Validation of Score Interpretations for the

BDI-2 Using Rasch Methodology

BY

ERICA MICHELLE LAFORTE
B.S., University of Wisconsin-Madison, 1995
M.A., Loyola University-Chicago, 2002

DISSERTATION

Submitted as partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Educational Psychology
in the Graduate College of the
University of Illinois at Chicago, 2014

Chicago, Illinois

Defense Committee:

Everett Smith, Jr., Chair and Advisor
Carol Myford
Daniel Maggin, Special Education
Fredrick Schrank, The Woodcock-Muñoz Foundation
Mark Ledbetter, Houghton Mifflin Harcourt

DEDICATION

This dissertation is dedicated to my beautiful daughters, Audrey and Lucy. My wish for you both is that you will have the imagination to dream big dreams, and the courage to make those dreams come true. I love you both.

ACKNOWLEDGMENTS

So many people have supported and encouraged me throughout this long journey. First and foremost, I give thanks to God, for it is only through His grace that all things are possible.

I would like to acknowledge and thank my advisor and committee chair, Dr. Everett Smith, for his never-ending patience, encouragement, and advice. Many advisors would likely have given up on a student who took as long as I did to persevere through a dissertation. Everett did not give up on me, and I know that I could not have completed this dissertation without his guidance.

My husband, George, stood by me every step of the way with saint-like patience. He never complained about the cost of dragging my graduate school career over many years, but always just quietly encouraged me to do the best I could do. He spent many evenings and weekends single parenting without complaint while I was locked away in my office, researching, analyzing, and writing. Most important, he never lost faith that I would someday actually get to this point. There is no way that I could have completed this project without his love and support.

My parents always encouraged me to pursue my educational goals from my early years, and their support and occasional "nudging" helped me to remember that this dissertation represents the achievement of a goal that I had from the time I was a little girl. I am so blessed to have parents who encouraged me to dream big.

My mom and Jerry deserve a special thank you for their gift of time during the last push to finish. They swooped in and rescued me when I needed it most, and I am so grateful.

ACKNOWLEDGMENTS (continued)

I am fortunate to have had many wonderful mentors and cheerleaders in my professional life. Dr. Fredrick Schrank first encouraged me to pursue graduate studies and has played a role in several of my professional milestones. At various times throughout my career, Fred has mentored me and provided me with opportunities, guidance, and support. He is the consummate professional and I will always strive to follow his example.

Several other colleagues have supported and encouraged me on at various times during my years as a graduate student: Melanie Bartels Graw, Dr. Kevin McGrew, David Dailey, Mary Ruef, Tommie Royce, Sharon Frey, Michael Custer, Dr. William Insko, and Dr. Mary Gevorkian. I learned from each of you and am grateful for your support and encouragement.

Johnna Gueorguieva has become a dear friend over the last several years, and her own perseverance to finish her Ph.D. has been an inspiration to me. Johnna is always willing to help me think through psychometric questions or share a bottle of wine (and sometimes both).

Jan Nickels's long-time friendship has seen me through my entire graduate career. Jan never once doubted that I would finish this degree. She generously shared her time and talent by editing an early version of my paper, and her confidence in me has been a source of strength to see this project through.

I will never forget the support and encouragement of Jean Newborg, author of the BDI-2. I am indebted to her for allowing me to be a part of the BDI-2 experience and for her blessing to pursue this research. I would also like to acknowledge and thank Riverside Publishing for permitting me to use the BDI-2 standardization dataset.

Elise Wilson at the UIC College of Education office was so helpful in walking me through the paperwork and requirements. Her patience and willingness to go out of her way to assist me has been a blessing.

Finally, I would like to express my sincere gratitude to my other dissertation committee members: Dr. Carol Myford, Dr. Mark Ledbetter, and Dr. Daniel Maggin. I appreciate your support, patience, and feedback. I am blessed to have had such an amazing group of professionals on my side.

EML

DISCLAIMER

I was employed by The Riverside Publishing Company—the publisher of the Battelle Developmental Inventory, Second Edition—from 1996 through 2010. During the period from 2000 through 2002 when the BDI-2 was in development, I served as Clinical Project Director and oversaw the early stages of product revision, including new item development and pilot testing, examiner recruitment and training, and planning for the standardization study. While I was not responsible for the BDI-2 development after the outset of the standardization study in 2002, I continued to support the project through 2010 by answering customer questions and assisting with planning and implementation of the web-based scoring system currently in use.

Through my affiliation with Riverside I have a unique relationship to the BDI-2. I was forthcoming with Riverside about my plans for this research; in fact, because the company believes that my research would provide important information to either support or refute the interpretations of the BDI-2 scores for use in identification, intervention, and reporting, Riverside granted me written permission to use the standardization dataset for this study (with all identifying information removed). The Examiner's Manual that accompanies the BDI-2 contains industry-standard technical information supporting use of the test scores for the stated purposes of the test. My purpose in conducting the analyses contained in this study is not to support or refute the validity information that has already been gathered and reported in the Examiner's Manual; rather, my objective is to add to the existing body of validity evidence by conducting additional analyses using a different validity framework and measurement model. Additionally, the information I gathered during this study will be shared with Riverside Publishing so that it may be considered if the test is revised in the future.

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

TABLE OF CONTENTS (continued)

LIST OF TABLES

LIST OF TABLES (continued)

LIST OF FIGURES

LIST OF ABBREVIATIONS

-2LL	-2 log-likelihood ratio
AE	Age equivalent
APR	Annual Performance Report
BDI	*Battelle Developmental Inventory*
BDI-2	*Battelle Developmental Inventory, Second Edition*
COSF	Child Outcomes Summary Form
CRT	Criterion-referenced test
CSS	Change-Sensitive Score
CTT	Classical Test Theory
DIF	Differential item functioning
ECO	Early Childhood Outcomes Center
FAPE	Free and appropriate public education
IDEA	Individuals with Disabilities Education Act of 1997
IDEA 2004	Individuals with Disabilities Education Improvement Act of 2004
MFRM	Many-facet Rasch measurement
NCLB	No Child Left Behind Act of 2001
NICHD	National Institute of Child Health and Human Development
NIH	National Institutes of Health
NRT	Norm-referenced test
OSEP	Office of Special Education Programs of the U.S. Department of Education
PCA	Principal components analysis
PLD	Performance-level descriptor

LIST OF ABBREVIATIONS (continued)

RDI	Relative Developmental Index
RMI	Relative Mastery Index
RPI	Relative Proficiency Index
SD	standard deviation

SUMMARY

Research has shown the positive impacts of early intervention for children who experience developmental delay. Recent legislation mandates early intervention for young children who have delays in the areas of cognitive, social and emotional, motor, speech and language, or adaptive skills development. Several challenges exist for professionals tasked with identifying children with developmental delay, designing intervention programs, and tracking the progress of the children who receive intervention services. Among these challenges are differences among states and jurisdictions in how developmental delay is defined, inconsistency in the content coverage and technical quality of the instruments available for assessing children, and ambiguity in the federally mandated reporting requirements.

The Battelle Developmental Inventory, Second Edition (BDI-2) is a measure of early childhood development that can provide a psychometrically sound solution to several of the challenges facing early childhood educators. In this study, I use the validity framework proposed by Wolfe and Smith (2007) and Rasch measurement analyses to gather evidence relevant to the structural, substantive, and generalizability aspects of validity for the BDI-2 Gross Motor subdomain scores. I first suggest several propositions which, if true, would provide support for the current uses of the BDI-2 test scores. For each proposition, I pose one or more research questions to guide my analyses. I utilize the BDI-2 standardization dataset for these analyses.

The results of my analyses provide evidence to support the structural and generalizability aspects of validity for the BDI-2 Gross Motor subdomain scores. The Rasch model assumptions of undimensionality and local independence are met. The item and examinee separation indices and separation reliabilities are high. The evidence I gathered relevant to the substantive aspect of validity suggests that examiners may not have used the three-category BDI-2 scoring system as

the test developer intended; however, an optimized two-category scoring system produced an examinee ability rank order that was nearly identical to the examinee ability rank order from the three-category scoring system. Additionally, I identified some anomalous examinee score strings in the dataset. Removal of these unexpected scores did not impact the rank-order of the item difficulty measures.

I. INTRODUCTION

A. Background and Statement of the Problem

The twentieth century saw the emergence of an entirely new field within education, devoted to the widespread and systematic testing of children; however, only since the passage of Public Law 94-142, the Education for All Handicapped Children Act of 1975, has there been an increased and focused interest in systematically identifying young children with developmental delays. Public Law 94-142 and its subsequent amendments and reauthorizations mandate identification of and early intervention services for young children who have delays in the areas of cognitive, social and emotional, motor, speech and language, or adaptive skills development, or who are at risk of developmental delay due to a medical condition or environmental factors. The current reauthorization of this law, the Individuals with Disabilities Education Improvement Act of 2004 (IDEA, 2004), contains Part C, Infants and Toddlers with Disabilities, which authorizes states to "develop and implement a statewide, comprehensive, coordinated, multidisciplinary, interagency system that provides early intervention services for infants and toddlers with disabilities and their families." Public Law 94-142 and its reauthorizations have driven much of the development in early childhood assessment for the past 35 years. After each reauthorization of the law, test developers have responded by tailoring their assessments to conform to the new requirements.

However, during this period of growth, changes in early childhood tests have often been reactionary. When laws change, test developers must quickly provide early childhood educators with tests that will comply with all the necessary eligibility and reporting requirements. Additionally, because states are allowed to establish and maintain their own Part C programs,

there is little agreement between states—or sometimes even within a state—regarding which assessments should be used and how those scores should be interpreted.

To illustrate these problems I will present two examples of challenges that have arisen from IDEA requirements. I will use these two examples to demonstrate the need for an early childhood developmental assessment that provides practitioners with valid information for decision making and reporting. This need leads to the research questions that have driven the present study.

1. Challenge #1: Defining Developmental Delay under IDEA

A national survey conducted in 2001 revealed that 62% of the children served by early intervention programs under IDEA Part C were eligible because of developmental delay (Scarborough et al., 2004). Under IDEA, individual states are responsible for determining how they will define developmental delay. The law allows states to choose which, if any, instruments they will use to make eligibility decisions and what level of functioning they will view as constituting developmental delay. Thus the definitions vary greatly from state to state. Some states have chosen to define delay as the difference between a child's chronological age and actual performance on a norm-referenced test, expressed as either a percentage (e.g., Maryland uses 25% delay in one or more developmental areas) or an absolute difference (e.g., Texas uses a graduated scale requiring a two-month, three-month, or four-month delay, depending on the chronological age of the child). Other states use norm-referenced cut scores (e.g., Nebraska uses 2.0 standard deviations below the mean in one developmental area or 1.3 standard deviations below the mean in two or more areas). Still other states (e.g., California, Colorado, Hawaii, Vermont, and West Virginia) allow the determination of delay to be made *only* by the qualitative "informed clinical opinion" of an early childhood professional, such as a pediatrician or

developmental specialist, or by a multidisciplinary team composed of parents and early childhood professionals (Shackelford, 2006). Most often, states allow practitioners a choice of two or more of these methods for determining eligibility. I outline each of these methods and discuss the consequences of each one in the next section.

a. Methods of Determining Delay

Of the four methods of determining delay identified by Shackelford (2006), three are based on norm-referenced inferences. These include the percent delay, absolute difference, and standard deviation methods. The fourth commonly used method, informed clinical opinion, relies on one or more professionals' judgment about whether a child's skills are developing on a trajectory within the range of what is considered typical.

i. Percent Delay Method

As of 2007, 37 of the 50 U.S. states allowed early childhood professionals to make an eligibility decision based on a child's percentage of delay. The percentage delay required for developmental delay diagnosis ranges from 15% to 50%, depending on the state (Shackelford, 2006). Some states require a higher percentage delay if the delay occurs in only one developmental domain rather than in multiple domains. Percent delay is calculated as

$$\frac{(\text{Chronological age in months}) - (\text{Age-equivalent score in months})}{(\text{Chronological age in months})} \times 100.$$

Practitioners obtain the age-equivalent score through the use of a norm-referenced developmental assessment. In most cases, the age-equivalent score is equal to the median score obtained by children of the same age in the norming sample. For example, if the median score

for all children 3 years, 6 months old (3-6) in the norming sample of an assessment is 23, then a child of any age who obtains a score of 23 on that assessment is said to have an age-equivalent score of 3-6.

While this method of determining delay may seem straightforward, it does have some problems. First, percent delay does not have the same meaning across the age span, as shown in Figure 1.

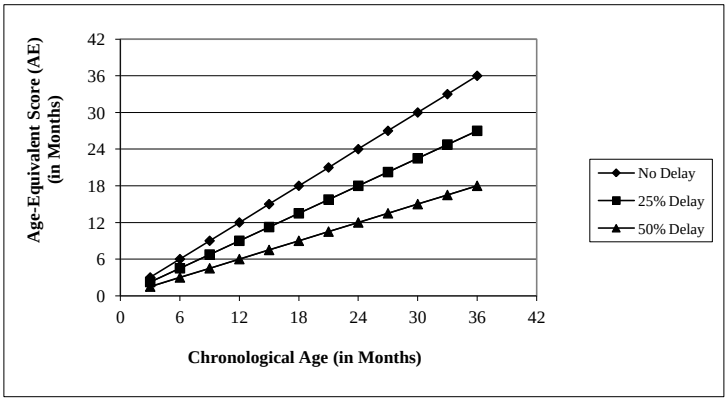

Figure 1. Age-equivalent scores necessary for 25% and 50% delay.

Note that as chronological age increases, so does the absolute number of months of delay required to reach the 25% and 50% delay thresholds, as the diverging trajectories of these two lines in the figure reveal. Thus, the older the child is, the more months of delay are required to qualify for eligibility under IDEA.

Also, this method does not take into account the measurement error associated with test scores. For a very young child, passing or failing one or two test items could make a significant difference in the resulting age-equivalent score. Because the absolute number of months of delay necessary for a younger child to reach a 25% delay threshold is so small—for example, just one-and-a-half months delay for a 6-month-old child—measurement error could indicate the presence of a significant delay when in fact there is none, or it could mask a delay that actually exists. Additionally, for children younger than 1 year old, the percent delay method is highly susceptible to fluctuations in the testing date. For instance, an examiner who administers a norm-referenced test to a baby girl who is exactly 5 months old and assigns an age-equivalent score of 3 months would calculate a 40% delay using the formula shown above. However, if the examiner had tested the child one day earlier—when she was still in her fourth month—the delay would have been only 25%. Finally, not all assessment instruments provide age-equivalent scores; therefore, in states that require the use of the percent delay method, diagnosticians have a limited selection of instruments available to them.

ii. Absolute Difference Method

Currently, several states employ what I will term the "absolute difference" method of determining developmental delay. According to this method, delay is defined as a minimum number of months of difference between a child's chronological age and the age-equivalent test score. States that use this method usually do so in conjunction with the percent delay method, and the less stringent of the two methods is the determining factor in the eligibility decision. However, one state (Texas) uses only the absolute difference method. To qualify for services in this state, children younger than 12 months need to show a 2-month delay, while 1-year-old children require a 3-month delay, and 2-year-old children require a 4-month

delay. One drawback of this method is the discrepancy in the relative delay at the tail ends of each age category. For instance, a two-month delay is much more significant for a 3-month-old child than for an 11-month-old child. Additionally, the eligibility decision for a child on the cusp of the next age category can be greatly impacted. For example, a child just short of his or her first birthday would qualify for intervention services with 18% delay (an absolute delay of two months), while that same child, if tested the very next day, would need a 25% delay (an absolute delay of three months) to qualify.

iii. Standard Deviation Method

Currently, 24 states allow use of the standard deviation (*SD*) method for determining developmental delay, often in conjunction with another method. In the *SD* method, a child can qualify for services if his or her score on a norm-referenced developmental assessment falls below a certain state-specified *SD* threshold. The most commonly used thresholds are 1, 1.5, or 2 *SD*s below the mean. As in the percent delay method, the requirements are often more stringent for delay in only one developmental domain, as opposed to two or more domains. An advantage of the *SD* method is that, theoretically, the likelihood of qualifying for services is equal across all ages within a state because, in a normal distribution of test scores, about 16% of children in each age cohort fall at or below –1 *SD*, and about 2% of children fall at or below –2 *SD*s. One political advantage of using the *SD* method is that it allows states to better estimate the number of children tested who may actually qualify for services in a given period. A disadvantage of this method is that, for populations that differ dramatically from the norm-referenced sample, a much larger-than-expected or much smaller-than-expected percentage of children may actually meet the criterion for delay.

iv. Informed Clinical Opinion Method

Thirty-four states allow the use of "informed clinical opinion" to determine developmental delay. In most cases, states offer this method as an alternative to one of the methods described previously when the use of a standardized test is not feasible. Some states require that a multidisciplinary team made up of education and health professionals as well as the child's parent(s) must make the eligibility decision. Other states require only that an individual with specified credentials or certifications, such as a physician or certified occupational, physical, or speech therapist, must make the judgment.

b. Problems with How Developmental Delay Is Currently Defined

One problem with the current system of state jurisdiction over IDEA Part C is that the eligibility requirements vary significantly from state to state (Shackelford, 2006). A child who qualifies for early intervention services in one state may not qualify in another. The differences between some states' eligibility requirements are so great that a child who qualifies as "not delayed" in one state could actually be considered "significantly delayed" in a neighboring state. For example, if a child scored 1.5 standard deviations below the mean on a measure of communication development, he or she may be considered delayed in a state with a cutoff of –1 *SD*. But if that child moved to another state with an eligibility requirement of –2 *SD*, he or she may no longer qualify for early intervention services. Because eligibility requirements have an impact on the number of children requiring services, some state agencies regularly revise their requirements in response to shifting budget constraints (Shackelford, 2006).

Even within states, diagnosticians may use any one of a number of assessment instruments to determine if a child is eligible for early intervention services, and they often use these instruments interchangeably. For example, the state of Illinois allows diagnosticians to use

any one of 12 separate instruments to assess communication skills development. While each approved instrument is designed to measure communication skills, the instruments are not equal in their technical quality, age of norms, or specific skills measured. Even different versions of the same test can yield dramatically different results. For example, Van Den Wymelenberg, Deitz, Wendel, and Kartin (2006) found that, in up to 43% of eligibility decisions, conflicts arose between results from the first and second editions of the Peabody Developmental Motor Scales, a test frequently used to make eligibility decisions in the motor domain.

Perhaps the most important shortcoming of the current methods of determining delay is that norm-referenced diagnostics alone do not provide information to early childhood diagnosticians and educators about the child's present *skills and abilities*. Standard scores, percentile ranks, and age-equivalent scores all compare a child's performance to the performance of his or her same-age peers, but these metrics do not provide information about which skills a child has mastered and which ones are only just emerging—information that is essential for early childhood professionals planning the types of intervention strategies that they will employ with the child.

The last of the four methods discussed above—informed clinical opinion—may provide information about the child's skills and abilities through required documentation, but in this case the child's present level of functioning cannot be referenced to the level of typically developing children. One professional's definition of "delayed" may not be the same as that of another professional, and the population whom each professional encounters most frequently is likely to influence his or her definition.

2. Challenge #2: Annual Progress Reporting

In 2003, the Office of Special Education Programs (OSEP) of the U.S. Department of Education provided a 5-year grant to the Early Childhood Outcomes Center (ECO) at the University of North Carolina to assist with developing a reporting system for states to monitor the progress of infants and toddlers with identified disabilities. OSEP initiated the effort in anticipation of new reporting requirements that took effect in 2007. Under these requirements, OSEP requires states to provide yearly documentation that the children whom their IDEA Part C programs serve are, in fact, showing an increased developmental trajectory. In this Annual Performance Report (APR), states must divide the population of infants and toddlers who exited early intervention services during the previous year into the following five categories or "buckets": (a) the percentage of children who did not improve functioning, (b) the percentage of children who improved functioning but did not show sufficient improvement to move nearer to the functioning level of same-age peers, (c) the percentage of children who improved functioning so as to approach more nearly the level of same-age peers but did not reach it, (d) the percentage of children who improved functioning sufficiently to reach a level comparable to same-age peers, and (e) the percentage of children who maintained functioning at a level comparable to same-age peers (IDEA, 2004). This report requires the aggregation of information obtained at the level of the individual child across two points in time. It thus poses two challenges: how to describe the child's developmental status at entry and exit, and how to describe the type and amount of change represented by the difference between entry and exit.

a. Describing Developmental Status

OSEP does not provide specific guidance to states as to how a child's developmental level should be assessed at each of the time points, other than to say that the

statement of present developmental level must be "based on objective criteria" (34 CFR §303.344(a)(1) through (2)). To assist states with this assessment, the ECO designed the Child Outcomes Summary Form (COSF), a 7-point rating scale for use by early childhood educators who need to determine a child's current developmental status. The ECO provides instructions to practitioners on how to complete the form using qualitative information about the child's developmental status. Alternatively, COSF users may translate the results from another early childhood developmental battery, using *z* scores, into COSF ratings of 1 through 7, with 6 and 7 representing "functioning comparable to same-aged peers" (*z* scores of –1.30 and above). The lower part of the *z*-score distribution (*z* scores of –1.31 and below) is broken down into the five remaining categories. The ECO describes four of these seven categories using labels indicating the degree of delay that the child exhibits; the intermediate three categories do not have labels.

The ECO and OSEP support use of the COSF instrument by states to comply with their annual progress reporting requirements. OSEP has acknowledged that tracking the progress of a group of children can be difficult when states use multiple instruments to assess those children's developmental status at entry and exit. According to the ECO, the COSF "can be used when different assessment instruments have been given to different children across the state and the results need to be placed on the same scale to be aggregated" (Early Childhood Technical Assistance Center, 2006, p. 1).

