholds'//;

DO i=1 TO &NI;
IF &ESTIM=1 THEN DO;
DISCRIM{i}=LAMBDA{i}*1/1.7;
DO k=1 TO &J; THR{k,i}=(T{k,i}/LAMBDA{i}); END;
END;
ELSE IF &ESTIM=2 THEN DO;
DISCRIM{i}=ROUND(LAMBDA{i}/SQRT(1-LAMBDA{i}**2),.001);
DO k=1 TO &J; THR{k,i}=ROUND((T{k,i}/LAMBDA{i}),.001);END;
END;
PUT @2 i @8 DISCRIM{i} @;
DO k=1 TO &j; PUT @(14+k*10) THR{k,i} @; END;
PUT /;
END;
RUN;
```

```
DATA THETA;
RETAIN THETA -&THETARANGE;
DO WHILE (THETA LT &THETARANGE..1);
SORT=1;
OUTPUT;
THETA=THETA+.1;
END;
RUN;

DATA ICC;
MERGE THETA TRANSFORMED_ESTIMATES; BY SORT;

ARRAY P{&J,&NI};
ARRAY PC{&NC,&NI};
ARRAY DISCRIM{&NI} DISCRIM1-DISCRIM&NI;
ARRAY THR{&J,&NI} THR1-THR&TT;

ARRAY DD{&NC,&NI};
ARRAY IC{&NC,&NI};
ARRAY INFO{&NI};
TINFO=0;
DO i=1 TO &NI;
DO k=1 TO &J; P{k,i}=(EXP(DISCRIM{i}*(THETA-
        THR{k,i})))/(1+(EXP(DISCRIM{i}*(THETA-THR{k,i})))); END;
PC{1,i}=1-P{1,i};
DO k=2 to &J; PC{k,i}=P{k-1,i}-P{k,i}; END;
PC{&NC,i}=P{&J,i};
DD{1,i}=-DISCRIM{i}*EXP(DISCRIM{i}*(THETA-
        THR{1,i}))/(1+EXP(DISCRIM{i}*(THETA-THR{1,i})))**2;
DO k=2 to &J; DD{k,i}=DISCRIM{i}*(P{k-1,i}*(1-P{k-1,i}))-(P{k,i}*(1-P{k,i}));
        END;

DD {&NC,i}=DISCRIM{i}*EXP(DISCRIM{i}*(THETA-
        THR{&J,i}))/(1+EXP(DISCRIM{i}*(THETA-THR{&J,i})))**2;
DO k=1 to &NC;
IF k=1 THEN INFO{i}=0;
IC{k,i}=DD{k,i}**2/PC{k,i};
INFO{i}=INFO{i}+IC{k,i};
END;
TINFO+INFO{i};
END;
RUN;

%MACRO ICCPLOT;
PROC PLOT DATA=ICC;
TITLE 'REPORT 6: OPTION RESPONSE FUNCTIONS';
```

```
TITLE2 ' ';
TITLE3 ' ';
PLOT PC&Q * THETA ='1';
PC&Q2 * THETA ='2';
PC&Q3 * THETA ='3';
PC&Q4 * THETA ='4'; /OVERLAY VAXIS=0 TO 1 BY .1;
QUIT;
%MEND;

%MACRO IIPLOT;
PROC PLOT DATA=ICC;
TITLE 'REPORT 7: ITEM INFORMATION CURVES';
TITLE2 ' ';
TITLE3 ' ';
PLOT INFO&Q * THETA ='*' / hAXIS= -3 TO 3 VAXIS = 0 TO 1;
QUIT;
PROC PRINT DATA=ICC;
RUN;
%MEND;

%MACRO TIPLOT;
PROC PLOT DATA=ICC;
TITLE 'REPORT 8: TEST INFORMATION CURVE';
TITLE2 ' ';
TITLE3 ' ';
PLOT TINFO * THETA = '*' / hAXIS= -3 TO 3;
QUIT;
%MEND;

%MACRO RUNPLOTS;
%DO Q=1 %TO &NI;
%DO K=1 %TO &NC;
%LET Q&K=%EVAL(&Q+(&K-1)*&NI);
%END;
%ICCPLOT;
%END;
%DO Q=1 %TO &NI;
%IIPLOT;
%END;
%TIPLOT;
%MEND;

%RUNPLOTS;
```

Appendix C

Matched Sample Statistical Tests of Differences in Test Information Curve Values

The original version of this study proposed that differences in the information test curve values would be tested using matched sample statistical tests across 61 points of the latent trait continuum (i.e., at 0.10 increments from -3 to +3). These tests were going to be conducted in Research Questions 1 and 2. However, multiple analyses were ineffective. The purpose of this appendix is to explain those analyses and demonstrate the reasons that similar analyses were not retained in the final study.

The explanations that follow include a discussion of: nonparametric matched sample tests used, matched sample *t*-tests used, and the distribution of the original and difference score data.

For both sets of item pairs, the information difference scores for each of the 61 latent trait points measured were first screened for normality using the Shapiro-Wilk test. The test was significant for both sets of item pairs: for the items matched for content and frequency, $W(N = 61) = .90$, $p < .01$; for the items matched for content only, $W(N = 61) = .90$, $p < .01$.

Because the difference scores were not normally distributed, they were first analyzed using a Wilcoxon Signed-Rank test. The difference between the BDI-Y and CDI information levels in the item set matched for content and frequency was significant, $Z = 521.50$, $p < .01$. The difference between the information levels in the item set matched for content only was also significant, $Z = 945.50$, $p < .01$. The procedures were repeated four more times for Research Question 2 (i.e., for the same two sets of items, but independently for each of the 9- and 12-year-old groups). The results from those analyses

showed that the data were also non-normally distributed and the tests of the difference scores were also highly significant: the two test values in the 9-year-old group were 945.50, and the two test values in the 12-year-old group were 370.50 and 929.50; all were significant at the <.01 level. The test findings indicated problems with the tests for two reasons. First, the high significance values seemed inconsistent with the visual depictions provided in Figures 8 and 9—the similar curves in Figure 8 and dissimilar curves in Figure 9 suggested that the difference tests for the Figure 8 data would be non-significant while the difference tests for the data in Figure 9 would be significant. Second, the test values in Research Question 2 analyses for the 12-year-old group were smaller than the test values for the 9-year-old group, which is inconsistent with the larger information value differences for the 12-year-old group than the 9-year-old group depicted in Figures 18 through 21 (in the *Results* section).