While the COSF may appear to laypersons to be an adequate solution to the challenge of describing a child's developmental functioning for purposes of reporting, it has serious psychometric shortcomings. Most importantly, the COSF does not address—and can seriously mask—the differences in the assessment instruments that practitioners may use to derive the COSF ratings. To derive a COSF rating from another instrument that is aligned to the OSEP

outcomes, one simply uses a *z*-score interpretation from the administration of the other instrument. However, assessment instruments are not all equivalent in terms of depth and breadth of content coverage, intended uses, technical quality, or alignment with the OSEP's three outcome areas. Placing *z*-scores from different instruments onto one 7-point scale implies that the ratings are equivalent, when in reality they are not. Additionally, although the ECO states that "The COSF process supports multidisciplinary 'best practices' in early childhood assessment and is consistent with the approach promoted by numerous professional organizations" (Enhance, 2013), there is currently no documented evidence to support the use of the COSF ratings for the purpose of classifying children's levels of development. The ECO acknowledges that "ideally, [validity] information would have been available before the instrument was released, but the OSEP reporting timeline did not allow for this" (ECO, 2012, p. 3). Despite a lack of evidence that the COSF produces valid and reliable information about children's levels of development, the ECO and OSEP are touting it as a solution to the OSEP reporting challenge; as of August 2012, 37 states were using the COSF to make determinations about the developmental status of the children entering and exiting their early intervention programs (Enhance, 2012).

b. Describing Progress from Entry to Exit

The five mutually exclusive OSEP reporting categories do not contain any information about what skills a child must be able to demonstrate to be considered "comparable to same age peers," nor do they provide any guidance on the amount and type of score change that is necessary for a child to demonstrate "improved functioning." The category descriptions that OSEP has chosen for the five reporting categories imply that the interpretations of "progress" should be based on norm-referenced scores. For example, the category (c) description "Improved functioning

to a level nearer to same-age peers" might imply that a child must exhibit an increase in standard score or percentile rank on a norm-referenced assessment from entry to exit, but that the exit score must still place the child below the mean score of his or her same-age peers. Similarly, the category (d) description "Improved functioning to reach a level comparable to same-age peers" might suggest that a child needs not only to exhibit an increase in standard score or percentile rank from entry to exit, but also to achieve a standard score or percentile rank at exit that would place the child within the "average" range for his or her age. Although these norm-referenced interpretations seem implicit in the use of the phrase "comparable to same-age peers," the instructions for compilation of the OSEP report do not explicitly require use of norm-referenced scores to make determinations of progress. In fact, practitioners in states using only the COSF as a stand-alone rating tool do not even have access to norm-referenced scores, because there are no existing norms for the COSF. For these states, the individuals or groups tasked with completing a COSF rating for each child must make the determination as to whether the child's function is comparable to that of his or her same-age peers based on their professional judgment.

3. Summary of the Problem

We have seen that numerous challenges face practitioners, policymakers, and local and state agencies using the results of early childhood assessment instruments to make decisions about delay identification and progress reporting. For those tasked with determining children's eligibility for early childhood intervention services, the challenges arise from inconsistency among jurisdictions in how delay is defined, inconsistency in the content coverage and technical quality of the available instruments, and the lack of useful information for intervention planning that can be gleaned from most norm-referenced score interpretations. As a result, the term "developmental delay" may have very different operational definitions,

depending on where a child lives. Even for children living within the same state, a diagnosis of a child's developmental status as delayed may depend on the particular instrument employed for the assessment, the background and experience of the individuals who participate in the multidisciplinary diagnostic team, and the amount of funding that the state has available to serve children with disabilities.

For the state agencies responsible for reporting on the progress of children served by early intervention programs, the challenges arise from ambiguity in how the reporting categories are operationalized and from the widespread use of COSF scores, which have not been subjected to the minimum industry-standard psychometric review and validation procedures. There is even inconsistency as to the derivation of COSF scores: some states use the COSF instrument as a stand-alone measurement tool, while other states use a published assessment battery to determine *z*-scores and then translate those *z*-scores into COSF ratings (using the procedure described earlier) for use on OSEP reports.

B. The Battelle Developmental Inventory, Second Edition

The Battelle Developmental Inventory, Second Edition (BDI-2) (Newborg, 2005a) presents a psychometrically sound solution to many of these early childhood assessment challenges. The BDI-2 is a standardized, early childhood developmental assessment that many practitioners and agencies are already using for eligibility assessment and longitudinal progress monitoring. Unlike most measures of early childhood development, the BDI-2 contains items for measuring a child's development in all five areas that are covered by IDEA: cognitive, physical/motor, communication, social/emotional, and adaptive. The instrument gives practitioners both norm-referenced and criterion-referenced information about the nature of a child's development in each of these areas. The BDI-2 normative information is based upon a

large, nationally representative standardization study that gives practitioners information about a child's level of development compared to his or her same-age peers. Additionally, the BDI-2 offers Change-Sensitive Scores (CSSs), derived through the use of the Rasch measurement model, which provide practitioners with meaningful information about a child's level of development that they can use for planning interventions, progress monitoring, and reporting. CSSs also offer practitioners information about how easy or difficult a child will find tasks that same-age peers find easy. In this way, scores from the BDI-2 are linked to the skills assessed by the test items, giving practitioners information that is valuable for planning interventions. In the following sections I provide an overview of the Rasch measurement model and an introduction to the BDI-2 CSS metric.

1. Objective Measurement with the Rasch Model

The Rasch model, introduced by Danish mathematician Georg Rasch (1960), allows test users to interpret the meaning of a test score in the context of the items on the test. Using the Rasch model, test developers place measures of item difficulty and examinee ability onto one common scale. The Rasch scale is an equal-interval metric, so a difference of one unit represents the same amount of growth or change across the entire range of the ability scale—a convenient feature that is not the case with a raw score or with some norm-referenced score scales such as a percentile rank scale. In those cases, a difference of one unit (or one point) can have substantially different meanings at the middle and extremes of the score distribution.

When measures of test item difficulty and examinee ability are placed onto a Rasch scale, test users can use the information to predict the examinee's chance of success on a particular item. Additionally, the properties of the Rasch model allow users to estimate item difficulties independent of the ability of the sample of examinees who took the test. They also permit

estimates of examinee ability regardless of which particular set of (targeted) items were included on the test. Test developers can use the Rasch model to assess the difficulty and quality of test items, to create multiple equivalent forms, to construct large calibrated item pools, and to establish score scales linked to the meaning of a test's content.

The first published test in the United States to utilize Rasch measurement principles for item and examinee calibration was the KeyMath Diagnostic Arithmetic Test (Connolly, Nachtman, & Pritchett, 1971; Woodcock, 1999). Since then, several authors of individually administered clinical assessments have used Rasch measurement principles for such processes as item development, form construction, equating, and scale development (Elliot, 1990, 2007; Elliot, Murray, & Pearson, 1979; Roid & Miller, 1997; Woodcock & Johnson, 1977, 1989; Woodcock, McGrew, & Mather, 2001, 2007). Within the area of early childhood testing specifically, several other test authors have utilized the Rasch model for test item development and form construction (Kaufman & Kaufman, 1983; Kaufman & Kaufman, 2004; Mardell & Goldenberg, 2011; Newborg, 2005b).

2. Using the Rasch Model for Test Score Interpretation

Although test developers now use the Rasch model routinely for test item development, very few have used it to create Rasch-based interpretations of clinical test scores (Woodcock, 1999). The BDI-2 (Newborg, 2005a) is the only published early childhood test that uses a Rasch-based metric for score interpretation. The BDI-2 author and development team employed the Rasch model throughout the test development process, especially for assessing the quality of items in the early stages of pilot testing, ensuring that the test covered the entire range of ability for the test, and calibrating the items for proper placement (in ascending order of difficulty) in the test books. More recently, the BDI-2 publisher developed the Rasch-based

Change-Sensitive Score[1] scale, aptly named to highlight its sensitivity to actual changes in an examinee's ability.

The CSS is a linear transformation of the Rasch scale, chosen so that distances along the scale have meaningful, easy-to-remember probability implications (Woodcock, 1999). The linear transformation sets the center point of the scale at 500 and eliminates negative numbers. A scaling factor of 9.1024 makes the probability implications easy to remember at several points along the scale. As with other Rasch-based scales, the distance between an examinee's ability measure and the difficulty measure of the test item, in CSS units, determines his or her chance of success on the item. For example, we expect that an examinee will have about a 50% likelihood of success on an item with difficulty exactly equal to his or her CSS ability measure, 90% likelihood of success on an item 20 points less difficult than the CSS ability measure, and only 25% success on an item 10 points more difficult than the CSS ability measure. An examinee's CSS Difference Score is the difference between the examinee's CSS and the median CSS for the examinee's same-age peers. Because the CSS Difference Scores are interval-level scores, the test authors can use these scores to report the Relative Developmental Index (RDI),[2] an important interpretive feature of the BDI-2 which represents "an individual's predicted level of success on those tasks performed with 90% success by average individuals at a given age or grade level" (Woodcock, 1999). An RDI of 40/90, for example, indicates that the examinee can perform with only 40% success the tasks that his or her age peers can perform with 90% success. In contrast, an RDI of 95/90 indicates that the

[1] In its derivation and application, the CSS is nearly identical to the *W*-score metric first introduced by Woodcock and Dahl (1971) and used extensively in the Woodcock-Johnson tests for more than 35 years. Many of the interpretive features of the BDI-2, including the CSS, "difference scores," Relative Developmental Index (RDI), and RDI-associated levels of development, borrow heavily from Woodcock (1978), McGrew, Werder, and Woodcock (1991), and McGrew, Schrank, and Woodcock (2007).

[2] The RDI is based upon the principles of the Relative Proficiency Index (RPI) currently used in the Woodcock-Johnson III, which is an extension of the Relative Mastery Index (RMI) first introduced by Woodcock (1973) and used in the Woodcock-Johnson Psycho-educational Battery (Woodcock, 1978) and the Woodcock-Johnson Psycho-educational Battery–Revised (Woodcock & Johnson, 1989).

examinee can perform with 95% success those tasks that his or her age peers can perform with 90% success. RDIs are valuable for interpreting an examinee's ability level because they describe relative quality of performance rather than simply one's rank in a group. Also, RDIs maintain their meaning across time, even if the ability of the population changes over time (Woodcock, 1999). Table I contains the RDI ranges and corresponding interpretations for examinee level of development for various CSS Difference Scores on the BDI-2.

The interpretive power of the RDI lies in its ability to describe an examinee's likelihood of success based on the skills measured by the test. Using the RDI, practitioners can make assertions about a child's level of development that do not rely on the particular shape of the distribution of scores in the norming sample. In many cases the RDI can detect changes or differences in examinee level of development that might be masked by the traditional norm-referenced metrics used in eligibility decisions, such as standard scores or differences in standard deviation.

To illustrate the practical use of the BDI-2 CSS, consider a 24-month-old female who was referred for evaluation for expressive communication delay. The early childhood practitioner administered the BDI-2 Expressive Communication scale. The child's raw score of 20 earned a scaled score of 3 and a percentile rank of 1. Her CSS was 417, which is 46 points less than the median CSS for 24-month-old children from the BDI-2 norming sample. Her RDI indicated that she would only have about 5% success on test items on which her same-age peers would have 90% success. Based on this information, the practitioner determined that the child likely has a moderate expressive language delay. The child was referred for intervention services with scheduled follow-up testing. Table II contains the BDI-2 scores from the child's initial evaluation at 24 months as well as the scores from the two scheduled follow-up testing sessions at 36 and 48 months.

TABLE I

BDI-2 CHANGE-SENSITIVE SCORE DIFFERENCES AND CORRESPONDING INTERPRETATIONS FOR EXAMINEE LEVEL OF DEVELOPMENT

CSS Difference Score	Relative Developmental Index (RDI) Range	Examinee's Level of Development	Examinee Will Find Test Items That Average Same-Age Peers Perform with 90% Success:
+31 and above	100/90	Very Advanced	Extremely Easy
+14 to +30	98/90 to 100/90	Advanced	Very Easy
+7 to +13	95/90 to 98/90	Age-Appropriate to Advanced	Easy
−6 to +6	82/90 to 95/90	Age-Appropriate	Manageable
−13 to −7	67/90 to 82/90	Mildly Delayed to Age-Appropriate	Difficult
−30 to −14	24/90 to 67/90	Mildly Delayed	Very Difficult
−50 to −31	3/90 to 24/90	Moderately Delayed	Extremely Difficult
−51 and below	0/90 to 3/90	Severely Delayed	Virtually Impossible

Note: From *Manual and Checklist. Report Writer for the WJ III* (p. 10), by F.A. Schrank and R.W. Woodcock, 2002, Rolling Meadows, IL: Riverside Publishing. Copyright 2002 by Riverside Publishing.

TABLE II

BDI-2 EXPRESSIVE COMMUNICATION SCORES FOR A CHILD REFERRED FOR EVALUATION

Score Metric	24 Months	36 Months	48 Months
Raw Score	20	31	41
Scaled Score (Mean = 10, SD = 3)	3	3	3
Percentile Rank	1	1	1
Change-Sensitive Score (CSS)	417	461	491
Relative Developmental Index (RDI)	5/90	11/90	27/90
Level of Development	Moderately Delayed	Moderately Delayed	Mildly Delayed

Note: Adapted from "Evaluating Growth Trajectories and Measurement Invariance for Early Developmental Domains," by M. Ledbetter, J. Betts, T. Boney, and M. Custer, presented at the annual conference of the National Association of School Psychologists, February 2013.

In Table II, we see that the child's raw score for the BDI-2 Expressive Communication Subdomain increases at each point in time, from a baseline of 20 points at 24 months to 31 points at 36 months and 41 points at 48 months. At all three points in time, the raw scores yield scaled scores of 3 and percentile ranks of 1. A practitioner relying on the normative information alone might (incorrectly) conclude that the intervention had had no positive effect on the child's level of expressive communication development. However, we can see in Table II that the BDI-2 CSSs at the three time points are 417, 461, and 491, indicating that the child has made actual gains in skill development. To determine the extent to which those gains affect her level of functioning, the practitioner utilizes the RDI and corresponding developmental levels. This information shows that the child has improved her functioning and is now 27% likely to have success on test items that her same-age peers can perform with 90% success; her developmental level is now within the mildly delayed range.

This example shows the power of the CSS metric for detecting growth and change in developmental skills, but it also shows how the CSS can help practitioners to describe a child's current level of functioning using criterion-referenced labels, or RDIs. This type of Rasch-based metric is ideally suited for use in identifying delay, tracking progress, and the aggregate reporting required of state agencies serving young children with developmental delays.

C. Purpose of This Study

The BDI-2 is a norm-referenced early childhood developmental assessment that can provide practitioners with valuable information for determining whether a child is manifesting delay in any of the five areas covered by IDEA, for planning interventions, and for reporting on the child's progress over time. The Examiner's Manual that accompanies the BDI-2 documents extensive validity evidence that the test publisher collected during the course of test standardization. Since the time when the BDI-2 was first published, the test developer has introduced the Change-Sensitive Score (CSS) metric into the scoring and reporting program for the test. The Rasch-based CSSs give test users information about a child's development beyond what the norm-referenced BDI-2 scores provide. In this chapter I have explained how the addition of the CSS scoring metric makes the BDI-2 a useful tool for addressing the measurement and reporting challenges facing early childhood practitioners. First, it allows practitioners to gauge a child's level of development based on the likelihood that the child will be able to perform theoretically identified milestone tasks. Unlike most norm-referenced methods of identification, the CSSs are not influenced by the shape of the score distribution from a particular normative sample; rather, they maintain their meaning across the entire range of ability. Additionally, knowing how difficult certain tasks are for children with disabilities can aid practitioners in targeting intervention plans to those skills that are just emerging. Finally, BDI-2

CSSs can reveal actual developmental growth that is masked by traditional standard scores and percentile ranks, providing important information for programs that must report children's progress to OSEP.

In this study I set forth propositions about the BDI-2 scores that, if supported, will provide further support for the scores' applicability to diagnosis, planning, and reporting. To investigate these propositions, I outline research questions, each one related to a specific use or interpretation of the BDI-2 scores. For each question, I describe some types of evidence that are not necessarily identified in the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999), but that would be beneficial in validating the BDI-2 CSSs for use in diagnosis, planning, and reporting. I establish Rasch-based methods for investigating each research question, following a framework that Wolfe and Smith (2007b) proposed. The results of my investigation provide validity evidence to support or refute the use or interpretation of the BDI-2 scores for identification, progress monitoring, and reporting.

In the literature review that follows, I document the history of early childhood assessment and describe the legislative decisions that have impacted the practice of testing young children. I review the most widely used developmental assessments and provide information about the features of each one. I then provide a detailed overview of the BDI-2, including its uses, administration, and scoring; the development of the first edition and the research related to its uses; and the development of the second edition. I discuss the current conceptions of validity and review the existing validity evidence for the BDI-2. Next, I introduce the propositions and research questions that focus this study. I discuss how my examination of evidence gathered to

answer these research questions might contribute to the existing body of validity evidence supporting the use of scores on the BDI-2 for determining developmental delay, planning intervention, and reporting progress. Finally, I investigate whether the results of my study suggest changes to future editions of the BDI-2.

II. REVIEW OF THE LITERATURE

A. Early Childhood Developmental Testing: An Historical Perspective

Early childhood developmental testing in this country evolved within the context of social, scientific, and political movements. First, the increases in school enrollment in the early part of the 20th century necessitated a method for classifying children according to their mental abilities. Second, a new field emerged within psychology that was devoted to the study of how children develop and learn. Finally, federal legislation passed in the latter part of the century prompted the mandatory testing of young children and required that assessment instruments measure development across *all* the domains—not just the cognitive domain.

1. The Emergence of Individual Testing

Early childhood developmental testing branched off from the individual intelligence testing that began in the early to mid-1900s. Historically, individual educational assessment was nearly synonymous with intelligence testing, or the assessment of an individual's intellectual capabilities and ability to learn. Intelligence testing in this country dates back to the early 1900s where it began in response to a need for a method of classifying individuals according to their mental ability. In the late 1800s and early 1900s, most U.S. states instituted compulsory school attendance laws for children. These laws, coupled with the high rate of immigration and the declining infant mortality rates, caused a sharp increase in school enrollments in the early 1900s. These enrollment increases made it necessary for educators to identify some measure of mental ability that would help them place children in appropriate school classes. Concurrently, this same social dilemma was playing out in France. In response, the French government commissioned a psychologist, Alfred Binet, and his colleagues to construct a test that would classify students according to their mental ability and identify children

in need of special education (Bracken, 1991). Binet's work in the early 1900s focused on identifying qualitative differences between the mental functioning of individuals. Along with a colleague, Theodore Simon, Binet developed the first test of general mental ability for French children in 1905. The test contained many features still common in contemporary intelligence tests: the items were arranged in ascending order of difficulty, were scored as either pass or fail, and were administered using standardized procedures. With the publication of this scale, Binet and Simon also introduced the concept of "mental age," a concept that is widely used in testing today. According to Binet and Simon's definition of mental age, a 7-year-old child who passes all the items that a typical 8-year-old child would pass is said to have a mental age of 8. This type of interpretation requires that the test developers establish what is "typical" for each age. Binet and Simon spent much time gathering data on the performance of French schoolchildren on their test items so that they could describe exactly what "typical" performance on the items was. This "normative" information is a quality that they felt distinguished their test from others at the time. In the words of Binet (1905/1916):

> The scale that we shall describe is not a theoretical work; it is the result of long investigations, first at the Salpêtrière, and afterwards in the primary schools of Paris, with both normal and subnormal children. These short psychological questions have been given the name of tests. The use of tests is today very common, and there are even contemporary authors who have made a specialty of organizing new tests according to theoretical views, but who have made no effort to patiently try them out in the schools. … We place but slight confidence in the tests invented by these authors and we have borrowed nothing from them. All the tests which we propose have been repeatedly tried, and have been retained from among many, which after trial have been discarded. We can certify that those which are here presented have proved themselves valuable. (p. 2)

A few years later, the American psychologist Lewis Terman translated the Binet test into English. In 1916, Terman, working with Theodore Simon, published the third edition of the scale, naming it the Stanford-Binet Scale. In this edition, Terman advanced the concept of mental

age by creating a new score, called the intelligence quotient, or IQ. Terman computed IQ using the following formula:

$$\frac{\text{Mental Age}}{\text{Chronological Age}} \times 100. \quad (2.1)$$

Terman felt that including the child's actual chronological age in the equation would help to put the mental age in a more appropriate context. (The multiplier simply removes the decimal from the result.) This method of computing IQ is methodologically problematic and is not the same one used today; however, the development of the IQ scale was an important step in making the results of the early intelligence tests accessible to nonpsychologists.

Intelligence testing continued to gain momentum in the United States during the first half of the 20th century. During World War I, the focus of IQ testing was extended beyond the school age when the U.S. Army began giving a group-administered measure of intelligence to army recruits to more appropriately assign them to military jobs. But it wasn't until the post-World War I years that psychologists began to be interested in testing preschool-age children (Bracken, 1991). The first tests for preschool children were simply downward extensions of existing intelligence tests, including the third edition of the Stanford-Binet, which extended the age range downward to 2 years. During the 1940s, several tests were published especially for measuring the intelligence of infants and preschool children. These included the Cattell Infant Intelligence Scale, the Northwest Infant Intelligence Scale, and the Leiter International Performance Scale. However, by the late 1940s and early 1950s, researchers began to realize that personal and social variables were also important components contributing to the overall functioning of children, and subsequent test development for early childhood focused on these factors of general functioning, as well as on intelligence (Bracken, 1991).

2. Child Development as a Discipline

In the early 1900s, as Binet's work on IQ testing was getting underway, an American researcher named Arnold Gesell was beginning to formally study the way children develop. Gesell was perhaps the first researcher to systematically document child development across the broad spectrum, including language, motor, and emotional development and personal hygiene skills in addition to intellectual development. Gesell spent much of his career studying and documenting the behaviors of the children who were brought to his Clinic of Child Development at Yale University during the early to mid-1900s. Gesell and his staff developed innovative methods for observing infants and young children at play, including a "device for segregative viewing"—a laboratory with a one-way mirror and a photographic observatory dome (Gesell, 1928). The result of Gesell's observational work was a description of the developmental steps or *gradients of growth* for the typical child at each age (Thomas, 1979). Gesell (1928) wrote, "It is assumed that fundamentally the laws of growth are universal" (p. 5). He believed that the order of development is predictable, governed by genetics, and can be assessed through the use of simple tests. Therefore, one goal of his work was to document as accurately as possible the typical behaviors in early childhood so that parents, teachers, and doctors could have a measure against which to gauge children's development.

In 1928, Gesell published his first version of a developmental assessment battery, titled Infancy and Human Growth. The instrument has been revised and updated a number of times, and the current version, the Gesell Developmental Observation, is still widely used today to screen young children for developmental delay and to assess school readiness.

B. Legislative Influences on Early Childhood Testing

President Gerald Ford signed Public Law 94-142, the Education for All Handicapped Children Act, into law in November 1975. The law was the first to require educational services for children with disabilities in American public schools. Among its features were the mandate that all children, regardless of disability, receive a free and appropriate public education (FAPE), and that identification and evaluation procedures be nondiscriminatory and tailored to the specific needs of the child. The first version of the law required that states provide these services to children ages 5 through 21, as long as the states provided the same access to education to other children of the same ages without disabilities.

In 1986, Public Law 94-142 was expanded through the passage of Public Law 99-457, the Education for All Handicapped Children Act Amendments, or the Preschool Law. As the nickname suggests, these amendments were added to expand services to young children ages 3 through 5 with disabilities. Public Law 99-457 established Part B, which required states to serve children ages 3 to 5, regardless of whether or not children without disabilities were receiving educational services at those ages, and Part H, which incented states to provide early intervention services to children with disabilities from birth to age 3. For Part B, states were required to comply with the provisions or risk losing federal funding. Part H was established as a voluntary program, and states were given the freedom to define eligibility requirements and to determine what type of intervention services they would provide.

Since 1986, there have been three more reauthorizations of the Act, including an amendment in 1990 that changed the name of the law to the Individuals with Disabilities Education Act, or IDEA. The passage of Public Law 94-142 and subsequent research in early childhood education and development spurred a momentum toward early intervention. In fact,

when the federal government reauthorized the law as the Individuals with Disabilities Education Improvement Act of 2004 (IDEA 2004), all 50 states were participating in the voluntary Part C (formerly Part H), which provides federal funding to states for the creation of early intervention programs for children with disabilities from birth through age 35 months.

C. The Importance of Early Educational Intervention

Educational researchers have shown that quality early intervention, which can be defined as "the process of anticipating, identifying, and responding to child and family concerns in order to minimize their potential adverse effects and maximize the healthy development of [children]" (Oser & Cohen, 2003, p. 3), has had significant positive impacts on the educational attainment of children with disabilities. Researchers have studied the impact of early intervention on many aspects of education and society. Studies have linked early intervention to higher academic achievement (Campbell & Pungello, 2000), higher IQ (Wasik, Ramey, Bryant, & Sparling, 1990), higher rates of college attendance (Campbell & Pungello, 2000), improved social and emotional functioning (Mahoney & Perales, 2003), and lower crime rates (Reynolds, Temple, Robertson, & Mann, 2002). Researchers reported positive impacts of early intervention on specific subgroups of children, including those with Down syndrome (Hernandez-Reif et al., 2006; Kumin, Von Hagel, & Bahr, 2001; Rynders & Horrobin, 1990), vision impairment (Beelman & Brambring, 1998), hearing impairment (Geers, 2002), autism (Mahoney & Perales, 2003; Sheinkopf & Siegel, 1998), and low birth weight or prematurity (Hill, Brooks-Gunn, & Waldfogel, 2003; Sparling & Lewis, 1991), as well as children with environmental risk factors (Grantham-McGregor, Powell, Walker, Chang, & Fletcher, 1994). Additionally, researchers in other disciplines have studied early intervention. For example, in economics, researchers have

attempted to study the return on investment achieved by early intervention and its impact on the local economy (Masse & Barnett, 2002).

D. Norm-Referenced Versus Criterion-Referenced Test Interpretations

Before I begin the discussion of the most commonly used early childhood developmental assessments, it is necessary to make an important distinction between norm-referenced and criterion-referenced interpretations of test scores. Most of the methods of determining developmental delay described in Chapter I rely on *norm-referenced* test score interpretations. Norm-referenced test score interpretations are those that compare the performance of one examinee to the performance of a group, termed the *normative group*. In most cases, the normative group is understood to be representative of the population as a whole, but the extent to which this is actually true depends on how well the test publisher followed census statistics or other population parameter information in the selection of the normative group. Commonly reported norm-referenced test scores include percentile ranks, standard or *z* scores, *T* scores, stanines, and age-equivalent scores. A percentile rank indicates the percentage of examinees in the normative group who scored below a particular score. A standard score, or *z* score, uses the standard deviation of the score distribution as the unit of measurement, so that the mean score is reported as 0, and a score of +1 is equivalent to a score that is one standard deviation above the mean of the scores. An advantage of the *z*-score scale is that when the test scores are normally distributed, one can directly compare an examinee's *z* score to the area under the normal curve to determine the corresponding percentile rank of the score. One disadvantage of *z* scores, however, is that in practice, when scores are normally distributed, approximately half of the scores fall below zero, which may confuse individuals who are trying to interpret the scores. The *T* score is a transformed standard score with a mean of 50 and a standard deviation of 10. *T* scores are all

positive, thereby making them easier to interpret than *z*-scores. The commonly known IQ scale is also a transformation of a standard scale. In the case of IQ scores, the scale mean is 100, and the standard deviation is 15. The stanine scale (the name is derived from "standard nines") divides the normal distribution into nine sections, each of which is one-half a standard deviation wide. Stanine scores are whole numbers, with a mean of 5 and a standard deviation of 2. Because one stanine represents a band of scores, the use of stanine scores may reduce the risk of test users over-interpreting small score differences (i.e., attributing too much meaning to small score differences between individuals, which is a risk when scores, rather than score bands, are reported) (Wiersma & Jurs, 1990). Finally, an age-equivalent score is the median raw score for a particular age. For example, if the median raw score for all examinees age 3 years, 5 months in the normative sample was 43, then an examinee who scores 43 on the test is said to have an age-equivalent score of 3 years, 5 months, regardless of that child's actual chronological age. Note that none of these interpretive statistics conveys information about an individual's performance in absolute terms; rather, they all describe the individual's performance in relation to a normative group.

Criterion-referenced test score interpretations, on the other hand, describe the performance of the examinee in terms of the skills and abilities measured on the test. Robert Glaser (1963) coined the term in his classic work on the measurement of achievement:

> Underlying the concept of achievement measurement is the notion of a continuum of knowledge acquisition ranging from no proficiency at all to perfect performance. An individual's achievement falls at some point on this continuum as indicated by the behaviors he displays during testing. … The degree to which this achievement resembles desired performance at any specified level is assessed by *criterion-referenced* [italics added] measures of achievement or proficiency. (p. 519)

Glaser went on to elaborate on the usefulness of these interpretations:

> The specific behaviors implied at each level of proficiency can be identified and used to describe the specific tasks a student must be capable of performing before he achieves one of these knowledge levels. … Along such a continuum of attainment, a student's score on a criterion-referenced measure provides explicit information as to what the individual can or cannot do. (p. 519)

Tests that provide criterion-referenced scores are typically constructed in one of two ways: (a) through the use of a well-specified content domain, or (b) through the use of instructional objectives (Wiersma & Jurs, 1990). Large-scale assessments based on learning standards have come into widespread use in recent years due to the requirements of the No Child Left Behind (NCLB) Act of 2001. The act, which is a reauthorization of the Elementary and Secondary Education Act of 1965, establishes a system of accountability whereby states are required to show that students are achieving proficiency in reading, math, and science, according to a set of well-defined learning standards. Since the passage of NCLB, nearly every state in the United States has developed its own content standards and has either developed or contracted with a test publishing company to develop custom assessments.

Another defining feature of tests that provide criterion-referenced score interpretations is the establishment of cut scores. Cut scores "separate a test score scale into two or more regions, creating categories of performance or classifications of examinees" (Cizek & Bunch, 2007, p. 13). These regions are then labeled in a manner according to the purpose for which the test was designed. Examples of performance categories, also called *performance levels*, are "Proficient/Not Proficient," "Pass/Fail," or, in an example of an assessment with three performance categories, "Does Not Meet Expectations/Meets Expectations/Exceeds Expectations." Typically, an examinee's test results are accompanied by *performance-level descriptors* (PLDs)—detailed descriptions of the measured skills and knowledge that examinees possess at each performance level. The examinee's performance level, together with the corresponding PLD, indicates exactly what that examinee knows and/or is able to do. Note how

this interpretive information differs from norm-referenced score interpretations, where an examinee's score is typically expressed as a number and is interpreted within the context of the performance of a comparative, or normative, group.

Although the terms *norm-referenced test* (NRT) and *criterion-referenced test* (CRT) are used extensively in the educational literature, Frisbie (2005) argued against this use. He suggested that "norm-referenced" and "criterion-referenced" should refer to the type of score interpretations that assessments provide and not to the assessments themselves, because scores from one instrument could provide both types of interpretations. He believed that researchers have misused these terms in the educational literature so much that many practitioners think that a test has to be *either* an NRT or a CRT, and cannot be both. In fact, Frisbie noted, the current *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) uses the terms "criterion-referenced test" and "criterion-referenced score interpretation" synonymously in the glossary, further contributing to the (false) notion that a test that provides norm-referenced scores cannot also provide criterion-referenced score interpretations (Frisbie, 2005).

E. Contemporary Early Childhood Developmental Tests

With the repeated reauthorizations of IDEA Part C and its five-domain structure for defining delay, test developers have generally followed suit and created early childhood batteries that measure performance in the five domains. Many of the assessments are administered by a trained examiner; others are checklists assessing the child's skills and abilities that the parent or caregiver completes. Most of the assessments used to determine early intervention eligibility yield norm-referenced score interpretations, while some yield criterion-referenced score interpretations. In this section, I briefly describe the most widely used, multi-domain early

childhood batteries. I provide the age range, content, and score information for each assessment, as well as information about the technical aspects of each test.

The most commonly used multi-domain assessments are the Battelle Developmental Inventory, Second Edition (to be discussed in a later section of this chapter); the Bayley Scales of Infant and Toddler Development, Third Edition; the Mullen Scales of Early Learning; and the Developmental Indicators for the Assessment of Learning, Fourth Edition.

1. Bayley Scales of Infant and Toddler Development, Third Edition

The third edition of this norm-referenced test, commonly known as the Bayley-III (Bayley, 2006), was published in 2006. It is used for children ages 1 through 42 months. It provides scores in the five domains covered by IDEA through the use of administered subtests for the Cognitive, Motor, and Language domains and the use of parent/caregiver checklists for the Social-Emotional and Adaptive Behavior domains. This is an improvement over the second edition of the test, which contained administered subtests for only the Cognitive (called "Mental" in that edition) and Motor scales. Although the Bayley-III contains assessments for all five domains, each domain can be administered as a stand-alone scale. The Bayley-III was normed on a sample of 1,700 children for the Cognitive, Motor, and Language scales, 456 children for the Social-Emotional scale, and 1,350 children for the Adaptive Behavior scale. The norm samples contained only typically developing children; however, children with developmental delay, risk factors for developmental delay, and identified health impairments were included in special validity studies. Samples were matched to the 2000 U.S. census demographic information (Bayley, 2006).

Scores provided on the Bayley-III include scaled scores, composite scores, and percentile ranks for each of the five scales. Additionally, the Cognitive, Language, and Motor scales

provide growth scores and developmental age-equivalent scores (Bayley, 2006). Validity evidence reported in the manual includes the results of studies correlating the scores on the Bayley-III with other early childhood developmental instruments; results of factor analysis studies which supported the three-factor (Cognitive, Motor, and Language scales) model underlying the test; and results of special group studies indicating lower mean scores for groups of children with developmental delay and other disabilities.

2. Mullen Scales of Early Learning: AGS Edition

The Mullen Scales of Early Learning: AGS Edition (MSEL-AGS) (Mullen, 1995) is an individually administered assessment instrument that is based on an information-processing model. In this model, performance on a task is analyzed through the component processes that the task requires, such as motor skills, visual reception, and language. As such, the test provides scores for Gross Motor, Visual Reception, Fine Motor, Receptive Language, and Expressive Language; taken together these areas cover three of the five domains under IDEA Part C—motor, cognitive, and communication. The MSEL-AGS was normed on a group of 1,849 children ages birth through 69 months (5 years, 9 months). The norm group was stratified by sex, race/ethnicity, father's occupation, and community type (rural/urban). The norm group contained only typically developing children; however, the authors reported the results of special studies with children who had developmental delay or other disabilities.

A trained examiner administers the items in the MSEL-AGS. Many items require assistance and/or information from a parent or caregiver. Some of the items are scored dichotomously (0,1), while others are scored using a variety of score points (0,1,2,3,4,5). Scores provided for the five scales are *T* scores and associated descriptive categories (i.e., average,

below average, etc.), percentile ranks, and age equivalents. For the composite, a standard score (Mean = 100, *SD* = 15), descriptive category, and percentile rank are provided.

In the test manual, the authors provide a variety of evidence to support the validity of the scores from the MSEL-AGS. The mean scores progress upward with age, thus supporting the developmental nature of the assessment. Intercorrelations of the scales and total scores as well as factor analysis results provide evidence for validity based on internal structure. Moderate correlations with the corresponding Bayley-III scale scores (Cognitive and Motor) provide evidence of validity based on relationships to other variables.

3. Developmental Indicators for the Assessment of Learning, Fourth Edition

Published in 2011, the Developmental Indicators for the Assessment of Learning, Fourth Edition (DIAL-4) (Mardell & Goldenberg, 2011) is a developmental screening instrument. It can be used to assess motor, concepts, language, self-help, and social development of children ages 2 years, 6 months through 5 years, 11 months. It is not a full-scale diagnostic battery, but rather a screening tool to be used to determine if further developmental testing is warranted. It is designed so that multiple groups of examiners can administer it to large groups of children in a setting that allows movement from one testing station to the next. The motor, concepts, and language areas are assessed using performance items; the self-help and social development areas are assessed using a parent-completed rating-scale instrument. The DIAL-4 test developers gathered a norming sample of 1,400 children stratified by age, sex, geographic region, race/ethnicity, and parent education level. Roughly 2% to 3% of the children in the norming sample were diagnosed with a speech and language impairment, 2% to 3% were diagnosed with developmental delay, and fewer than 1% were diagnosed with some other impairment (e.g., autism, orthopedic impairment) (Mardell & Goldenberg, 2011). The DIAL-4

provides cutoff scores for each assessment area. Examiners may choose the percent of the population that they would like to refer for further testing—16%, 10%, 7%, 5%, or 2%, corresponding to z scores of approximately –1, –1.3, –1.5, –1.7, and –2.0—and then they can use either the raw score or a scaled score to assess whether a particular child's performance falls below the cut for that rate of referral. Norms are provided for 2-month age intervals.

The DIAL-4 manual provides a summary of the validity evidence gathered during the standardization study. Intercorrelations between the DIAL-4 areas and DIAL-4 total score are reported and range from .19 between the social and motor areas and between the language and self-help areas to .88 between the concepts area and the DIAL-4 total score. Correlations between the DIAL-4 scores and other early childhood screening and full-scale assessment instruments are also reported. Finally, the authors establish that the DIAL-4 score patterns for children with specific disabilities are as expected, thereby supporting the use of the screener for children with disabilities (Mardell & Goldenberg, 2011).

4. Battelle Developmental Inventory, Second Edition

The Battelle Developmental Inventory, Second Edition (BDI-2) (Newborg, 2005a) is an individually administered, norm-referenced assessment measuring developmental milestones in children from birth through age 7 years, 11 months. The BDI-2 is based on the concept that children develop according to a standard set of milestones, and that these developmental milestones emerge in roughly the same sequence, and at roughly the same ages, for typically developing children. Also, the acquisition of each skill is usually dependent upon having mastered the preceding skills along the developmental continuum.

a. Uses of the BDI-2

The primary uses of the BDI-2 are identification of developmental delay, assessment and monitoring of the typically developing child, planning for instruction and intervention, and evaluation of programs serving young children (Newborg, 2005b).

i. Identification of Developmental Delay

Under Part C of the current IDEA legislation, each state or jurisdiction must provide early intervention services to "any child under 3 years of age who needs early intervention services because the child is experiencing developmental delays, as measured by the appropriate diagnostic procedures in 1 or more of the areas of cognitive development, physical development, communication development, social or emotional development, and adaptive development" (IDEA, 2004). Because the BDI-2 aligns well with these five categories of development, it is often used as a tool in making determinations of delay. As discussed in Chapter I, there is no universal definition of "developmental delay"; therefore, each state must set its own criteria for defining delay. Many states use performance on a standardized test that is below a certain level as a qualifying factor. For example, some states require that a child be functioning at a 20% delay in one or more of the five developmental areas to qualify. In this case, a practitioner would use the age-equivalent score metric to make this determination. Other states require that a child receive a score that is at least 1.5 standard deviations below the mean in any area to qualify. In this case, a practitioner could use the *z*-score metric. Because the BDI-2 results can be reported in a variety of score metrics, including age-equivalent scores, percentile ranks, *z* scores, and Developmental Quotient (DQ) scores, it is a widely accepted tool for making eligibility determinations.

ii. Assessment and Monitoring of the Typically Developing Child

Because the BDI-2 spans the age range from birth through 7 years, 11 months, it can provide users with a longitudinal record of development for the critical early childhood period (Newborg, 2005b). Comparing norm-referenced scores from multiple BDI-2 administrations can help practitioners assess how a child's skill level has changed over time, relative to his or her same-age peers. Additionally, practitioners and parents can use information about a child's performance on specific items to see in which domains a child has progressed and which milestones the child has mastered since the previous assessment. Because the BDI-2 items are ordered from least to most difficult, and because the examiner only administers those items that are developmentally appropriate for the child, there may be very little overlap of the items administered from one testing session to another. This is especially true if the time between testing sessions is long enough for the child to make significant developmental progress.

iii. Planning for Instruction and Intervention

Under IDEA, children with disabilities are required to have individualized education programs (IEPs) or, for children under three years of age, individualized family service plans (IFSPs). The IEP/IFSP is a document a team of professionals in collaboration with a child's parents prepares. It contains a description of the child's present functioning, outlines measurable goals and how these will be assessed, and contains a description of the services that teachers and other professionals will provide to the child. Professionals can use the BDI-2 to assist them in writing IEP/IFSP goals. Each BDI-2 item contains a behavioral objective (the "milestone") and a procedure for assessing the behavior. Professionals can rewrite these behavioral objectives as IEP/IFSP measurable goals, and then use the scoring criteria to

operationalize the skills or behaviors that the child will exhibit when the goals have been attained (Newborg, 2005b).

iv. Evaluation of Programs Serving Young Children

Researchers used the original version of the BDI to evaluate the progress of groups of children that educational programs serve (e.g., see Hundert, Mahoney, Mundy, & Vernon, 1998; Markowitz & Larson, 1988; Zeece & Wang, 1998). Due to its breadth of coverage and wide age range, researchers often used the BDI as an outcome measure in research on early intervention strategies (e.g., see Cohen & Mannarino, 1996; Fallon, 1994; Kouri, 2005; Sayers, Cowden, Newton, & Warren, 1996; Tingey, 1991; Whitmore, Ford, & Sack, 2003). Finally, because the early childhood field widely accepted the original BDI as a measure of development in early childhood, it was often used as a target measure in the validation of other, similar instruments (e.g., see Farmer-Dougan & Kaszuba, 1999; McLean, McCormick, & Baird, 1991; Saylor, Boyce, Peagler, & Callahan, 2000).

b. History of the BDI

In 1973, the U.S. Office of Education contracted with the Battelle Memorial Institute to create an assessment to help evaluate the effectiveness of a network of early childhood education programs the federal government funded. These early childhood programs served children from birth through age 8 with speech impairment, mental retardation, learning disabilities, deafness, visual impairment, emotional disturbance, and other disabilities. The results of the new assessment would allow the government to track the developmental progress of the groups of children that these programs were serving, thus creating a standard measurement tool to judge the relative effectiveness of the programs (Newborg, 2005b).

The development team at the Battelle Memorial Institute studied other published early childhood assessments to compile lists of the most commonly assessed developmental skills. They then grouped the skills together into five domains of development—Motor, Cognitive, Communication, Adaptive, and Personal-Social Skills—and sequenced them in order of typical child development. Experts in each developmental area then reviewed the lists of skills to determine which skills were critical milestones. The development team defined milestones as behaviors or skills that (a) are important for the development of normal functioning, (b) occur frequently in the child-development literature, (c) practitioners agree are milestones, and (d) are receptive to educational intervention (Newborg, 2005b).

After identifying the milestone skills and behaviors, the development team divided the skills in each domain into smaller, more specific units called subdomains. Next, they wrote items to assess each skill. Each item contained a statement of the target skill or behavior and a standard procedure for assessing the skill or behavior. Additionally, the development team assigned each item to an age range from birth through age 8. The item set was then pilot tested, and the development team used the results of this pilot test to revise items, adjust item order, and finalize the scale. In 1984, DLM/Teacher Resources published BDI. In 1992, Riverside Publishing purchased the publishing rights to the BDI.

Although the original purpose of the BDI was to evaluate programs serving children with disabilities, early childhood education professionals quickly became interested in the instrument as a measure for identifying developmental delay in individual children. There were, however, some criticisms of the instrument and cautions for its use that appeared in the research literature in the late 1980s and 1990s. This type of literature falls into two main categories: (a) research on the use of the BDI with specific subpopulations, and (b) research on the technical adequacy of

the instrument. In the first category, most of the available research literature from this time period focused on use of the BDI with populations of children with disabilities. In many cases, this research found that practitioners could not interpret the BDI scores in the same way for children with and without disabilities. For example, Snyder and Lawson (1993) found that the five-factor structure as reported in the BDI manual did not hold when examiners administered the test to a group of children with severe disabilities such as Down syndrome, cerebral palsy, and spina bifida. These researchers asserted that for this population, the five BDI domains should not be used as stand-alone assessments, as the test authors suggest in the manual. In another example, researchers studying children with severe motor delays found that the children in their study scored lower on the BDI Adaptive Domain than a sample of children with less severe motor delays from another study (Johnson, Cook, & Kullman, 1992). They concluded that the high reliance on motor skills for the Adaptive items could affect scores on this domain, and that these scores should be interpreted with caution for children with motor impairments.