Reasoning that *t*-tests can be relatively robust to violations of non-normality but potentially less sensitive to sample differences, matched-sample *t*-tests were also conducted. The pattern of findings was similar to the pattern of findings when the Wilcoxon Signed-Rank test was used. For Research Question 1, *t*-test values were all significant at levels of <.01. When the absolute differences for Figure 8 were tested, the associated test statistic and effect size were larger than the other two tests. The test values in Research Question 2 were also all significant at levels of <.01. Inconsistent with the visuals provided in Figures 18 through 21, the effect sizes and test values in Research Question 2 were larger for the 9-year-old group than the 12-year-old group.

An examination of the data distribution patterns and the standard deviation formulas revealed that the reason for the incongruities between the test values and the

figures is the shape of the original data distributions. To illustrate the discussion, the information curves from Research Question 1 (Figures 8 and 9) are presented again here:

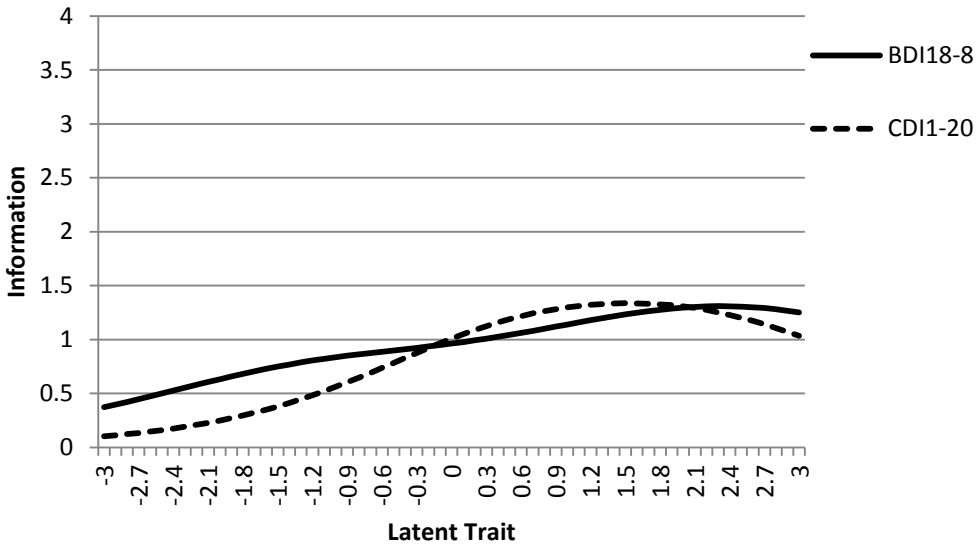

Figure 8. Information plot of items matched for **content and frequency**.

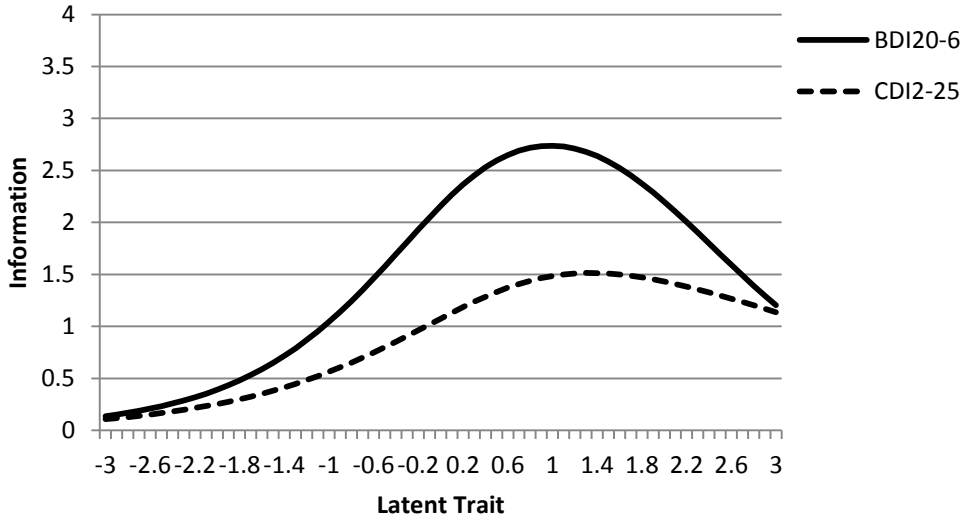

Figure 9. Information plot of items matched for **content only**.

Note that in Figure 8, the curve lines are nearly parallel to each other, while the curves in Figure 9 reach a much larger maximum difference. Stated differently, the values for both curves in Figure 8 have a limited range and therefore vary little from the means for each curve. In Figure 9, the values for the BDI-Y have a much greater range and therefore vary from the mean much more than the values in Figure 8.

According to the formula for a matched-sample statistical test:

$$t = \frac{\sum d}{\sqrt{\frac{n(\sum d^2) - (\sum d)^2}{n-1}}} \qquad [C1]$$

In this formula, *d* refers to difference scores between the two sets of data. The numerator represents the sum of the difference scores, while the denominator represents the standard deviation of those scores. As the standard deviation of a set of scores increases, the *t*-test value decreases. The difference scores for Figures 8 and 9 are presented in Figures C1 and C2, respectively.

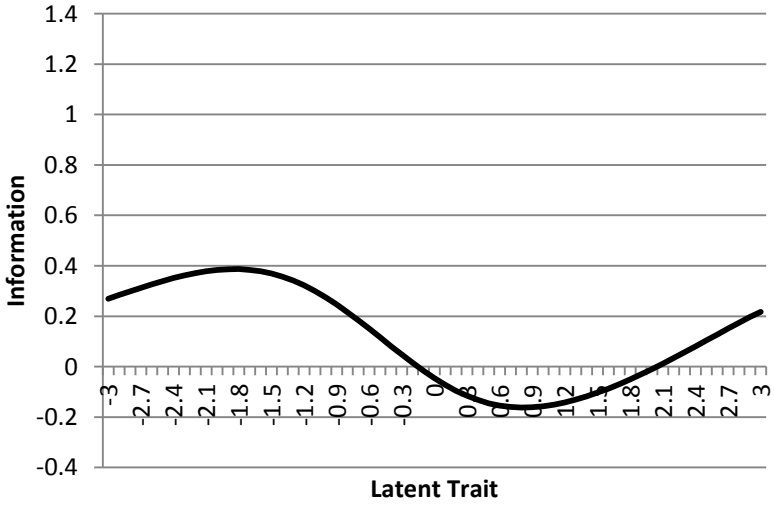

Figure C1. Difference scores for items matched for **content and frequency**

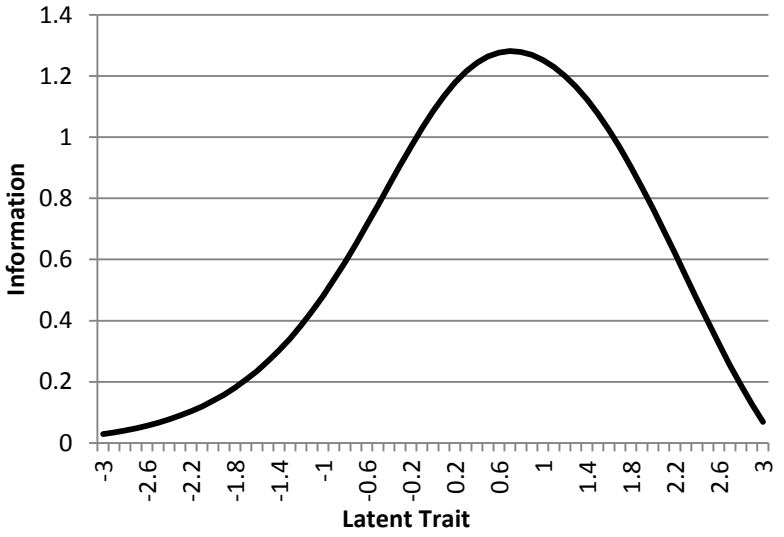

Figure C2. Difference scores for items matched for **content only**