In the second category of literature—research focused on the technical adequacy of the BDI—many researchers have noted the lack of continuity in the norms due to the large age spans used in the development of the norms. Boyd (1989) and McLinden (1989) both noted that for children whose chronological ages are at the very low or high end of their respective norm groups, scores on the BDI could differ dramatically when the child's performance is compared to the two adjacent sets of age norms. Additionally, some reviewers commented on the small number of items at each level in some subdomains and the apparent gaps in item difficulties and steep item gradients (Bailey, Vandiviere, Dellinger, & Munn, 1987; Bracken, 1987). All of these issues can lead to large standard score differences in relation to relatively small raw-score point differences, and can skew the interpretation of the BDI results.

c. Development of the BDI-2

In 2000, Riverside Publishing began the process of revising and renorming the BDI. The development team first undertook an extensive review of the research literature to determine how practitioners were using the BDI and what its strengths and weaknesses were relative to other early childhood developmental assessments. Next, the company contacted existing customers to find out what they liked about the instrument and what suggestions they had for improving it. The BDI-2 development team also paid on-site visits to practitioners who were using the original BDI for large-scale screening events to see firsthand what the strengths and weaknesses of the instrument were for that purpose.

From these multiple sources, the development team identified the following goals for the revision: (a) update the test items and manipulative "toys" to address current technology and societal norms, (b) add a Spanish adaptation and translation of items and materials, (c) update test materials to include colorful artwork and a more organized presentation of items, (d) develop more detailed instructions for administering items, (e) simplify the score category descriptions to make scoring items easier, (f) restructure subdomains within the five domains, (g) add more difficult items to eliminate ceiling effects for high-functioning 6- and 7-year-old examinees, and (h) develop software for scoring the assessment and producing individual and aggregate reports.

For new items, the item writing and pilot testing followed a process similar to that conducted in the development of the original BDI items. The development team, together with the test author, identified gaps in the item continuum for each subdomain. These gaps were places along the continuum where there was a large increase in item difficulty from one item to the next. To fill these gaps, the author wrote new items, and the development team pilot-tested

them on small samples of children to ensure that the instructions were clear and that the score category descriptions covered all possible responses.

The development team also reviewed the existing items and made edits to clarify instructions and score category descriptions. For items containing a parent interview option for administration, the team rewrote the instructions into script format to ensure a more standardized administration. All existing artwork was re-rendered in full color.

The final tryout edition of the BDI-2 contained 466 items across the five domains. The tryout study took place in the fall and winter of 2001 to 2002. In this study, paid examiners administered the BDI-2 to approximately 850 children from birth through age 7 years, 11 months. The sample contained approximately equal numbers of males and females and contained children from nine different states in the United States. Participants from all five major racial/ethnic groups (White, Black, American Indian and Alaskan Native, Asian, and Hispanic) were included in the study.

Psychometricians analyzed the data from the tryout study using conventional Classical Test Theory (CTT) methods and Rasch item analyses. Conventional analyses included an examination of the item gradients and item difficulties across age groups, reliability analyses, and differential item functioning. The psychometricians used Rasch analysis to calibrate the items. Because the BDI-2 items need to be ordered in the test books from easiest to most difficult, psychometricians also examined the Rasch difficulty calibrations to ensure that the item order was correct.

Approximately 45 of the examiners who participated in the tryout study completed a questionnaire about the test. This questionnaire contained items pertaining to the instructions, score category descriptions, and materials in the test. Examiners were also asked to comment on items

that they felt were too easy or difficult, or items that they felt contained gender or racial/ethnic bias. This information, together with the information obtained from the item analysis, enabled the development team to make final edits to the BDI-2 prior to the standardization study. The types of changes that the author and development team made to the test at this point were minor (e.g., changes in the manipulatives kit to make certain toys durable or child-friendly, minor edits to the scripts of some interview questions to eliminate regional slang, and minor edits to a few score category descriptions to provide more detailed scoring criteria).

The BDI-2 standardization study began in the fall of 2002 and continued through the summer of 2003. During this time, the development team identified a norming sample of 2,500 children from birth through age 7 years, 11 months old from whom BDI-2 data was gathered. The sample was representative of the U.S. population on the stratification variables of sex, race/ethnicity, geographic region, and socioeconomic level (Newborg, 2005b). The sample was divided into 20 age groups. The first 8 age groups spanned 3 months each (0 to 2 months, 3 to 5 months, and so on, up to 23 months). The remaining age groups spanned 6 months each (24 to 29 months, 30 to 35 months, and so on, up to 95 months). Each age group contained 125 children. The development team chose the age groups to account for the rapid changes in development from birth through age 2 years and to allow for a large representative sample of children from the infant and toddler ages (Newborg, 2005b).

As part of the criterion-related validity studies for the standardization, examiners administered other early childhood assessments along with the BDI-2 to some of the children. These additional assessments included the Bayley Scales of Infant Development, Second Edition, the Denver Developmental Screening Test-II, the Preschool Language Scale, Fourth Edition, the Vineland Social-Emotional Childhood Scales, the Comprehensive Test of

Phonological Processing, the Wechsler Preschool and Primary Scale of Intelligence–Third Edition, and the Woodcock-Johnson III Tests of Achievement.

In addition to the 2,500 standardization cases, examiners tested 300 children representing special groups during the BDI-2 standardization study. These special groups included children with autism, cognitive delay, developmental delay, motor delay, premature birth, and speech/language delay.

Psychometricians analyzed the standardization dataset using both CTT and Rasch methods. They used the information they obtained to check the quality of the items and scales. Criteria for retaining items included high item discrimination, appropriate difficulty for targeted age groups, differentiation of responses from low- to high-ability examinees, and proper progression of scores across age ranges (Newborg, 2005b). At this point, the test development team dropped from the final item set items that did not meet these stringent requirements.

From this final item set, the psychometricians created raw score-to-scale score and scale score-to-percentile rank conversion tables for each age group and each subdomain. Then, they created composite domain scores by summing the appropriate subdomain scale scores. They used these composites to create the developmental quotient (DQ) conversion tables found in the BDI-2 Examiner's Manual. The BDI-2 was published in the fall of 2004.

d. Structure of the BDI-2

The BDI-2 is organized into five separate scales, or "domains": Cognitive, Personal-Social, Communication, Adaptive, and Motor. These five domains correspond to the five areas in which a child may be identified as having a developmental delay under IDEA. Within each domain, items are further divided into smaller scales called "subdomains." Each of these subdomains is a hypothesized unidimensional construct measuring a specific set of

developmental skills. Table III outlines the subdomains, types of skills assessed, and number of items in each BDI-2 domain and subdomain.

TABLE III
SUBDOMAINS, SKILLS ASSESSED, AND NUMBER OF ITEMS IN EACH BDI-2 DOMAIN AND SUBDOMAIN

Domain	Subdomain	Skills Assessed	Number of Items
Adaptive	Self-Care	Eating, toileting, dressing, grooming, preparation for sleep	35
	Personal Responsibility	Initiating play and other activities, carrying out tasks, avoiding common dangers	25
Communication	Receptive Communication	Responds to different tones of voice, responds to *who* or *what* questions, identifies initial sounds in words	40
	Expressive Communication	Produces vowel sounds, articulates clearly, speaks in sentences	45
Personal-Social	Adult Interaction	Responds physically when held, is aware of other people, helps adults with tasks	30
	Peer Interaction	Shares toys, plays cooperatively with other children, recognizes similarities and differences among all children	25
	Self-Concept and Social Role	Expresses emotions, is aware of differences between males and females, copes effectively with aggression, criticism, or teasing	45
Motor	Gross Motor	Walks without support, walks up and down stairs, throws ball and hits target, hops on one foot	45
	Fine Motor	Picks up objects, traces designs, ties a simple knot, cuts paper with scissors	30
	Perceptual Motor	Puts objects in a bottle, stacks cubes, copies letters, numbers, and words, writes in script	25
Cognitive	Attention and Memory	Follows auditory and visual stimuli, recites poems or songs, locates hidden objects in a picture	30
	Reasoning and Academic Skills	Names and matches colors, demonstrates basic math skills, uses simple logic to answer questions	35
	Perception and Concepts	Compares objects, sequences events in time, puts together a puzzle, groups and sorts objects	40
Total Number of Items			450

e. Administration of the BDI-2

i. Test Materials

The BDI-2 comprises five item books (one for each domain), a record form used for recording the scores, a stimulus book, presentation cards, a student workbook (necessary for the administration of some of the Cognitive and Motor items), and an examiner's manual containing instructions for administration and technical information about the test. Additionally, a set of manipulatives (toys) is required for the administration of many items. Examples of manipulatives included in the test kit are a doll, stacking cups, blocks, and a ball.

ii. Order of Administration

Examiners can administer the BDI-2 domains in any order. Examiners can use their professional judgment when determining the order of domain administration. For example, an examiner might choose to administer the items from the Cognitive Domain earlier in the testing session when a child is most alert and engaged. Likewise, an examiner may decide to administer the items from the Motor Domain midway through the testing session to allow the child an opportunity to move around and stretch.

iii. Basal and Ceiling Rules

The BDI-2, like many other clinically administered assessments, uses basal and ceiling rules to minimize testing time. The age range for the items in most BDI-2 subdomains spans from birth to age 8 years. However, the examiner needs to administer only those items that are developmentally appropriate for the child being tested. Because the items are arranged within each subdomain in ascending order of difficulty, the use of basal and ceiling rules makes this possible. A *basal* is defined as the point below which a child would most likely

pass, or receive a score of 2 for every item. A *ceiling* is defined as the point above which a child would most likely fail, or receive a score of 0 for every item.

For each subdomain of items administered, testing begins at a start point based on the child's chronological age. From there, the examiner must assign the child a score of 2 on three consecutive items to establish the basal. If the examiner does not assign the child a score of 2 on three consecutive items, the examiner must administer items in reverse order until the child passes three items. At this point, the examiner has established a basal. The examiner then resumes testing forward until the child scores 0 on three consecutive items. At this point, the examiner has established a ceiling and can discontinue testing in that subdomain.

iv. Item Administration Procedures

In consideration of the flexibility required when testing very young children, the BDI-2 allows for three different types of administration procedures:

- Structured: The examiner follows a set of instructions printed in the test manual and uses test materials and stimuli to elicit a response from the child. The examiner then judges the response based on a scoring rubric and assigns a score of 0, 1, or 2 for the item.
- Observation: The examiner observes the child in a home, school, or daycare setting over a period of time and scores the item 0, 1, or 2 based on how often the target behavior occurs.
- Interview: The examiner interviews the parent or caregiver about the child's typical behavior using a script provided in the test manual. Questions are designed to elicit information about the *frequency* and *quality* of the target behavior. The

examiner scores the item 0, 1, or 2 based on the parent or caregiver's responses to the interview questions.

All items are administered using at least one of these three options; some items have two or three options for administration. The examiner has the choice of which option to use, depending on factors such as the age and mood of the child. The Examiner's Manual recommends that when provided, the Structured administration option should be used if possible. (For some items—especially those given to the youngest children—a Structured administration procedure may not even be provided. In these cases, the examiner usually has a choice between an Observation and an Interview procedure.)

The inclusion of up to three different administration options for some items speaks to the flexibility of the instrument for assessing very young children and children with severe developmental delays. Because infants and very young children may be more affected by mood, hunger, and fatigue than older children during a testing situation, and because this population may not have the same achievement motivation for testing situations as typically developing, older children, the multiple administration options are provided to allow examiners a better chance of collecting accurate information about a child's developmental functioning than the Structured administration procedure alone might provide.

f. Scoring the BDI-2

i. Item Scoring

Most BDI-2 items are scored 0, 1, or 2.[3] Because each item measures a different skill or behavior, the scoring rubric for each item is different. In general, though, a score of 2 means that a child has mastered the skill, or displays the behavior *in most situations* when it is appropriate. A score of 1 indicates that the skill is emerging, or that the child displays the behavior *sometimes* when it is appropriate. A score of 0 is given when the child *rarely or never* displays the behavior, or cannot perform the skill being assessed. A clear scoring rubric is included with each item so that the examiner can quickly and accurately score the item during the testing session.

ii. Available Scores

The BDI-2 yields scores at the subdomain, domain, and total test level. For the subdomains, the available scores are age equivalents (AEs), percentile ranks (PRs), and scaled scores. The scaled scores for the subdomains are normalized standard scores ranging from 1 to 19, with a mean of 10 and a standard deviation of 3 for each subdomain. At the domain level, the available scores are developmental quotients (DQs) and PRs. The DQ is a normalized standard score with a mean of 100 and a standard deviation of 15. At the total test level, a BDI-2 Total DQ and PR are available.

[3] A note regarding the effect of examiner severity (or leniency) is in order for readers who may be accustomed to considering this factor within the measurement model. In clinically administered psychological and developmental tests such as the BDI-2, examiner severity is assumed to be equal across examiners. This tradition is most likely rooted in practical necessity. Because most data-gathering methods for individually-administered tests involve interaction between an examiner and an examinee, it is practically impossible to produce score matrices that include a crossover of examiners and examinees. Therefore, interrater reliability data from a sample of examiners and examinees is typically presented as a proxy for variability in examiners' ratings (i.e., strictness or leniency).

F. Validity

The sixth edition of *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) asserts that validity is "the most fundamental consideration in developing and evaluating tests" (p. 9). According to Messick (1989), validity is "an inductive summary of both the existing evidence for and the potential consequences of score interpretation and use" (p. 13). Messick's definition helped shape the current conception of validity and contains two elements that are interesting in light of how validity research has evolved over time. First, validity refers to a *unified body of evidence* that summarizes how useful the test scores are for the intended purpose(s). The accumulation of validity evidence occurs over time, first through documentation that the test developer provides in support of the test scores for their proposed use(s), and subsequently through the evaluations of test users. Messick (1989) acknowledged this when he stated that "over time, existing validity evidence becomes enhanced (or contravened) by new findings, and projections of potential social consequences of testing become transformed by evidence of actual consequences and by changing social conditions" (p. 13). Second, validity is not an all-or-nothing property. There are various degrees of validity. Further, one cannot claim that a test itself is valid or invalid. Rather, validity refers to the extent to which the scores and their inferences are appropriate for the *context and manner* in which the test is to be used (AERA, APA, & NCME, 1999).

1. Defining Validity Using the *Standards for Educational and Psychological Testing*

The currently accepted definition of validity as stated in the *Standards for Educational and Psychological Testing* is "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (AERA, APA, & NCME, 1999,

p. 9). Thus, validity evidence is accumulated and evaluated for the proposed uses and interpretations of a test's *scores*, not for the test itself. Although validity is now accepted to be a unitary concept, the *Standards* define a variety of sources of evidence that practitioners can use to evaluate the proposed interpretation of a test score for an intended purpose. The body of evidence that supports the use and interpretation of a test score for a specific purpose can be described as a "validity argument." Not all types of validity evidence are germane to all tests. The types of evidence needed to establish a validity argument depend upon the proposed uses and interpretations of the particular test's scores. Below I summarize the five general types of evidence as outlined in the *Standards* and discuss how each type of evidence might support the use of scores from a particular test for a particular purpose.

a. Evidence Based on Test Content

The content of a test includes the format of the test items, tasks, or questions, and administration and scoring procedures. Thus, test developers gather much of the validity evidence based on test content during the development process, when the developers are writing, reviewing, piloting, and revising the items and administration procedures. Careful documentation of these activities is an important step in gathering validity evidence related to content.

Further validity evidence related to test content is obtained by examining the relationship between the content of a test and the construct the test is intended to measure. One way to examine validity evidence based on test content is to assess (quantitatively or qualitatively) how well the test represents the domain of content it is intended to measure. The items on a test should adequately cover the important aspects of the underlying construct, without including other, non-related aspects from other constructs. As an example, one might expect that a test designed to measure a person's knowledge of fishing would contain items about the different

types of fish; methods of fishing; types of fishing poles, lures, and sinkers; and perhaps even about fishing boats and nets. If only one or two of these aspects were covered, the test may not provide an adequate assessment of a person's full knowledge of fishing. In this case, the test would underrepresent the construct it was intended to measure. If, on the other hand, the fishing test also contained items about scuba diving and snorkeling, our snapshot of a person's fishing knowledge would be blurred by his or her performance on the items not directly related to fishing. In this case, the test would contain *construct-irrelevant variance*. Both construct underrepresentation and construct-irrelevant variance pose threats to validity.

Content-related evidence of validity also encompasses evidence supporting the use of the test for a specific purpose. If one were gathering validity evidence related to content for the fishing test used in the example above, one might examine how the scores from the test will be used. Perhaps the owner of a chartered fishing tour company wanted to use the fishing test to screen potential employees. He consults with an expert on the subject of fishing; after close examination the expert finds that the set of items on the test appropriately covers the range of knowledge and skills that a fishing guide needs to possess, without covering topics unrelated to fishing. This information, then, would be one piece of evidence to support the use of the fishing test scores for screening potential tour guides.

b. <u>Evidence Based on Response Processes</u>

For certain types of tests, information about the processes in which examinees engage may be helpful in constructing a validity argument. Researchers use this information to determine the fit between the intended construct and the actual response strategies that examinees or raters use (AERA, APA, & NCME, 1999). For instance, if test developers were creating a test to measure reasoning, they might gather information about response

processes by questioning examinees about the cognitive strategies they used during the test. Alternatively, for a test that utilizes raters or scorers, evidence of validity based on response processes might include quantitative data showing that the raters used the scoring criteria appropriately and consistently when rating all examinees.

c. Evidence Based on Internal Structure

The internal structure of a test refers to the relationship between a test's items and parts, and how they conform to the construct upon which the test is built. To examine validity evidence based on internal test structure, it is necessary to understand the theoretical framework underlying the test construct. If one assumes the test is unidimensional, then the items should correlate highly with each other. On the other hand, if one assumes that the test is multidimensional and composed of two or more subparts, then the items in each subpart should correlate highly with each other and less so with items from the other subparts of the test. If a theoretical framework included the hypothesis that an item or group of items would function differently for one or more subgroups, one could use differential item functioning to determine whether this was the case or not.

d. Evidence Based on Relations to Other Variables

This type of validity evidence can be gathered by examining a test's relationship with another, related criterion. This is referred to in the current *Standards* as a "test-criterion relationship." In contrast to earlier conceptions of criterion-related validity, which often didn't take into account the quality of the criterion measure, the current definition in the *Standards* stresses that in order for the study of a test-criterion relationship to be useful, the criterion measure must be relevant, reliable, and valid for the interpretation of the test score for a given purpose (p. 14). The criterion variable in a validity study could be categorical, such as

membership in a group, or it could be continuous, such as scores on another test or measure that is designed to measure the same (or different) construct for the same population. Test-criterion relationships can be predictive, such that one can use scores from one measure to accurately estimate performance on the second, related measure administered at a later time. Alternatively, the study of the relationship between two measures could be concurrent, where both measures were administered at or about the same time.

Another way to examine validity evidence based on relations to other variables is to study how well test scores correlate with other similar or different measures. If we were designing a test measuring knowledge of algebra, we would hope that scores from our test would show strong correlations with scores from other well-constructed algebra tests. This type of study of the relationship between a test and another measure of the same or similar construct might provide *convergent* validity evidence. In contrast, we would probably hope that our algebra test would not show strong correlations (positive or negative) with scores on a vision screening, since there is no research linking skill in algebra to vision. This type of study—the examination of the relationship of a test's scores to a measure of another, unrelated construct—might provide evidence of *divergent* validity.

e. <u>Evidence Based on Consequences of Testing</u>

The validity of a test's score interpretation may be called into question when one can trace the score differences between examinees to construct underrepresentation or construct-irrelevant components (AERA, APA, & NCME, 1999). Validity evidence based on consequences of testing, therefore, would show that differences in test performance are directly related to differences in the abilities of the examinees on the construct being measured and not some other intervening factor. Additionally, if claims supporting a test's use include benefits

other than those directly related to the test scores themselves (such as improved motivation or efficiencies), either intended or unintended, then evidence based on consequences of testing could support these claims with qualitative information or quantitative data.

2. An Additional Framework for Examining Validity Evidence

The sources of validity evidence outlined previously provide a general framework to guide the building of a validity argument. In this section I describe an additional, more specific validity framework that Wolfe and Smith (2007b) proposed. This framework addresses validity specifically through the use of Rasch models, and it draws upon the properties of Rasch models to extend the scope of evidence gathering beyond what a Classical Test Theory (CTT) approach to collecting validity evidence provides. This is the framework I use in this study to answer the research questions outlined later in this chapter. Wolfe and Smith agreed with the idea of a unified concept of validity but drew upon the terminology and classification system that Messick (1995) proposed, combined with additional types of validity evidence that the Medical Outcomes Trust (Scientific Advisory Committee of the Medical Outcomes Trust, 1995) identified. Below, I provide a brief introduction to Rasch measurement models, followed by a summary of the types of validity evidence that Wolfe and Smith outlined. For each type of validity evidence, I discuss how it relates to the framework outlined in the *Standards,* and how using a Rasch measurement approach to analyze data can contribute additional validity evidence beyond what is provided through the use of a CTT approach to test data analysis.

a. The Rasch Measurement Models

In traditional measurement theory, referred to herein as Classical Test Theory (CTT), the difficulty of a dichotomously scored item is regarded as the proportion of examinees in some sample who respond correctly to the item. The correlation of this proportion

with the total test scores is thought to reveal the ability of the item to discriminate between high- and low-ability examinees. Finally, an examinee's ability is explained by his or her standing relative to peers in the particular sample (Wright & Stone, 1979). The practical concerns about this measurement method are that the difficulty of an item is dependent upon the abilities of the particular sample of examinees that take the item, and the ability of an examinee is dependent upon the difficulty of the particular test he or she takes. Wright and Stone called this type of measurement "uncomfortably slippery" (p. xi).

In an attempt to overcome the shortcomings of CTT, Danish mathematician Georg Rasch (1960) first introduced the use of the logistic function in the analysis of dichotomously scored test items. This model contains two parameters: the ability of the examinee, B_n, and the difficulty of the item, D_i. It is written as

$$P_{ni} = \frac{e^{B_n - D_i}}{1 + e^{B_n - D_i}}. \tag{2.2}$$

where P_{ni} is the probability of examinee n correctly responding to item i (Rasch, 1960). Using only these two pieces of information, the model predicts what will happen when an examinee encounters an item. If the examinee's ability is higher than the difficulty of the item (i.e., if $B_n - D_i$ is positive), then P_{ni} will be greater than .5. Conversely, if the examinee's ability is less than the difficulty of the item (i.e., if $B_n - D_i$ is negative), then P_{ni} will be less than .5. Finally, items that are perfectly targeted to an examinee's ability (i.e., $B_n = D_i$) result in a P_{ni} equal to .5.With the Rasch model, one can use observed item scores to estimate the difficulties of the items on a test, regardless of the abilities of the particular examinees who have taken the test. Likewise, one can use item scores to estimate the abilities of the examinees, regardless of the difficulties of the particular set of items chosen to appear on the test.

For polytomously scored items, Wright and Masters (1982) later expanded the dichotomous model to its partial credit form:

$$P_{nik} = \frac{e^{B_n - D_{ik}}}{1 + e^{B_n - D_{ik}}}, \qquad k = 1,2, \ldots m_i, \tag{2.3}$$

where P_{nik} is the probability of examinee *n* receiving a score in category *k* for item *i*, and D_{ik} is the difficulty of the *k*th category transition, or the point where a score in category *k* becomes more probable than a score in category *k*-1. The partial credit model is appropriate for use with items that have different score categories, and therefore different step difficulties (i.e., different values of D_{ik}). These two models and a number of other derivations—collectively referred to as the Rasch family of measurement models—have three advantages that distinguish them from other measurement models, if the data fit the model. First, the estimates of the parameters (item difficulty and examinee ability) are reported on the same interval scale in a common metric (referred to as the "logit" scale). Second, these parameter estimates are neither sample nor test dependent; in other words, the item difficulty estimates are freed from the distribution of examinee ability, and the examinee ability measures are freed from the distribution of item difficulty. Third, one can use the probability expression above to produce expected score values, which one can then compare to actual scores to produce measures of fit for items and examinees.

b. Evidence Relevant to the Content Aspect of Validity

This type of evidence is analogous to evidence based on test content in the *Standards*, and Wolfe and Smith (2007b) described evidence based on content in much the same way as the *Standards* do. The focus is on the relevance and representativeness of the test content in relation to the construct that the test proposes to measure. Expert review and documentation of the development process are essential when providing evidence of validity related to content. In

addition, Wolfe and Smith described how one can use the point-measure correlation and the item mean-square fit indices in the Rasch framework to assess the technical quality of the test items.

The point-measure correlation provides a correlation between the vector of scores on an item and the Rasch ability measures for those examinees that examiners scored. A strong positive point-measure correlation for an item indicates that, in general, examiners assigned higher scores to examinees with high ability measures and lower scores to examinees with lower ability measures. In the context of instrument construction, items with high point-measure correlations are desirable because they contribute useful information for separating high- and low-ability examinees.

The Rasch mean-square item fit indices provide a means for assessing how well the data fit the model. Under the Rasch model, as described previously, we can hypothesize how an examinee will perform on an item based on the examinee's ability, B_n, and the difficulty of the score category, D_{ik}. In the case of the BDI-2, for example, when an examinee's ability measure is lower than the step difficulty measure of a 2-point score, the examinee is less likely to receive a score of 2 on that item than if his or her ability measure is greater than the difficulty measure of the 2-point score. Following this, for every item, the model provides an expectation about what will happen when an examiner scores on examinee on an item. For example, if an examinee's ability measure is greater than the step difficulty measure for a 2-point score on that item, then it is most likely that the examiner will assign a score of 2 to the examinee's performance on the item. Fit indices allow us to analyze the discrepancy between what is expected (under the model) and what is observed (in the data). Rasch item mean-square fit indices are calculated by taking the sum of the squared residuals (the expected minus observed scores) and averaging them over the total number of examinees the examiners scored. The expected value of the item mean-

square fit indices is 1.0. Item mean-square fit values close to 1.0 imply that the examiners, in general, assign scores for the item in a way that is consistent with the model expectations. In other words, if the item is fairly difficult, the examiners tend to assign higher scores to only the higher-ability examinees, while the examiners tend to assign lower scores to the lower-ability examinees. If an item's mean-square fit index is higher than 1.0, the item may be measuring something other than the intended construct, may be vague or poorly written, or may contain scoring criteria that are unclear. For instance, in the case of the BDI-2, item mean-square fit indices greater than 1.0 may indicate that examiners do not understand the scoring criteria for 2-, 1-, and 0-point scores. If the item mean-square fit index is much less than 1.0, examiners may not have used the scoring criteria in a way that distinguished between higher- and lower-ability examinees. Rather, they may have overused some of the categories, giving many examinees the same score on the item. While this does not necessarily degrade the quality of the item's measurement, it may cause overinflation of test reliability statistics (Linacre, 2012). Thus, Rasch mean-square item fit indices provide validity evidence related to test content.

c. Evidence Relevant to the Substantive Aspect of Validity

Wolfe and Smith (2007b) described the substantive aspect of validity in much the same way the *Standards* describe evidence related to response processes. The focus is on the analysis of the item responses themselves and, for test items that call upon an examinee to exercise specific cognitive processes, the degree to which examinees actually engage in those cognitive processes. Wolfe and Smith suggested some types of Rasch-based validity evidence related to the response processes of examinees: analysis of Rasch examinee fit statistics, confirmation of the theoretical item hierarchy, distractor analysis for multiple-choice items, and rating scale functioning for polytomously scored items.

Similar to the item fit statistics described previously, Rasch examinee fit statistics show whether or not examinees are responding to the items in a way that is consistent with what the model expects. In general, according to the Rasch model, higher-ability examinees are more likely than lower-ability examinees to receive high scores on difficult items. If the data show unexpected patterns of scores for certain examinees, it may be an indication that those examinees have specialized knowledge or skills, or targeted developmental weaknesses. Additionally, it may be an indication of differential item functioning (DIF) (Wolfe & Smith, 2007b).

One unique feature of output from a Rasch analysis is that items are ordered by difficulty along a continuum, and this ordering is invariant over the entire ability continuum, if the data fit the model. Using this item hierarchy, test developers can evaluate whether the intended item hierarchy indeed matches the one that the examinee scores reveal. In other words, do the items that were intended to be difficult actually require a higher ability to get a higher score than those that were intended to be easy? Or, in the context of my study, are the items that measure higher-level developmental skills more difficult than the items that measure lower-level developmental skills? If the item hierarchy is consistent with the theory on which the test is based, this provides evidence relevant to the substantive aspect of validity.

For assessments containing polytomously scored items, Wolfe and Smith (2007b) suggested that an examination of the rating scale functioning can provide important information about whether or not examiners are using the score scale in a manner that is consistent with the intentions of the assessment developer. In both the rating scale model and the partial credit model, the examiner's assignment of a score in each successive ordered category implies that the examinee has exhibited more of the trait being measured than does a score assigned in the next lower category. Each BDI-2 item has its own rating scale. In other words, although the scores of

2, 1, and 0 generally correspond with skills that are "fully emerged," "emerging," and "not yet emerged," the criteria for assigning these scores differ for each item. In this context of the BDI-2, then, when an examiner assigns an examinee a score of 2 on a BDI-2 item, it implies that the examinee exhibits a higher degree of development on the skill that item measures than an examinee who received a score of 1 on the item. If the data suggest that this is not the case (i.e., if examiners tend to assign higher-ability examinees a score of 1 on an item more frequently than a score of 2), then this would indicate that, for some reason, examiners are not assigning scores in a way that is consistent with what the test's underlying "developmental milestones" theory would assert. On the other hand, ordered score categories, with each score category being the most likely to be assigned at some point along the continuum of the underlying latent trait, would provide evidence related to the substantive aspect of validity.

d. Evidence Relevant to the Structural Aspect of Validity

In the Wolfe and Smith (2007b) framework, this type of validity evidence very closely mirrors the *Standards*' evidence related to internal structure. In addition to the subtest intercorrelations, however, Wolfe and Smith asserted that a Rasch dimensionality analysis can contribute additional evidence related to the test structure. A requirement of the more commonly used Rasch models is that the data are unidimensional; that is, that the test items measure one and only one dominant dimension. In fact, according to Wright and Masters (1982), this is a requirement of all meaningful measurement. In their words, "The idea of measurement contains the image of a single line of inquiry, one dimension, along which objects can be positioned on the basis of observations which add up" (p. 8). If a set of test items measures more than one dimension, the purpose of the measurement, they argued, may not be met. A dimensionality analysis can indicate the degree to which the items in a test measure a single,

measurable trait. In the Rasch model, point-measure correlations and item-fit indices can assist with identifying items that contribute to multidimensionality. Smith (2004) presented evidence that Rasch standardized fit statistics, used in conjunction with a principal components analysis of residuals, can successfully identify which items, if any, contribute to multidimensionality. If this type of analysis indicates that a set of items measures one distinct dimension, this supports the unidimensionality assumption of the Rasch model and, in turn, provides evidence relevant to the structural aspect of validity.

Wolfe and Smith (2007b) also noted that indicators that the measurement model's requirements are satisfied can contribute to the structural aspect of validity. In the case of the Rasch model, the requirement of unidimensionality was noted above. In addition to unidimensionality, the Rasch model also assumes local independence; that is, it requires that the score to one item not be dependent upon the score to another item, after controlling for ability. Violations of local independence may present threats to the structural aspect of validity.

e. Evidence Relevant to the Generalizability Aspect of Validity

Wolfe and Smith (2007b) argued that evidence for this aspect of validity could extend beyond the traditional test reliability analyses that are common in CTT. They asserted that this type of validity evidence should not only account for the invariance of item calibrations across time and contexts, but also the invariance of examinee calibrations, or "examinee measures," across time and contexts.

The Rasch model's approach to providing this type of validity evidence involves conducting an analysis of item calibration invariance. Differential item functioning (DIF) occurs when an item has difficulty measures that vary across contexts or subgroups of examinees. If DIF is present, it may mean that the item is biased against one or more groups of examinees,

indicating that the item may not be a useful indicator of the measured trait. In the context of the BDI-2, DIF would indicate that examiners systematically assign lower scores to children from one or more subgroups on the item. Ruling out DIF on a set of test items lends support to the generalizability aspect of validity; that is, the test items maintain their meaning across subgroups of examinees.

The Rasch model can also be used to assess the internal consistency of test scores. The Rasch model provides internal consistency estimates for both examinees and items. Examinee internal consistency tells us how well examinee performance on the items spreads out the examinees along an ability continuum. Item internal consistency tells us how well the measurement of the examinees succeeds in separating the items so that a meaningful hierarchy can be established.

f. Evidence Relevant to the External Aspect of Validity

This type of validity evidence is similar to the *Standards*' relations to other variables; however, Wolfe and Smith (2007b) provided an additional source of validity evidence that one can obtain through the use of a Rasch model approach. They asserted that the capacity of a test to detect whether examinees' performance has changed over time is an example of evidence related to the external aspect of validity. Change in examinee performance can occur over time or as a result of introducing some intervention. The Rasch index for determining the extent to which an instrument is sensitive to change is the examinee strata. This index indicates how many distinct levels of examinee ability are distinguishable, given the ability measures obtained from examinee performance on a set of items. In other words, if the item difficulties were well dispersed along the continuum from low to high, the number of separate strata of examinee ability would be higher than if the item difficulties were more homogeneous. A high

examinee strata index provides evidence related to the external aspect of validity, with high examinee separation indices indicating that the test items would be more likely able to detect change in ability over time, or following an intervention.

g. Evidence Relevant to the Consequential Aspect of Validity

The *Standards'* discussion of the validity evidence related to the consequences of testing primarily focuses on discerning between the intended and unintended consequences of test use, and examining the unintended consequences for evidence of construct underrepresentation and construct-irrelevant variance components. Wolfe and Smith (2007b) asserted that the processes for setting cut scores (or "standard setting") are also important pieces of validity evidence related to the consequences of testing. Cut scores derived through a standard-setting process ultimately determine an examinee's performance level—and the interpretation of his or her ability—on a criterion-referenced test. Therefore, issues of bias, fairness, and distributive justice in the setting and application of cut scores "have a direct relationship with the consequential aspect of validity" (Wolfe and Smith, 2007b, p. 224). The authors argued that the standard-setting process must be carefully planned, carried out, and documented, and that this careful planning and execution must provide evidence related to the consequential aspect of validity.

h. Evidence Related to the Interpretability Aspect of Validity

One of the major advantages of the Rasch approach over the CTT approach to analysis is in the interpretability of the test scores. Woodcock (1999) described the "powerful interpretation features that become accessible when person abilities and item difficulties have been calibrated on a common Rasch scale" (p. 105). In a Rasch analysis, the meaning of an examinee's ability measure is directly related to the content of the items.

Although the interpretability aspect of validity is not mentioned in the *Standards*, Wolfe and Smith (2007b) suggested that the clear communication of a test score by assigning "qualitative meaning to quantitative measure" (p. 227) contributes to validity.

For example, a Rasch-based "Kidmap" (Wright, Mead, & Ludlow, 1980) is a visual display that contains both norm-referenced and criterion-referenced test information for an examinee. The criterion-referenced information includes the Rasch-derived probabilities of success on each item, as well as information about unexpected scores to items (i.e., hard items that the examiner scored as full credit, and easy items that the examiner scored as no credit) and the examinee's position with respect to specified cut scores. The end user can also superimpose the ability measure for the examinee onto a norm-referenced distribution. For example, a traditional normal score distribution with standard deviations would allow the end user to see where the examinee's score falls with respect to the normal curve.

Wolfe and Smith (2007b) suggested that Kidmap-type displays are useful tools that can contribute evidence to the interpretability aspect of validity by allowing clear communication of not only traditional norm-referenced information about an examinee's performance, but also criterion-referenced information (i.e., the examinee's performance on individual items relative to the item content through a description of the specific skill assessed by the item, the model-expected probability of full-credit score on each item, and the examinee's performance in relation to any specified cut scores).

3. Evidence for the Validity of the BDI-2 Scores

Researchers conducted extensive validity studies with the BDI-2 during the development period. The BDI-2 Examiner's Manual (Newborg, 2005b) contains descriptions of and results of these studies. The validity evidence documented in the BDI-2 Examiner's Manual

is organized according to the framework that the *Standards* defined. In this section, I will briefly summarize that evidence.

a. Content-Related Evidence of Validity for the BDI-2

Because the BDI-2 assesses widely accepted developmental milestones in children, evidence of validity based on test content necessarily includes expert judgments about the breadth and depth of the item set for adequately measuring a child's development. The experts need to agree that the items that the test author chose to represent each of the five developmental domains are a fair representation of the milestones that all typically developing children achieve, that attainment of these milestones is a necessary requirement for typical development, and that these milestones are represented in an order consistent with the order they occur during a child's development. At several times during the course of the BDI-2 item development, the test author consulted with experts in the various areas of cognitive development, occupational therapy, physical therapy, and speech/communication to ensure that the items included were indeed important developmental milestones, that the items were grouped appropriately into domains and subdomains, and that the item content was adequately measuring the target skill or ability. Where necessary, the test author made changes to the items or to the subdomain structure to reflect the recommendations of the expert reviewers.

The BDI-2 items each allow for up to three different administration procedures—Structured, Interview, and Observation—designed to elicit the information necessary to score the item. Examiners may choose which administration procedure to use for each item. The testing situation itself dictates the choice of administration procedure. For example, some children may not behave in a formal testing situation as they would behave in a familiar situation. In these cases, the examiner may choose to use the parent/caregiver interview procedure to administer

appropriate items. Additionally, a formal testing environment may not necessarily elicit all behaviors. In these instances, the item instructions may allow the examiner to observe the child in a natural setting for a period of time, or interview the parent or caregiver about the child's typical behavior. Documented evidence of validity based on test content in the BDI-2 Examiner's Manual includes a study to support the equivalence of the administration procedures (Pomplun & Custer, 2004). This study employed a many-facet Rasch model analysis to examine the 275 items in the BDI-2 Tryout Edition that offered more than one administration procedure. This study provided strong evidence that the administration procedure that the examiner chose did not introduce construct-irrelevant variance by changing the difficulty of the item.

b. Validity Evidence Supporting the Internal Structure of the BDI-2

The BDI-2 Examiner's Manual (Newborg, 200b) contains a variety of validity evidence related to the internal structure of the test. The first is factor-analytic evidence to support the domain and subdomain structure of the test. This type of evidence is especially important for the BDI-2, given that the subdomain structure is intended to match the legislative requirements for placing young children in intervention programs. In the study reported in the examiner's manual, the researchers proposed and tested four different models to see which model best fit the data. The results demonstrated that the five-factor model was the best of the four, and the factor loadings of the subdomains on their respective domains were in all cases adequate, and in many cases, substantial. This study's results provide support for the claim that the five BDI-2 domains are indeed five separate domains, each made up of narrower subdomains.

Another type of validity evidence related to internal test structure included in the BDI-2 manual is the presentation of growth curves for the five domains. These curves demonstrate that

the BDI-2 is sensitive to differences in age-related changes in developmental rates, as the literature on development hypothesizes, and that these trajectories plateau where anticipated for each of the five domains. Wolfe and Smith (2007b) would consider both of these types of validity evidence as relevant to the structural aspect of test validity.

c. Validity Evidence Supporting the BDI-2's Relationship to Other Variables

In their framework Wolfe and Smith (2007b) referred to a test's relationship to other variables as external validity evidence. This is the area in which the BDI-2 Examiner's Manual provides the most extensive documentation. The BDI-2 Examiner's Manual (Newborg, 2005b) presents several studies in which researchers compare scores on the BDI-2 with scores on other early childhood developmental assessments. In some cases, the criterion assessment was not a full-scale battery covering all five developmental domains, but rather a more narrowly focused battery covering only one (or a few) of the domains that the BDI-2 measured. In these cases, the researchers used the relevant BDI-2 domains for comparison, and they provided correlations as convergent validity evidence for the BDI-2. In some additional cases, researchers compared the BDI-2 domains that do not measure the same constructs as the criterion test's domains. Low correlations between scores on BDI-2 domains and scores on non-corresponding domains of other early childhood assessments provided divergent validity evidence.

d. Validity Evidence Supporting the BDI-2 Response Processes

When building a validity argument for clinical tests, it is important to demonstrate that the test scores can distinguish between different subgroups of examinees, especially those subgroups whose members the test purports to identify. In the case of the BDI-2,

the test is intended to identify those children with developmental delay in any of the five domains. To support these score interpretations, the BDI-2 Examiner's Manual presents studies in which examiners administered the BDI-2 to children with diagnosed autism, cognitive delay, developmental delay, motor delay, and speech delay. The manual reports the classification accuracy of the BDI-2 for each of these groups of children. Classification accuracy can be described using a sensitivity value and a specificity value. The sensitivity value is the probability that a child with a delay will be correctly identified using BDI-2 scores. Sensitivity values ranged from .75 for the motor delay group to .93 for the autism group. The specificity value is the probability that a child without a delay would be classified as typically developing using BDI-2 scores. Specificity values ranged from .75 for the speech and language delay group to .91 for the autistic group. The high sensitivity and specificity values support the use of the BDI-2 as an instrument for identifying children with these types of delays.

e. Additional Evidence for the Validity of the BDI-2 Scores

Since the publication of the BDI-2 in 2005, several studies have added information to the body of validity evidence for the use of the test scores in a variety of settings. Elbaum, Gattamora, and Penfield (2010) evaluated the utility of the BDI-2 Screening Test scores for use in states' child outcomes measurement systems. They were interested in two questions: (1) What are the psychometric characteristics of the BDI-2 Screening Test when used for a sample of children referred for evaluation? (2) How accurately does the BDI-2 Screening Test classify children as having or not having developmental delay? To answer the first question, the researchers administered the BDI-2 Screening Test to 142 children referred for evaluation in early childhood. The researchers utilized Rasch measurement principles to study the first research question, and determined that the acceptable item fit statistics and high point-measure

correlations of the screening items contribute usefully to the measurement of the construct. To investigate the second research question, the authors computed sensitivity and specificity values for each domain and age range in the BDI-2 Screening Test. They concluded that the choice of cut score (-1.0 SD, -1.5 SD, or -2.0 SD) impacts the sensitivity and specificity values, and recommended that for a referred population, a cut score of -1.5 SD maximized sensitivity while minimizing false positive referrals.

Sipes, Matson, and Turygin (2011) studied the use of the BDI-2 as a screening tool to identify young children with autism spectrum disorders (ASD). Using a sample of 1,301 children referred for early childhood evaluation, they found that a BDI-2 cutoff score of -1.5 SD correctly identified 94% of the children in the subgroup of the sample who had been identified as having ASD. The researchers concluded that the BDI-2 can be a useful measure for identifying children who may be at risk for ASD, even when it is administered as part of non-specific clinical evaluation.

Matson, Hess, Sipes, and Horovitz (2010) compared the BDI-2 domain and total scores of 28 toddlers who were born prematurely, who had Down syndrome, or who had been diagnosed with global developmental delay. They found unique characteristics of the BDI-2 score profiles for each group of children. All children had significantly lower mean BDI-2 scores than average children their age; however, the children with Down syndrome and global developmental delay had relatively lower BDI-2 total scores than the children who were born prematurely. Additionally, children with global developmental delay had significantly lower BDI-2 Personal-Social Domain scores than the other two groups, and both the global developmental delay and Down syndrome groups had significantly lower BDI-2 Motor Domain scores than the premature group. This research provides support for the BDI-2 total and domain profile scores for use in identifying children with different types of disabilities.