As can be seen in these figures, the difference scores in Figure C1 vary less than the difference scores in Figure C2. As a result, the standard deviation for the Figure C2 scores will be larger than the standard deviation for the C1 scores. The associated test value for Figure C1 will therefore be higher than the test value for Figure C2, even though the differences associated with C2 are greater overall.

The use of a pooled standard deviation was also considered as a possible way of avoiding the inflated standard deviation values associated with Figure C2:

$$s_p^2 = \frac{\sum_{i=1}^{k}(n_1-1)s_i^2}{\sum_{i=1}^{k} n_i-1} \qquad [C2]$$

Formula C2 states that for a set of items (k), the numerator is the sum of each original item set's (i.e., as opposed to the difference scores) weighted standard deviation, where the weight is the number of items in the set minus 1 (n-*1*). That sum of weighted standard deviations is then divided by the sum of each item's n-*1*. As with the non-pooled standard deviation formula, however, greater variance in a set of scores results in greater standard deviations. Because the original data in this study also cover a wide range of values,

using a pooled standard deviation did not eliminate the problem of increased standard deviations for curves with higher maximum information levels.

Because any test based on tests of differences from the mean would face similar problems, further statistical analyses were not pursued. Alternatives to such statistical tests were considered, such as the differential item functioning (DIF) work performed by such authors as Meade (2010), whose work was used in this study in Research Question 2. Meade reviewed the work of a number of authors, and presented a set of DIF statistics that summarized those earlier works. These statistics were designed by their authors to analyze differences between two groups' parameter estimates for a single set of items. Although using those tests to compare parameters across two sets of items was considered (i.e., by entering BDI-Y and CDI parameters into the model instead of two groups' parameters), the test calculations require a consistent number of parameter estimates (i.e., threshold) in both sets of item responses. Thus, analyses using these statistics were also not possible.

In the end this unsuccessful line of inquiry points to the need for further research to quantify and evaluate differences in item and test information values. The current study was forced to rely on the visual analyses that are commonly seen in IRT studies of item and scale functioning (e.g., Carle, Millsap, & Cole, 2008; Olino et al., 2012; Pupura & Lonigan, 2009; Rauch, Schweizer, & Moosbrugger, 2008)