G. Defining the Research Questions

In this section, I introduce the propositions and research questions that I use to focus my study. The BDI-2 Examiner's Manual contains extensive documentation of the validity evidence that the author and publisher gathered during the development of the test; however, Wolfe and Smith (2007b) suggested that researchers can gather additional types of validity evidence if they use the Rasch model to analyze test data. The following propositions, and associated research questions, focus on how these types of additional evidence gleaned through Rasch analysis either support or refute the use of the BDI-2 Gross Motor Subdomain scores for identification, progress monitoring, and score reporting.

Proposition 1) BDI-2 Gross Motor (GM) Subdomain scores are useful for accurately describing mastery or non-mastery of gross motor developmental milestones in childhood.

Research Question 1) Does substantive validity evidence support the current uses of BDI-2 Gross Motor subdomain scores to make inferences about children's development?

According to Wolfe and Smith (2007b), substantive validity evidence provides information about "the degree to which theoretical rationales relating to … item content … adequately explain the observed consistencies among item responses" (p. 207). In the case of the BDI-2 Gross Motor subdomain scores, evidence to support the substantive aspect of validity would show that the scores are an accurate measure of a child's level of mastery of the underlying gross motor milestones. If this is true, examiners will assign higher scores to children who demonstrate mastery of the developmental skills assessed by the test, and lower scores to children who demonstrate less development or ability.

Proposition 1.1) BDI-2 Gross Motor subdomain score categories are appropriate for obtaining information about a child's current level of gross motor development.

Research Question 1.1) Do the examiners use the BDI-2 score categories of 2, 1, and 0 as expected to label behaviors that are fully emerged, emerging, and not yet emerged?

The BDI-2 is based on a developmental model, where lower scores represent earlier stages of development and higher scores represent higher levels of development. In developmental assessments, "the model of development on which the assessment is based can be used as a basis for validation … by allowing one to explicitly compare theory-based predictions concerning item difficulties with empirical difficulties" (Wolfe & Smith, 2007a, p. 107). Nearly all of the BDI-2 items are scored 2, 1, or 0, allowing children to receive partial credit for those items on which they can display even some emergent skill or ability. Because of the wide variability in children's acquisition of developmental milestones, early childhood diagnosticians, who may be hesitant to withhold credit for an item on which a child can succeed "somewhat," find this feature of the BDI-2 attractive. This three-category scoring system, however, is only useful if the score categories function in the expected way. For example, for each BDI-2 item, the likelihood of a child scoring 1 rather than 0, and 2 rather than 1, should become greater as the child's developmental level increases. If, for example, the score of 1 is never most probable along the continuum of development, then examiners are not using the item scoring rubric as intended in accordance with the developmental milestone theory; either the category of 1 describes only a very small range of the continuum, or very few children actually received a score of 1 on that item. By examining each Gross Motor item's category thresholds and probability curves, I attempt to identify items that examiners may not interpret in the way they were intended when written. If the rating scale structure of each item seems to function adequately in practice, this would provide evidence relevant to the substantive aspect of validity.

Additionally, I examine the item mean-square fit statistics to determine whether there are any values less than 1.0, which may help identify Gross Motor items on which the examiners have not used the full range of score categories. Taken together with the information on rating scale structure gathered above, evidence regarding the item fit statistics may contribute to the substantive validity of the BDI-2 score interpretations.

Proposition 1.2) The BDI-2 Gross Motor subdomain item hierarchy accurately represents the underlying milestone theory of development.

Research Question 1.2) In the standardization dataset, are there any anomalous examinee score strings that may have degraded the quality of the item calibrations?

The test publisher utilized a stratified sampling plan during the BDI-2 standardization study. The psychometricians who developed the test norms utilized all the data the examiners gathered during the course of the study; this amounted to 250 cases per age group. No cases were dropped, and no mention is made in the BDI-2 Examiner's Manual about whether psychometricians analyzed the data for anomalous score strings. However, as Wolfe and Smith (2007b) noted, anomalous score strings can degrade the quality of item calibrations. Using Rasch-based methods suggested by Wolfe and Smith, I inspect the examinee fit statistics and examinee score strings in the standardization dataset for the Gross Motor items, looking specifically for examinees whose unusual score strings may have degraded the quality of the item calibrations.

Proposition 2) The BDI-2 Gross Motor subdomain scores provide meaningful measures of the distinct, domain-specific abilities that contribute to a child's development.

Research Question 2) Does structural validity evidence support the use of BDI-2 Gross Motor subdomain scores to make inferences about children's development?

The domain structure of the BDI-2 makes it attractive to practitioners responsible for identifying developmental delay in the areas of Cognitive, Motor, Communication, Personal-Social, and Adaptive development, which are specified under IDEA. Additionally, BDI-2 subdomain scores can provide practitioners with information about a child's relative strengths and weaknesses within each of these developmental domains. This developmental "profile" feature of the BDI-2 assists practitioners in planning goals and interventions focused specifically on those areas of concern to help children advance their development. Finally, BDI-2 subdomain scores are well-suited to the requirements for annual OSEP progress reporting because they map directly to the three OSEP outcomes (Early Childhood Technical Assistance Center, 2007). If the BDI-2 Gross Motor subdomain scores are to be useful for the purposes of diagnosis, planning, and reporting, however, users need to be confident that the BDI-2 Gross Motor subdomain structure and scoring model adequately represent the underlying developmental abilities. According to Wolfe and Smith (2007b), structural validity evidence "appraises the fidelity of the scoring structure to the structure of the construct domain" (p. 213), and test developers may gather this type of evidence through the use of correlational and dimensionality analysis of the test's subparts. The BDI-2 Examiner's Manual (Newborg, 2005b) provides evidence that the structure of the BDI-2 conforms to theoretical understanding about the relationships between these developmental areas through the presentation of intercorrelations among the domain and subdomain scores, exploratory factor analysis, and structural equation modeling results. Wolfe and Smith (2007b) suggested that test developers and users can provide additional evidence to support the structural aspect of validity through an examination of the assumptions underlying a test's scoring model. In the case of the BDI-2, two assumptions of the Rasch partial-credit model are unidimensionality and local item independence.

Research Question 2.1) Do the BDI-2 subdomain data from the BDI-2 Gross Motor subdomain represent one dominant underlying dimension?

Some states require that examiners carry out separate assessments, for example, for gross motor and fine motor skills. This is because a child could conceivably have a delay in one area with normal functioning in another, and a composite (i.e., motor domain total) score may "mask" the lower-performing area. However, in the BDI-2, the narrow subdomains contain many fewer items than the more general domains. In order to be confident about the eligibility decisions made through the use of BDI-2 subdomain scores, early childhood diagnosticians need to be confident that the items they are administering in the narrow test subdomain tap into one and only one underlying developmental trait, and that the subdomain scores have sufficient reliability for making decisions about the abilities of individual children.

I gather evidence to support (or refute) Gross Motor subdomain score interpretations through a Rasch dimensionality analysis. Because items that significantly misfit the model may be measuring something other than the intended construct, I first examine item fit statistics for evidence of multidimensionality. I then use the iterative method that Smith (2004) described to investigate the structure of the BDI-2 Gross Motor subdomain using a principal components analysis (PCA) of standardized residuals. If the results from the analyses of the Gross Motor subdomain data do not suggest that there is a measurable second factor, then those results provide evidence related to the structural aspect of validity in support of the subdomain score interpretation.

Research Question 2.2) Do the data from the BDI-2 Gross Motor subdomain satisfy the Rasch model requirement of local independence?

The Rasch model requirement of local independence states that, after controlling for the examinee level on the underlying trait, an examinee's score on one item should not be dependent

upon the examinee's score on another item. The extent to which this requirement is met for the BDI-2 provides evidence to support (or refute) the use of the Gross Motor subdomain scores for making inferences about a child's gross motor development. Many BDI-2 items measure varying levels of the same type of skill; for example, multiple items measure feeding skills in various ways across the developmental continuum from infancy through later childhood. These items are not necessarily stand-alone items; they are "clustered," or related, by virtue of the underlying skill they are measuring. Using Rasch standardized residuals, I normalize a distribution of Pearson correlations of all pairs of items, and then compare each item pair to a predetermined value. If pairs of "clustered" item residuals show significantly higher correlations than the stand-alone item pairs, the clustered items may present violations of the local independence requirement. I assess the extent to which these violations, if they exist, pose threats to the structural aspect of validity for the BDI-2.

Proposition 3) BDI-2 Gross Motor subdomain scores are precise enough to locate an examinee's ability on the scale and reveal changes in ability over time.

Research Question 3) Does evidence relevant to the generalizability aspect of validity support the use of BDI-2 Gross Motor subdomain scores to make inferences about children's development?

Recall that Wolfe and Smith (2007b) suggested that this type of validity requires conducting other kinds of analyses beyond the traditional CTT reliability analyses, and that, by examining item calibration invariance and internal consistency estimates under the Rasch model, one can obtain additional sources of validity evidence to support the generalizability of the scores. The BDI-2 Examiner's Manual (Newborg, 2005b) reports that the publisher conducted a DIF analysis as part of the test development process. For this reason, I do not examine DIF in the

present study. I instead focus my investigation of generalizability evidence for the Gross Motor subdomain on the interpretation of Rasch estimates of examinee and item internal consistency, as suggested by Wolfe and Smith (2007b).

Research Question 3.1) Are the BDI-2 Gross Motor subdomain scores sufficiently reliable for making inferences about children's development?

Reliability is defined as the consistency of test scores across multiple administrations of the test. Reliability indices typically range from 0 to 1, with values closer to 1 indicating more stability in measures across multiple administrations. The Rasch model provides separation reliability indices for both examinees and items. Examinee-separation reliability indices tell us how well the set of items separates the group of examinees. According to Linacre (2012), examinee separation reliability of .80 or higher indicates that the test items are able to separate the sample of examinees into three or more distinct ability groups. Because the purpose of the BDI-2 is to identify examinees whose skills are not emerged, emerging, and mastered, it is important that the test be sufficiently reliable to separate the examinees into at least as many distinct ability groups. Likewise, Rasch item separation reliability indices tell us how well the items are spread across the continuum. Low item separation reliability would indicate that the sample size is too small (or the range of item difficulty is too constrained) to accurately locate the BDI-2 Gross Motor items on the item difficulty continuum (Linacre, 2012).

H. Chapter Summary

In this chapter, I reviewed the literature on the history of early childhood assessment and described how legislative decisions have influenced assessment practices. I then discussed the types of developmental assessments available for use with young children, and described some of the information that these instruments can provide. I introduced the instrument that is the

focus of this study, the Battelle Developmental Inventory, Second Edition, and described its development, administration, and scoring. Later in the chapter, I discussed validity from both the framework of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) and the framework that Wolfe and Smith (2007b) proposed. I also described the current validity evidence available for the BDI-2. Finally, I introduced the propositions and research questions that guide my study. In the next chapter, I introduce the methods that I followed to address each of these research questions.

III. METHOD

In this study, I used the data from the Battelle Developmental Inventory, Second Edition (BDI-2) standardization study to investigate the research questions that I posed in Chapter II. This investigation provides new validity evidence to support (or refute) the current uses of the BDI-2 for its documented purposes. Where the results of this investigation provided insight on potential improvements to the BDI-2, I synthesized the information I gathered during the analyses to suggest modifications to the test.

A. Instrument

For this study, I used the standardization data from the BDI-2 Gross Motor Subdomain. This subdomain consists of 45 items intended to measure a child's gross motor development from birth through age 7 years, 11 months. Each item contains a description of the skill to be measured, a list of the materials necessary to administer the item, standard procedure(s) to be used in the item administration, and a detailed scoring rubric. Table IV contains a list of the developmental skills that the Gross Motor items measure.

TABLE IV
DEVELOPMENTAL SKILLS MEASURED IN THE BDI-2 GROSS MOTOR SUBDOMAIN

Item	Developmental Skill Measured
1	The child maintains an upright posture at adult's shoulder without assistance for at least 2 minutes.
2	The child holds his or her head erect for 1 minute when held.
3	The child lifts his or her head and holds it up for 5 seconds while lying in the prone position.
4	The child lifts and turns his or her head from side to side while lying in a prone position.
5	The child brings his or her hands together at the midline.
6	The child turns his or her head freely from side to side while supported in a sitting position.
7	The child holds his or her head parallel to the body when pulled from a supine to a seated position.
8	The child moves his or her arms when a toy is in sight.
9	The child puts objects into his or her mouth.
10	The child moves an object from hand to mouth.
11	The child turns from a prone to a supine position unassisted.
12	The child intentionally secures a nearby object while in a prone position.
13	The child sits without assistance for at least 5 seconds.
14	The child makes stepping movements when held in an upright position.
15	The child moves 3 or more feet by crawling.
16	The child pulls himself or herself to a standing position while holding on to a solid object without adult assistance.
17	The child moves from a standing position to a sitting position while holding on to a solid object.
18	The child walks 3 or more steps with assistance.
19	The child stands in an upright position without support for 30 or more seconds.
20	The child creeps or crawls up 4 steps without assistance.
21	The child walks without support for 10 feet without falling.
22	The child moves from a sitting position to a standing position without support.
23	The child moves from a supine to a standing position using smooth, coordinated movements without support or assistance.

TABLE IV
DEVELOPMENTAL SKILLS MEASURED IN THE BDI-2 GROSS MOTOR SUBDOMAIN (CONTINUED)

Item	Developmental Skill Measured
24	The child maintains or corrects his or her balance when moving from a standing position to other, nonvertical positions.
25	The child walks up 4 stairs with support.
26	The child walks down 4 stairs with support.
27	The child runs 10 feet without falling.
28	The child kicks a ball forward without falling.
29	The child walks up and down stairs without assistance.
30	The child walks backward 5 feet.
31	The child throws a ball 5 feet forward with direction.
32	The child jumps forward with feet together.
33	The child walks forward 2 or more steps on a line on the floor, alternating feet.
34	The child walks down stairs without assistance, alternating feet.
35	The child imitates the bilateral movements of an adult.
36	The child bends over and touches the floor with both hands.
37	The child catches an 8-inch ball from 5 feet away, using both hands.
38	The child walks in a straight line, heel-to-toe, for 4 or more steps.
39	The child hops forward on one foot without support.
40	The child stands on each foot alternately with eyes closed.
41	The child walks a 6-foot line on the floor, heel-to-toe, with eyes open.
42	The child skips on alternate feet for 20 feet.
43	The child throws a ball and hits a target with the dominant hand.
44	The child jumps rope without assistance.
45	The child throws a ball and hits a target with the nondominant hand.

The considerable size of the complete BDI-2—450 items across 13 subdomains—made it impractical for me to conduct all my analyses on the entire test. Instead, I focused my research on one specific subdomain. I chose to examine the BDI-2 Gross Motor Subdomain in this study for three reasons. First, the Gross Motor Subdomain relies heavily on the Structured administration procedure, which is the procedure preferred by the test developers. Although Pomplun and Custer (2004) found that that the administration procedure that the examiner chose during the item tryout did not change the difficulty of the items, I wanted to minimize the chance of introducing construct-irrelevant variance into my study by choosing a subdomain in which the majority of the examiners used the Structured procedure to collect the standardization data. Indeed, 42 of the 45 items in the Gross Motor Subdomain offer a Structured administration procedure, and 17 of the items offer *only* the Structured procedure. For the 25 items offering the Structured procedure plus at least one other procedure, examiners in the standardization study used the Structured procedure most often (in 19 of the 25 items), evidence that it actually is the procedure that examiners prefer and use most often in this subdomain. Other BDI-2 subdomains, such as those in the Personal-Social Domain, rely more heavily on the Interview and Observation administration procedures.

Second, the items contained in published assessments of gross motor development are relatively consistent across assessments; there seems to be a high degree of agreement among early childhood developmental experts as to what constitutes "typical" gross motor development. In other words, this domain doesn't appear to be driven by specific theories as are other areas of development. Rather, there appears to be general agreement among developmental assessments that certain milestones should appear at specified ages (or ranges of ages) in early childhood, and the measurement of these milestones is very similar among published instruments measuring

gross motor skills. Finally, there appears to be consistency in item content across multiple editions of most available instruments for assessment of gross motor skills, including the two editions of the BDI. Therefore, this developmental domain may be less susceptible to trends in measurement techniques than, for example, the cognitive domain—where new and changing theories of brain development and cognition have, over time, impacted the way that cognitive development is conceptualized and measured.

B. Sample

The BDI-2 was standardized on 2,500 children who ranged in age from a few hours old to 7 years, 11 months, 29 days old. The sampling plan was designed to obtain 20 distinct 3-month age groups across the 8-year age span of the instrument. Each age group contains 125 children (e.g., 125 children ages 0-3 months, 125 children ages 4-6 months, and so on, up to 95 months of age). Within each age group, the sample is stratified by sex, race, ethnicity, region of the country, and socioeconomic level, with percentages matched to 2001 U.S. Census Bureau publications (Newborg, 2005b).

C. Data

1. Data Collection Procedures

The test publisher gathered the BDI-2 standardization data over a 14-month period in 2003 and 2004. The publisher hired independent examiners for the purpose of data collection. Examiners were typically professionals in the field of early childhood education and/or development, and a majority of the examiners had at least some prior experience administering the first edition of the BDI. Project staff trained examiners in groups on the administration and scoring of the BDI-2 items. Each 2-day training session consisted of item-by-

item administration instructions, information on selecting examinees and obtaining consent, and instructions about exclusionary criteria.

The publisher maintained control of the demographic makeup of the sample by specifying exact characteristics of examinees to be tested. To do this, the publisher provided examiners in each region of the country with a list of "target" examinees sorted by the demographic stratification variables. Examiners were then responsible for selecting examinees who matched the targets based on age, sex, race, ethnicity, and socioeconomic level. Once an examinee was tested, his or her "target" was removed from the list of available targets, so that other examiners would not be able to choose that exact same target.

Because the purpose of the standardization study was to collect information on the performance of typically developing children on the BDI-2 items, the authors and the publisher made the decision to exclude children from the study who had known disabilities or delays, or medical risk factors for delays. Additionally, only children from English-speaking families were included in the study. Examiners were instructed to administer the items exactly as written; no special testing accommodations were allowed during the data gathering.

Before testing began, the examiner obtained written consent from the examinee's parent or legal guardian. A trusted adult (parent or caregiver) was allowed to stay with the child during testing, if necessary, to help alleviate the child's fears or anxieties about the test session; however, examiners instructed adults to avoid unnecessary talking or "prompting" during testing. After testing was complete, the examiner provided the child's parent or guardian compensation in the form of a gift card for a local retailer.

2. Characteristics of the Dataset

The data for this study are a subset of the complete BDI-2 standardization dataset. The data file includes demographic information and scores for all 2,500 children tested as part of the standardization study; however, I utilized the scores for the Gross Motor Subdomain for this study. For each Gross Motor item, the data file contains scores of 2, 1, or 0 (corresponding to developmental skills that are "fully emerged, " "emerging," and "not yet emerged"). Additionally, for the 26 items that offer a choice of administration procedures, the file includes a code to designate whether the examiner administered the item using the Structured, Observation, or Interview procedure. The publisher removed all examinee names from the dataset prior to releasing the file to me. Examinees are distinguished from one another in the file using only unique 4-digit identification numbers.

As I explained in Chapter II, the BDI-2 employs the use of basal and ceiling rules to minimize testing time and to ensure that examinees do not encounter items that are much too easy or much too difficult. As a result, each BDI-2 examinee responds only to a subset of items in each domain, depending on the examinee's chronological age and developmental functioning. All the items below an examinee's basal level and all items above an examinee's ceiling level appear as "missing." He and Wolfe (2012) studied the effect of treating non-administered items above the examinee's ceiling level as missing responses during the calibration process, and found that the examinee ability parameters were more accurately recovered than when the non-administered items were treated as incorrect or fractionally correct, especially under maximum likelihood estimation procedures.

D. Analyses

In this section, I describe the analyses that I conducted to investigate the research questions I outlined in Chapter II. I use the validity framework that Wolfe and Smith (2007b) proposed. This framework addresses validity using Rasch methodology. Because the BDI-2 items each contain a unique set of score categories, the partial credit form (Wright and Masters, 1982) is the appropriate Rasch model for this application. The partial credit model is written as:

$$P_{nik} = \frac{e^{B_n - D_{ik}}}{1 + e^{B_n - D_{ik}}}, \qquad k = 1, 2, \ldots m_i, \tag{2.3}$$

where P_{nik} is the probability of examinee n receiving a score in category k for item i, and D_{ik} is the difficulty of the kth category transition, or the point where a score in category k becomes more probable than a score in category k-1. The partial credit model is appropriate for use with items that have different score categories, and therefore different step difficulties (i.e., different values of D_{ik}). For the Rasch analyses, I used the WINSTEPS computer software (Linacre, 2012).

I chose to use Wolfe and Smith's framework because it addresses validity in some unique ways that publishers do not typically consider during the development of most individually administered early childhood developmental tests, including the BDI-2. As I discussed in Chapter II, most of the validity evidence the author and publisher gathered and documented during the development of the BDI-2 is based on CTT analyses. By using the Wolfe and Smith (2007b) framework, I hoped to add to the already-existing body of validity evidence for the current uses of this test.

Because I believe they are the most relevant for the current uses of the BDI-2, I gathered three types of validity evidence: substantive evidence, structural evidence, and evidence based on generalizability. My analyses were designed to gather these three types of validity evidence.

Research Question 1) Does substantive validity evidence support the current uses of BDI-2 Gross Motor Subdomain scores to make inferences about children's development?

Recall from Chapter II that Wolfe and Smith's (2007b) definition of substantive validity evidence is similar to the evidence related to response processes cited in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999), but Wolfe and Smith suggested that examination of Rasch examinee fit statistics and rating scale functioning may provide important additional pieces of substantive validity evidence that CTT analyses do not provide. Below I pose two specific questions related to substantive validity that I addressed using Rasch methodology.

Research Question 1.1) Do the examiners use the BDI-2 score categories of 2, 1, and 0 as expected to label behaviors that are fully emerged, emerging, and not yet emerged?

The BDI-2 utilizes a 2-, 1-, or 0-point scoring system for each item. Although the possible scores for all items are the same, each item has its own scoring rubric, so that, for example, the requirements for a score of 2 differ from item to item. For most items, a score of 2 requires that the child displays mastery of the particular skill, a score of 1 requires that the child can display some emergence of the skill, and a score of 0 means that the child has not displayed any emergence of the skill.

Wolfe and Smith (2007b) suggested that an examination of rating scale functioning can indicate whether the examiners are utilizing the scale in a way that is commensurate with the intentions of the test developer. If so, this provides evidence for the substantive aspect of validity. For the BDI-2, this would require that for each item, examiners assign scores of 2 to children who display mastery of the skill, scores of 1 to children who display some emergence of the skill, and scores of 0 to children who do not display any emergence of the skill.

Linacre (2004) presented eight guidelines for the examination of rating scale functioning. I focused on the first five of these guidelines in my analysis because they are the most commonly used in practice. I analyzed the BDI-2 data using the partial credit model (Wright & Masters, 1982). Under the Partial Credit Model, the rating scale for each item is modeled separately; therefore, I applied the following guidelines that Linacre proposed to evaluate the rating scale for *each item* in the BDI-2 Gross Motor subdomain:

1. *Ensure that each score category of each item contains a minimum of 10 observations.* According to Linacre (2004), when the frequency is low for a given category, the threshold estimates for that category can be imprecise and unstable. Linacre suggested that category frequencies of 25 to 100 times the number of steps are optimal. The BDI-2 data contain at least this many scores for the 2 and 0 score categories for most items; however, I anticipated that there would be some items in the dataset for which there were fewer than 10 scores in score category 1. Because a score of 1 is only possible for examinees who were administered the item, and because the item is only administered if it is targeted to the examinee's ability, there will generally be lower frequencies for score category 1 than for the other two categories. For each item, I checked the frequencies of 2, 1, and 0 scores and noted score categories with frequencies less than 10.
2. *Analyze the observation distribution within each item.* Linacre (2004) suggested that a uniform distribution of scores across all the categories is best for calibration. For each item, I calculated the percent of the total scores in each of the score categories 2, 1, and 0.

3. *Analyze average measures within each category of each item.* If higher scores imply more of the measured trait or skill, then the average measures of examinees in each category should monotonically increase with each score from 0 to 1 to 2. For each item, I documented the average examinee measure for each score category, noting where the values did not increase monotonically from 0 to 2.
4. *Examine the mean-square outfit statistic for each score category of each item.* Outfit statistics greater than 2.0 indicate that there is more unsystematic variance than systematic variance (Linacre, 2004). I examined each category of each item to look for unexpected category usage, as indicated by outfit statistics greater than 2.0. In these cases, I inspected the data to see whether removing individual aberrant scores (or examinees) may resolve the issue.
5. *Ensure that category thresholds advance for each item.* Higher-ability examinees should have a greater probability of scoring 2 on any item than lower-ability examinees. The likelihood of scoring 0, 1, or 2 on any item can be graphed such that the three curves corresponding to the probability of receiving a score in each scoring category 0, 1, or 2 is modal at some point on the continuum. If this is not the case (i.e., if a score is never most probable), then the category thresholds for the item are disordered. This problem can be caused by irregular category frequencies, such as when a higher score is rarely assigned or when higher-ability examinees are assigned a score of 0 (Linacre, 2004). I examined the category thresholds for each item to determine if disordering occurred in the dataset.

In addition to examining the rating scale structure using Linacre's guidelines, I also examined the item mean-square fit statistics to determine whether any were less than 1.0. Typically, mean-square fit values less than 1.0 indicate overfit; in other words, the scores are too predictable (possibly due to a violation of local independence), indicating better-than-expected fit to the model. This does not usually degrade the measurement, but can artificially inflate reliability indices. However, Smith (1996) suggested that for polytomous items such as those in the BDI-2, item mean-square fit indices less than 1.0 might indicate that examiners are not utilizing the full range of score categories for that item.

Research Question 1.2) In the standardization dataset, are there any anomalous examinee score strings that may have degraded the quality of the item calibrations?

Recall from Chapter II that the Rasch model estimates what will happen when an examinee encounters an item. According to the model, higher-ability examinees should have a greater chance than lower-ability examinees of receiving a higher score on any item. Likewise, lower-ability examinees should be less likely than higher-ability examinees to receive a higher score on any item. One can identify unexpected scores (i.e., high-ability examinees receiving low scores on easy items or lower-ability examinees receiving high scores on more difficult items) by studying the examinee fit statistics. Identification of unexpected scores provides a procedure for evaluating the validity of each examinee's scores (Wright & Stone, 1979). Highly unexpected scores on the BDI-2 items might indicate that an examinee was tired, uncooperative, did not understand the instructions, or had specialized skills or abilities. Alternatively, highly unexpected scores might actually be indications of errors in the data that are independent of examinee behavior, such as scanning errors or examiner scoring errors. In either case, if the purpose of the Rasch analysis is for scale construction during the test development process, then these

unexpected scores do not contribute to the valid estimation of the item difficulties. In test development, it is common practice to evaluate examinee fit statistics for evidence of potentially invalid scores and then, if reasonable justification can be provided for doing so, selectively remove unexpected scores and re-estimate item difficulties. Once reasonable data fit to the model is achieved, the optimal item difficulty estimates can be used as anchor values in a final Rasch analysis including all previously removed scores to obtain examinee ability measures.

While there are no "rules" for acceptable fit statistics, I used a criterion of examinee mean-square outfit statistic > 1.5 to analyze the fit of the examinee score strings. Because the outfit statistic is sensitive to unexpected scores that are outside of an examinee's targeted ability range, it is relatively easy to identify and diagnose. In the context of the BDI-2, large (> 1.5) mean-square outfit values indicate that an examiner assigned a much lower or much higher score on an item than would be expected by the model, given the examinee's overall ability. In cases where examinees' score strings contained a number of unexpected high or low scores, I removed the unexpected score(s) from the dataset. In each case, I reran the analysis to determine if the score removals resulted in higher quality measurement as indicated by better overall model fit.

At each point in my analyses, I documented edits I suggested to the dataset or score category descriptions. After each round of edits, I compared the value of the change in the -2 log-likelihood ratio (-2LL) and the associated degrees of freedom to the critical value of χ^2 to determine if the edits resulted in better overall model fit. Additionally, I used the -2LL value to compute Akaike's Information Criterion (AIC) (Akaike, 1973) and the Bayesian Information Criterion (BIC) (Kass & Raftery, 1995; Raftery, 1995; & Schwarz, 1978). AIC and BIC are model selection criteria that account for model complexity in their evaluation of data-to-model fit. Lower values of AIC and BIC indicate better model fit. Table XVI in the Appendix contains

a description of each analysis run I performed in this study, including edits I made to the dataset, overall examinee and item mean-square outfit statistics, change values for -2LL and associated degrees of freedom, critical χ^2 values, and values of AIC and BIC.

Research Question 2) Does structural validity evidence support the use of BDI-2 Gross Motor Subdomain scores to make inferences about children's development?

Research Question 2.1) Do the data from the BDI-2 Gross Motor Subdomain represent one dominant underlying dimension?

A dimensionality analysis can indicate the degree to which the items in a test measure a dominant, measurable trait. Although the Rasch model requires unidimensionality, according to Smith (2004), "unidimensionality should not be viewed as a dichotomous yes or no decision, but rather as a continuum" (p. 576). Evidence that the test measures one underlying trait provides support for the structural aspect of validity. I investigated the degree to which unidimensionality is present in the BDI-2 standardization dataset. Because the publisher asserts that the BDI-2 subdomains measure one latent trait across the entire age continuum, I conducted the dimensionality analyses using the entire Gross Motor dataset and not separately for each age range.

I first examined the point-measure correlation statistic for each item in the Gross Motor Subdomain. The point-measure correlation is similar to the point-biserial correlation obtained from a CTT analysis. The point-biserial is a correlation between the score and the total raw test score. Higher correlations indicate that the examinees with the higher total scores received higher scores on the item, and the examinees with lower total scores received lower scores on the item. By contrast, for the Rasch-based point-measure correlation, the examinee's Rasch ability measure—rather than total raw score—is the continuous variable. Items with negative point-measure values don't correlate well with the Rasch ability measure.

Next, I examined the Rasch item mean-square outfit statistics for additional evidence of multidimensionality. Recall from Chapter II that the Rasch model allows us to predict how an examinee will perform on an item based on the examinee's ability, B_n, and the difficulty of the score category, D_{ik}. For every item, the model provides an expectation about what will happen when an examinee encounters an item. The expected value of the item mean-square outfit statistic is 1.0. Item mean-square outfit values close to 1.0 imply that the examiners, in general, assigned scores in a way that is consistent with the model expectations. In other words, the examiners tended to assign higher scores to the higher-ability examinees and lower scores to the lower-ability examinees. If an item's mean-square outfit statistic is higher than 1.0, the item may be measuring something other than the intended construct. For each BDI-2 Gross Motor item, I examined the mean-square item outfit statistic and looked for values greater than 1.5 (Smith, 1996). Values that are greater than 1.5 may indicate a departure from unidimensionality; in other words, those items may be measuring something different than the other items on the test.

Finally, I conducted a principal components analysis (PCA) of the residuals from the Rasch analysis to determine if there were additional factors in the data beyond the first factor extracted. Smith (2004) provided an overview of how WINSTEPS handles this analysis. If the Rasch model assumption of unidimensionality holds, the data should indicate the presence of only one primary factor, or latent trait. Using the residuals from the Rasch analysis (i.e., the difference between the predicted and actual values), I ran a PCA (WINSTEPS automates this process). The PCA can identify whether there are second or subsequent factors in the data that the first factor cannot explain, indicating a potential departure from unidimensionality. To help interpret the results from the PCA, I created five simulated datasets as comparisons. The simulated datasets provide a baseline value for the percent of variance that would be explained by chance alone when the data

perfectly fit the model. I created these simulated datasets based on the item and examinee estimates from the empirical BDI-2 Gross Motor data, and I calibrated the data in WINSTEPS. Following the same process that I used with the actual BDI-2 data, I ran a PCA on the simulated data. Finally, I compared the eigenvalues from the empirical and simulated datasets to determine if the percentage of the variance that the first principal component accounted for in the empirical data was similar to the percentage of variance accounted for in the first principal component of the simulated datasets, which would provide additional evidence that the BDI-2 Gross Motor subdomain items measure a single, dominant trait.

For each of these source of validity evidence—point-measure correlations, item mean-square outfit statistics, and the results from the PCA of the residuals—substantiation of unidimensionality provides evidence for the structural aspect of validity for the BDI-2.

Research Question 2.2) Do the data from the BDI-2 Gross Motor Subdomain satisfy the Rasch model requirement of local independence?

I investigated whether the BDI-2 Gross Motor data satisfied the requirement of local independence using the Fisher's *Z* index (Shen, 1997). I first obtained the Rasch standardized residuals for each observation of examinee *n* on item *i*:

$$d_{ni} = \frac{Observed_{ni} - Expected_{ni}}{SE_n}. \tag{3.1}$$

Using the ICORFILE command in WINSTEPS (Linacre, 2012), I produced a table of correlations of the residuals for each item pair. I then computed Fisher's *Z* indices to normalize the Pearson correlations obtained from the WINSTEPS table:

$$Z_{ij} = \frac{1}{2}\log\left(\frac{1+r_{ij}}{1-r_{ij}}\right). \tag{3.2}$$

BDI-2 Gross Motor items can be classified as either stand-alone or clustered. Stand-alone items measure a single skill or trait that no other item measures; clustered items measure a skill or trait that another item also measures. An example of a stand-alone item is Item 44 ("The child jumps rope without assistance"). This is the only BDI-2 Gross Motor item that measures skill at jumping rope. An example of an item cluster would be Items 18, 20, 25, 26, 29, and 34. All of these items measure a child's ability to move up and down stairs, starting with the basic skills of creeping or crawling, and progressing to walking up and down stairs without assistance. Following the methodology that Shen (1997) proposed, I used the level of dependence among the stand-alone items as a reference; by nature of the normalized distribution of correlations obtained using Fisher's *Z* indices, the mean of the correlations of stand-alone item residuals is zero. Then, I compared the Fisher's *Z* index for each pair of clustered items to the mean of the Fisher's *Z* indices for the stand-alone items. Values that are at least two standard deviations higher than zero in comparison to the Fisher's *Z* indices for stand-alone items may indicate a violation of the assumption of local independence. The absence of significantly dependent item pairs provides evidence that the requirement has been met, which in turn would provide evidence related to the structural aspect of validity.

Research Question 3) Does evidence relevant to the generalizability aspect of validity support the use of the BDI-2 Gross Motor subdomain scores to make inferences about children's development?

Research Question 3.1) Are the BDI-2 Gross Motor subdomain scores sufficiently reliable for making inferences about children's development?

I examined the Rasch examinee and item estimates of internal consistency at the subdomain level. High (i.e., 0.8 or higher) examinee and item estimates of internal consistency

provide evidence relevant to the generalizability aspect of validity supporting the propositions that subdomain scores are sufficiently reliable for making decisions about the performance of individual children and for interpreting the item hierarchy, respectively.

At each step in the study, when the results of my analyses revealed that changes to the data, score category descriptions, or item content would improve the model fit or increase the value of the BDI-2 score interpretations, I made the changes and documented the change in item and examinee fit, overall model fit, and interpretation. Because the BDI-2 is a published instrument, I do not anticipate that my suggested modifications will actually be implemented in the test; they serve only to inform possible improvements in future editions of the test.

IV. RESULTS

To investigate the research questions posed in Chapter III, I used the Partial Credit Model in WINSTEPS (Linacre, 2012) to analyze the examinee scores for all Gross Motor items. The results from that analysis are shown in Table XVII in the Appendix. The results from all subsequent analyses are also reported in Table XVII.

A. Research Question 1.1

Research Question 1.1) Do the examiners use the BDI-2 Gross Motor score categories of 2, 1, and 0 as expected to label behaviors that are fully emerged, emerging, and not yet emerged?

For this analysis, I used Linacre's (2004) guidelines for evaluating rating scale functioning as outlined in Chapter III. Before proceeding with investigation of the rating scale functioning, however, I first checked the item point-measure correlations to determine whether all items were positively correlated with the overall measure. The correlations were all positive and ranged from .28 to.79, with a mean of .68. All but three of the 45 items had point-measure correlations above .50, indicating that each Gross Motor item was positively correlated with the total measure.

Of the 2,500 examinees included in the BDI-2 Gross Motor dataset, 126 had extreme high or low total raw scores.[4] Extreme raw scores are not estimable within the Rasch model; they represent infinitely high or low measures on the underlying trait (Linacre, 2012). These examinee raw score strings do not contribute useful information to the calibration of the items or the examinees, so I removed these examinees manually prior to performing the subsequent analyses.

[4] Because the BDI-2 items are not all administered to every examinee, each examinee encounters a unique set of items. This necessarily means that the value of an extreme high score varies by examinee. For each examinee, an extreme high raw score would be a score of 2 for every item administered. The value of an extreme low score is always 0, or no credit for any administered item.

Additionally, I examined the summary statistics for item and examinee fit. The average mean-square examinee outfit value was .83 with a standard deviation of 1.60. The average mean-square item outfit value was 2.47, with a standard deviation of 3.31. The high value of the mean-square outfit and standard deviation indicates an unexpected amount of variability in the item and examinee outfit statistics. This result could be caused by a few very unexpected scores in the data file. Prior to examining the functioning of the rating scales, I wanted to ensure that the data file was clean and free from obvious data entry and examiner scoring errors. Recall that the mean-square outfit statistic is sensitive to unexpected scores on items that are either well above or below the examinee's ability; thus, unexpected scores of 2 on harder BDI-2 items for low-ability examinees or unexpected scores of 0 on easy BDI-2 items for high-ability examinees could affect the overall fit of a given item. An examination of the score patterns for the most misfitting examinees can reveal aberrant scores that might adversely impact the item fit. To determine whether a few unexpected scores of 0 or 2 might be contributing to the large standard deviations in the fit statistics, I examined the score patterns for the most misfitting examinees in the dataset. I was interested not only in identifying the examinees with unexpected scores, but also in determining how to best remove the problematic data: by removing individual scores, thereby creating missing data values within an examinee's score string, or by removing the entire score string for an examinee.

Of the 2,374 examinees with non-extreme scores, 284 had mean-square outfit statistics greater than 1.5; of these, 206 were 2.0 or greater, 137 were 3.0 or greater, and 40 were 9.9 or greater. Additionally, 11 items had mean-square outfit statistics of 1.5 or greater. From an output of every examinee's score on every item encountered, I examined the scores to determine whether the examinee misfit was caused by only one unexpected score, or by more than one. For each examinee-

by-interaction, WINSTEPS provides a standardized residual. This residual indicates how unlikely the observed score is, given the examinee's overall ability, and is calculated as

(Observed Score-Expected Score) / Square root (Variance).

This standardized residual approximates a *z* distribution; therefore, I considered values outside the range +/- 2.0 to indicate scores that were very unlikely. There were 866 scores associated with a standardized residual outside this range. Using the absolute value of this standardized residual, I ordered the unexpected scores for all examinees on all items. Next, I started with the highest absolute values of the standardized residual and located the unexpected scores in the data file. For the first several unexpected scores, the problems were apparent immediately. For example, the score string for items 1 through 32 for Examinee 2820 appeared as:

02222222222222222222222220020200.

Clearly, this is a higher-ability examinee as evidenced by the scores of 2 (full credit) on Items 2 through 25, Item 28, and Item 30. However, two other anomalies are apparent in this examinee's score string. First, the examiner did not follow the administration rules by utilizing the appropriate starting point for the child's age. This examinee was 1 year, 4 months old at the time of testing. According to the suggested starting points table in the administration manual, the examiner should have begun testing with Item 20, not with Item 1. Therefore, this examiner administered several items that were too easy for the child. Additionally, because this examinee scored 2 on Items 2 through 25, it is highly unlikely that he or she would have actually failed Item 1. More likely, the score of 0 on Item 1 was an examiner scoring error or a scanning error.[5]

[5] The publishing company inputted the test data from the BDI-2 standardization study into an electronic file through the use of a scannable test protocol. Examiners filled in "bubble" circles corresponding to the scores of 2, 1, and 0. Although the company implemented quality-control measures to maximize the accuracy of the data input process, a small chance of scanning errors existed due to erasures, stray marks, or smudges on the scannable test protocol documents. Thus, the data file might occasionally contain a score of 2, 1, or 0 for an item that was not administered by the examiner. The use of item and examinee fit statistics, including the standardized residuals for all scores, helps to identify and diagnose these possible scanning errors so that the scores can be removed from the dataset.

For each unexpected score, I first ascertained whether the score seemed like an examiner scoring or scanning error. Then I examined the item content and the examinee's age to determine whether an edit was warranted. In most cases the unexpected scores were scores of 1 or 0 that appeared somewhere in a string of 2s for examinees for whom the ability estimate was much higher than the item's difficulty estimate. Ultimately I identified 170 scores that appeared that they could have been examiner scoring or scanning errors. These scores came from 75 unique examinees, and accounted for fewer than 1% of all scores in the dataset. Table V reports demographic information for these 75 examinees and for the entire sample of 2,500 examinees from the standardization study. The subgroup of 75 examinees was similar to the entire standardization sample for all demographic characteristics except ethnicity. A higher percentage of the examinees with unexpected scores were Hispanic ($\chi^2 = 5.302$, 1 d.f., $p = 0.0213$).

TABLE V

DEMOGRAPHIC CHARACTERISTICS OF EXAMINEES WITH UNEXPECTED SCORES AND ENTIRE STANDARDIZATION SAMPLE

Demographic Variable	Percent in Subgroup of 75 Examinees with Unexpected Scores (n = 75)	Percent in Standardization Sample (n = 2,500)	χ^2 (d.f.)	p
Sex				
Male	49%	50%	0.006 (1)	.9388
Female	51%	50%		
Race				
White	61%	64%	2.31 (1)	.6304
Not White	39%	36%		
Ethnicity				
Not Hispanic	71%	81%	5.302 (1)	.0213
Hispanic	29%	19%		
Mother's Education				
Less than H.S.	18%	18%	3.733 (3)	.2917
H.S.	23%	31%		
Some College	28%	28%		
College +	31%	23%		
Father's Education				
Less than H.S.	7%	12%	5.263 (3)	.1535
H.S.	26%	34%		
Some College	33%	27%		
College +	34%	27%		

After removing these 170 scores, four additional examinees had scores that became extreme (maximum). I removed these four examinees and reran the data in WINSTEPS (Run 2 in Table XVII). The overall fit of the items improved (mean-square outfit value = .74), and the standard deviation of the mean-square outfit value decreased significantly (*SD* = .41). The overall examinee mean-square outfit value decreased to .71, as did the standard deviation of the mean-square outfit value (1.13). The change in -2LL was greater than the critical χ^2 value, and AIC and BIC both decreased from the first analysis. This indicates that the removal of a few

unexpected scores improved the overall fit of the data to the model. After this data cleaning process, I proceeded to examine the rating scale functioning using Linacre's (2004) guidelines.

Guideline 1: Ensure that each score category of each item contains a minimum of 10 observations. Table VI reports score frequencies for each category of each item in the BDI-2 Gross Motor subdomain. Items 8, 9 and 16 had fewer than 10 scores for category 1. Item 9 also had fewer than 10 scores for category 0. Category frequencies less than 10 may result in less stable estimation of the category thresholds, and they might indicate that a different—possibly collapsed—scoring system might be more appropriate for those items, or that more data should be collected in an attempt to obtain scores in the infrequently used categories. In the case of the BDI-2 standardization dataset, though, each item in the dataset had several hundred scored responses; thus, low category usage is not likely due to a shortage of data.

TABLE VI

FREQUENCY OF EACH SCORE CATEGORY FOR BDI-2 GROSS MOTOR ITEMS, RUN 2

	Total # of	Score of 2		Score of 1		Score of 0	
Item	Scores	N	%	N	%	N	%
Item 1	247	202	82%	35	14%	10	4%
Item 2	252	185	73%	40	16%	27	11%
Item 3	255	189	74%	26	10%	40	16%
Item 4	259	174	67%	28	11%	57	22%
Item 5	255	156	61%	25	10%	74	29%
Item 6	459	361	79%	20	4%	78	17%
Item 7	440	326	74%	34	8%	80	18%
Item 8	317	283	89%	**8**	3%	26	8%
Item 9	461	447	97%	**7**	2%	**7**	2%
Item 10	413	322	78%	21	5%	70	17%
Item 11	390	264	68%	29	7%	97	25%
Item 12	374	253	68%	27	7%	94	25%
Item 13	354	234	66%	16	5%	104	29%
Item 14	312	191	61%	25	8%	96	31%
Item 15	290	155	53%	20	7%	115	40%
Item 16	255	151	59%	**7**	3%	97	38%
Item 17	221	130	59%	22	10%	69	31%
Item 18	424	334	79%	11	3%	79	19%
Item 19	397	281	71%	20	5%	96	24%
Item 20	384	294	77%	10	3%	80	21%
Item 21	383	260	68%	16	4%	107	28%
Item 22	570	468	82%	24	4%	78	14%
Item 23	740	607	82%	53	7%	80	11%
Item 24	726	555	76%	63	9%	108	15%
Item 25	710	532	75%	46	6%	132	19%
Item 26	671	503	75%	46	7%	122	18%
Item 27	585	408	70%	41	7%	136	23%
Item 28	539	364	68%	50	9%	125	23%
Item 29	503	260	52%	49	10%	194	39%
Item 30	596	265	44%	92	15%	239	40%
Item 31	624	399	64%	57	9%	168	27%
Item 32	807	602	75%	59	7%	146	18%
Item 33	754	509	68%	73	10%	172	23%
Item 34	742	491	66%	71	10%	180	24%
Item 35	716	471	66%	98	14%	147	21%
Item 36	1,167	992	85%	58	5%	117	10%
Item 37	698	356	51%	104	15%	238	34%

TABLE VI
FREQUENCY OF EACH SCORE CATEGORY FOR BDI-2 GROSS MOTOR ITEMS, RUN 2 (CONTINUED)

	Total # of	Score of 2		Score of 1		Score of 0	
Item	Scores	N	%	N	%	N	%
Item 38	1,127	809	72%	112	10%	206	18%
Item 39	1,232	864	70%	74	6%	294	24%
Item 40	1,091	632	58%	199	18%	260	24%
Item 41	977	545	56%	169	17%	263	27%
Item 42	898	398	44%	109	12%	391	44%
Item 43	660	306	46%	173	26%	181	27%
Item 44	624	175	28%	122	20%	327	52%
Item 45	519	63	12%	105	20%	351	68%

Because the BDI-2 scales are designed to measure effectively a wide range of examinee ability (i.e., from low-functioning infants through high-functioning 8-year-olds), I expected some items to be extremely easy for most of the examinees who encountered them and other items to be extremely hard for most of the examinees who encountered them. For this reason, I expected to find that the easiest items had very low frequencies for score category 0, and that the hardest items had very low frequencies for score category 2. I did notice this pattern to be true for the easiest items, but not necessarily for the hardest items; this might indicate that the test has a ceiling for the highest-ability 7- and 8-year-old children. More surprising to me were the extremely low frequencies across most items for score category 1. Examiners very seldom assigned a score of 1 to examinees; frequencies for score category 1 ranged from 7 for Items 9 and 16 (2% of the scores for each of those items) to 199 (18% of the scores) for Item 40. These relatively low score counts for score category 1 suggest that examiners did not assign the score of 1 for items to indicate that a child showed an emerging skill or ability; rather, most examiners seemed to be utilizing a two-category scoring system, assigning scores of 2 (skill fully emerged) or 0 (skill not yet emerging) for most items.

Guideline 2: Analyze the observation distribution within each item. As indicated by Table VI and as noted above, the frequency of scores in category 1 across all items was relatively low compared to the frequency of scores in categories 2 and 0. Linacre (2004) suggested that although uniform category usage is optimal for calibration, a non-uniform distribution with peaks at the extreme categories might be substantively meaningful. For instance, the bimodal distribution of scores in these data might suggest that the score category description for the category 1 may not be sufficiently different from the descriptions for the categories 0- or 2 for many items. Most of the items in the BDI-2 Gross Motor domain require the examinee to

perform some task, and the examiner assigns a score based on the child's performance. As a result, the score category descriptions for most of the Gross Motor items involve measurements of distance, time, or number of successful trials. It may be that the scores of 1 cover a very narrow band on the ability continuum; for example, if a child can successfully perform a skill one time (thus meriting a score of 1), perhaps very little additional skill or development is necessary to perform the skill successfully multiple times (for a score of 2). As another example, many of the items require the examiner to measure the distance that a child can successfully walk, hop, or run. In this case, if a child can walk, hop, or run for a few feet (for a score of 1), then perhaps it is just as likely that the child will be able to successfully walk, hop, or run a longer distance (for a score of 2).

Guideline 3: Analyze average measures within each category of each item. The "average measure" of a category is the mean ability, in logits, of the examinees who scored in a given category on an item. Average measures are sample-dependent. If higher score categories imply more of a measured skill, and if the average measures of the score categories are monotonically increasing, then there is evidence that that particular set of data supports the theoretical ordering of the scoring categories (Linacre, 2004). Table VII presents average measures, standard errors, and mean-square outfit statistics for each item in the BDI-2 Gross Motor subtest. All items had monotonically increasing average measures, indicating that for the examinees in the BDI-2

TABLE VII

AVERAGE MEASURES, STANDARD ERRORS, AND MEAN-SQUARE OUTFIT STATISTICS FOR SCORES OF 0, 1, AND 2 FOR BDI-2 GROSS MOTOR ITEMS, RUN 2

	Score of 0			Score of 1			Score of 2		
Item	Avg. Meas.	S.E.	Outfit MNSQ	Avg. Meas.	S.E.	Outfit MNSQ	Avg. Meas.	S.E.	Outfit MNSQ
1	-17.08	0.34	1.88	-16.70	0.25	0.51	-11.82	0.26	0.96
2	-17.34	0.22	1.06	-15.99	0.22	1.00	-11.25	0.27	0.86
3	-17.23	0.15	0.65	-15.70	0.25	0.48	-11.23	0.26	0.84
4	-16.95	0.14	0.44	-14.98	0.19	0.28	-10.79	0.27	1.04
5	-15.84	0.15	1.31	-14.36	0.34	1.79	-9.98	0.28	1.13
6	-15.42	0.12	0.32	-13.65	0.28	0.12	-7.12	0.20	0.55
7	-14.56	0.15	1.36	-12.51	0.32	0.73	-6.62	0.20	0.65
8	-15.81	0.31	0.73	-12.71	0.33	0.04	-6.11	0.22	**2.76**
9	-12.53	0.68	**2.52**	-10.99	0.87	0.95	-3.02	0.23	**2.42**
10	-13.98	0.16	0.49	-11.76	0.31	0.46	-6.48	0.20	0.65
11	-12.17	0.18	**2.52**	-10.21	0.25	0.39	-5.57	0.21	**2.12**
12	-11.86	0.16	0.56	-9.54	0.18	0.13	-5.35	0.21	0.55
13	-10.86	0.15	1.30	-9.32	0.20	0.12	-5.03	0.21	0.33
14	-9.66	0.14	0.59	-8.01	0.31	1.35	-4.44	0.23	0.87
15	-8.72	0.10	0.61	-7.66	0.21	0.25	-3.56	0.24	0.77
16	-7.91	0.10	0.96	-6.46	0.42	0.17	-3.00	0.21	0.42
17	-7.20	0.15	0.27	-4.55	0.16	0.27	-1.81	0.28	1.35
18	-6.30	0.18	0.53	-4.01	0.26	0.08	1.38	0.19	0.76
19	-4.81	0.16	0.43	-3.64	0.16	0.10	2.36	0.18	0.39
20	-4.43	0.17	1.84	-2.86	0.27	0.10	2.16	0.18	0.90
21	-3.43	0.12	**2.54**	-1.21	0.25	0.19	3.12	0.16	0.34
22	-2.73	0.13	0.63	-1.03	0.21	0.11	4.39	0.11	0.54
23	-0.61	0.12	0.80	2.18	0.20	0.55	6.23	0.10	0.63
24	0.36	0.14	0.76	3.03	0.17	0.48	6.53	0.10	1.17
25	1.20	0.14	0.72	3.23	0.17	0.30	6.76	0.09	0.58
26	2.19	0.12	1.27	3.68	0.17	0.29	6.90	0.09	0.56
27	3.82	0.08	0.52	5.30	0.21	1.23	7.48	0.09	0.86
28	4.33	0.09	0.48	5.62	0.15	0.46	7.90	0.09	0.75
29	5.68	0.08	1.60	6.82	0.15	0.89	8.75	0.10	1.00
30	4.50	0.10	**3.72**	6.31	0.17	1.82	8.03	0.11	1.40
31	6.47	0.08	0.71	8.17	0.16	0.84	10.50	0.09	0.89
32	6.91	0.08	0.49	8.64	0.17	0.51	11.69	0.08	0.62
33	7.92	0.09	0.65	9.50	0.12	0.35	12.22	0.08	0.63
34	8.54	0.10	1.63	9.65	0.15	0.66	12.32	0.08	0.79

TABLE VII
AVERAGE MEASURES, STANDARD ERRORS, AND MEAN-SQUARE OUTFIT STATISTICS FOR SCORES OF 0, 1, AND 2 FOR BDI-2 GROSS MOTOR ITEMS, RUN 2 (CONTINUED)

	Score of 0			Score of 1			Score of 2		
Item	Avg. Meas.	S.E.	Outfit MNSQ	Avg. Meas.	S.E.	Outfit MNSQ	Avg. Meas.	S.E.	Outfit MNSQ
35	8.78	0.10	1.21	10.08	0.11	0.49	12.64	0.08	0.69
36	10.05	0.09	1.05	11.17	0.16	0.51	14.39	0.06	1.52
37	10.01	0.09	**2.27**	11.51	0.13	1.29	13.16	0.09	1.47
38	10.87	0.07	0.72	12.09	0.09	0.23	15.07	0.06	0.81
39	10.26	0.07	0.47	11.79	0.11	0.16	14.91	0.06	0.79
40	11.46	0.08	1.14	13.70	0.10	1.09	15.38	0.06	1.59
41	12.33	0.06	0.62	14.13	0.09	0.74	15.76	0.05	1.00
42	13.47	0.07	1.47	14.88	0.12	1.21	16.13	0.05	1.28
43	14.17	0.09	0.80	15.35	0.08	0.96	16.38	0.05	0.98
44	14.84	0.06	0.90	16.05	0.08	1.07	16.71	0.06	1.04
45	15.56	0.05	0.88	16.69	0.09	0.90	16.80	0.10	1.49

standardization study, each successive score category of every item represented more gross motor skill or ability than the next lower category.

Guideline 4: Examine the mean-square outfit statistic for each score category of each item. Mean-square outfit statistics greater than 2.0 indicate that there is more unsystematic than systematic variance in the data, while statistics less than 1.0 indicate little systematic variance in the data. Outfit statistics are particularly sensitive to unexpected scores that are off target (i.e., much lower or much higher than an examinee's ability), and high outfit, or misfit, is often easy to remedy by removing these anomalous scores. Using Linacre's (2004) suggested criterion of outfit value > 2.0 to indicate score category misfit, several of the score categories for items shown in Table VII exhibited misfit (score category outfit statistics greater than 2.0 are indicated by boldface type in the table). For Items 9, 11, 21, 30, and 37, score category 0 was misfitting; for Items 8, 9, and 11, score category 2 was misfitting. In these instances, misfit is likely caused by a few unexpected off-target scores due to either unusual examinee skill profiles or errors in data entry or examiner administration. I reexamined the standardized residual statistics for all the scores for all examinees and identified 19 additional scores with residual statistics outside the range +/-2.0, indicating that the scores were very unexpected given the examinees' overall ability measures. In Run 3, I removed these unexpected scores from the analysis. The change in -2LL was greater than the critical χ^2 value, and both AIC and BIC decreased, indicating better overall data fit to the model. The fit of score category 0 for Items 9, 11, and 21 improved significantly. The fit of score category 2 for Items 8 and 11 improved significantly, and the fit of Items 30 and 37 improved moderately. Score category 2 for Item 9 still exhibited significant misfit, even after the removal of the additional unexpected scores. Table VIII displays the mean-square outfit

statistics for the six items with category misfit before and after the removal of these unexpected scores in Run 3.

TABLE VIII

FIT OF SCORE CATEGORIES BEFORE AND AFTER REMOVAL OF ADDITIONAL UNEXPECTED EXAMINEE SCORES FOR ITEMS 8, 9, 11, 21, 30, AND 37, RUN 3

Item	Score Category	MNSQ Outfit Statistic Prior to Removal	MNSQ Outfit Statistic After Removal
8	2	2.76	0.81
9	0	2.52	0.52
9	2	2.42	2.36
11	0	2.52	1.47
11	2	2.12	0.96
21	0	2.54	0.30
30	0	3.72	1.95
37	0	2.27	1.54

In addition to the misfitting score categories, several of the score categories in Table VII exhibited overfit as indicated by mean-square outfit statistics less than 1.0. For these items, examinee score strings were overly predictable or Guttman-like; while these items do not degrade the quality of measurement, they do not add much information to the examinee ability measures, and they may also contribute to the artificial expansion of the scale's overall range and artificially increase estimates of internal consistency.

Guideline 5: Ensure that category thresholds advance for each item. At every point along the ability continuum, higher-ability examinees should have a greater probability of scoring 2 on any item than lower-ability examinees. Likewise, each score category (2, 1, and 0) should be modal at some point over the range of ability measured by the test, indicating that the category is

most likely to be observed at some point along the scale. The "step calibrations" are the thresholds along the ability continuum at which it becomes more likely that an examiner will assign an examinee that score category than the next lower score category. Another way to describe these thresholds is as the points along the ability continuum where there is equal probability that an examiner will assign an examinee either of two adjacent score categories. If a BDI-2 Gross Motor item exhibits disordered step calibrations, this means that the examiners will never be most likely to assign a score of 1 to an examinee. In this case, the probability curve for the item will show a flattened curve for score category 1, with its peak falling below the intersection of the probability curves for the score categories of 2 and 0. We can examine the step calibrations and the probability curves for each item to determine if category thresholds advance monotonically. Table IX displays the item measures, item standard errors, step calibrations for the category transitions from 0 to 1 and from 1 to 2, and score category standard errors for each BDI-2 Gross Motor item.

An examination of Table IX reveals that all but nine of the BDI-2 items have disordered step calibrations (disordered steps are indicated by boldface type in the table). The items with increasing step thresholds were Items 1, 2, 7, 23, 24, 40, 41, 43, and 45. The other 36 items all showed disordering of the step thresholds. Figures 2 and 3 show examples of the score category probability curves for two BDI-2 Gross Motor items. Figure 2 (Item 1) contains an item with properly ordered category thresholds. Note from Table IX that for Item 1, the score of 1 is most probable for examinees with ability measures between -19.16 logits and -17.34 logits. In contrast, Figure 3 (Item 18) shows how, when category thresholds are disordered, the score of 1 is never most probable for any examinee.

TABLE IX
ITEM DIFFICULTY MEASURES, STANDARD ERRORS, CATEGORY TRANSITION MEASURES, AND CATEGORY STANDARD ERRORS FOR THE BDI-2 GROSS MOTOR ITEMS, RUN 3

Item #	Item Difficulty Measure	Item S.E.	Transition from 0 to 1		Transition from 1 to 2		Logit Distance between Category Transition Measures
			Measure	S.E.	Measure	S.E.	
1	-18.25	0.18	-19.16	0.38	-17.34	0.23	1.82
2	-17.12	0.15	-17.74	0.27	-16.5	0.22	1.24
3	-16.77	0.14	-16.69	0.26	-16.84	0.23	**-0.15**
4	-16.05	0.14	-15.96	0.24	-16.13	0.23	**-0.17**
5	-14.92	0.14	-14.71	0.24	-15.13	0.24	**-0.42**
6	-14.23	0.15	-13.78	0.25	-14.68	0.25	**-0.90**
7	-13.11	0.14	-13.31	0.24	-12.92	0.23	0.39
8	-13.77	0.26	-13.58	0.49	-13.95	0.4	**-0.37**
9	-13.11	0.3	-12.61	0.63	-13.62	0.44	**-1.01**
10	-12.61	0.14	-12.21	0.26	-13.02	0.24	**-0.81**
11	-10.5	0.13	-10.2	0.23	-10.8	0.21	**-0.60**
12	-10.11	0.13	-9.71	0.23	-10.51	0.21	**-0.80**
13	-9.34	0.12	-8.21	0.22	-10.46	0.21	**-2.25**
14	-8.41	0.11	-7.73	0.2	-9.08	0.2	**-1.35**
15	-7.38	0.12	-6.57	0.21	-8.19	0.23	**-1.62**
16	-6.39	0.14	-4.68	0.27	-8.1	0.27	**-3.42**
17	-5.34	0.14	-4.92	0.25	-5.76	0.23	**-0.84**
18	-4.63	0.13	-3.28	0.24	-5.98	0.23	**-2.70**
19	-3.15	0.13	-2.46	0.22	-3.85	0.23	**-1.39**
20	-3.27	0.13	-1.81	0.23	-4.73	0.24	**-2.92**
21	-1.61	0.14	-0.68	0.24	-2.54	0.24	**-1.86**
22	-1.09	0.14	-0.69	0.24	-1.48	0.23	**-0.79**
23	1.36	0.12	1.14	0.21	1.58	0.17	0.44
24	2.47	0.1	2.38	0.19	2.56	0.15	0.18
25	3.22	0.09	3.72	0.17	2.72	0.15	**-1.00**
26	3.68	0.09	4.28	0.17	3.07	0.15	**-1.21**
27	5.05	0.08	5.98	0.15	4.13	0.14	**-1.85**
28	5.62	0.08	6.28	0.15	4.97	0.14	**-1.31**
29	7.16	0.08	7.94	0.14	6.38	0.14	**-1.56**
30	6.36	0.08	6.47	0.13	6.26	0.13	**-0.21**
31	8.09	0.08	8.63	0.14	7.55	0.14	**-1.08**
32	8.47	0.08	8.98	0.15	7.96	0.14	**-1.02**
33	9.63	0.08	10.03	0.14	9.24	0.13	**-0.79**
34	10	0.07	10.49	0.14	9.5	0.12	**-0.99**
35	10.21	0.07	10.31	0.13	10.11	0.12	**-0.20**
36	10.9	0.07	11.67	0.14	10.12	0.12	**-1.55**
37	11.57	0.07	11.79	0.12	11.35	0.11	**-0.44**

TABLE IX
ITEM DIFFICULTY MEASURES, STANDARD ERRORS, CATEGORY TRANSITION MEASURES, AND CATEGORY STANDARD ERRORS FOR THE BDI-2 GROSS MOTOR ITEMS, RUN 3 (CONTINUED)

Item #	Item Difficulty Measure	Item S.E.	Transition from 0 to 1		Transition from 1 to 2		Logit Distance between Category Transition Measures
			Measure	S.E.	Measure	S.E.	
38	12.35	0.07	12.61	0.12	12.08	0.11	**-0.53**
39	12.12	0.06	12.91	0.12	11.33	0.11	**-1.58**
40	13.38	0.06	13.03	0.11	13.73	0.09	0.70
41	14.14	0.06	14.07	0.11	14.21	0.09	0.14
42	15.36	0.06	15.97	0.1	14.76	0.1	**-1.21**
43	15.57	0.06	15.35	0.11	15.79	0.1	0.44
44	16.69	0.06	16.92	0.1	16.46	0.11	**-0.46**
45	17.74	0.07	17.72	0.11	17.77	0.15	0.05

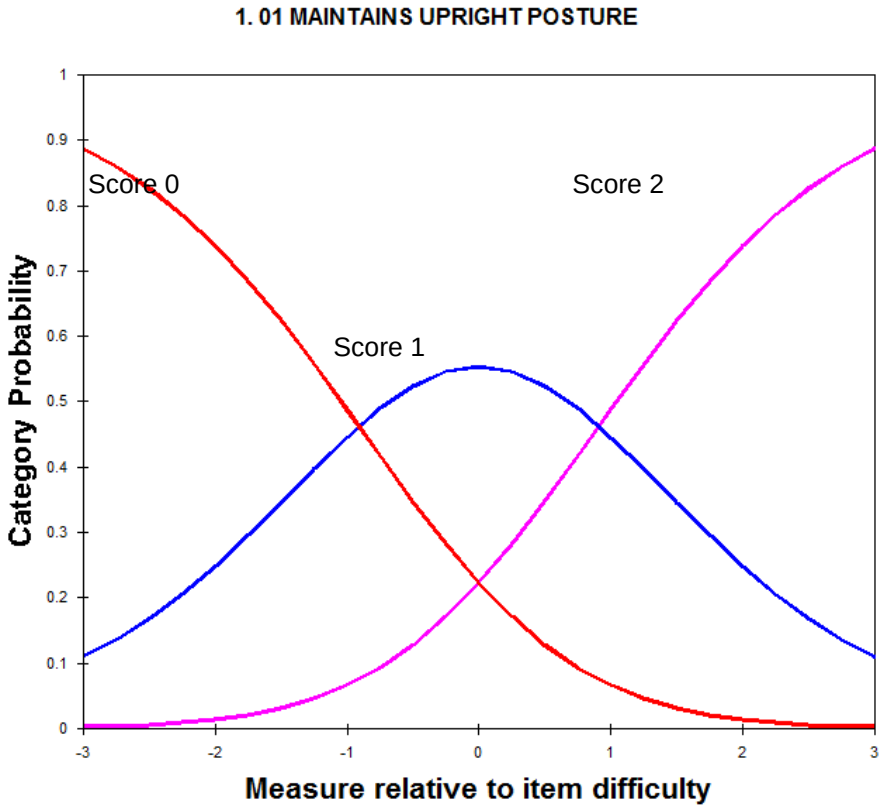

Figure 2. Score category probability curves for Item 1 ("Maintains upright posture").

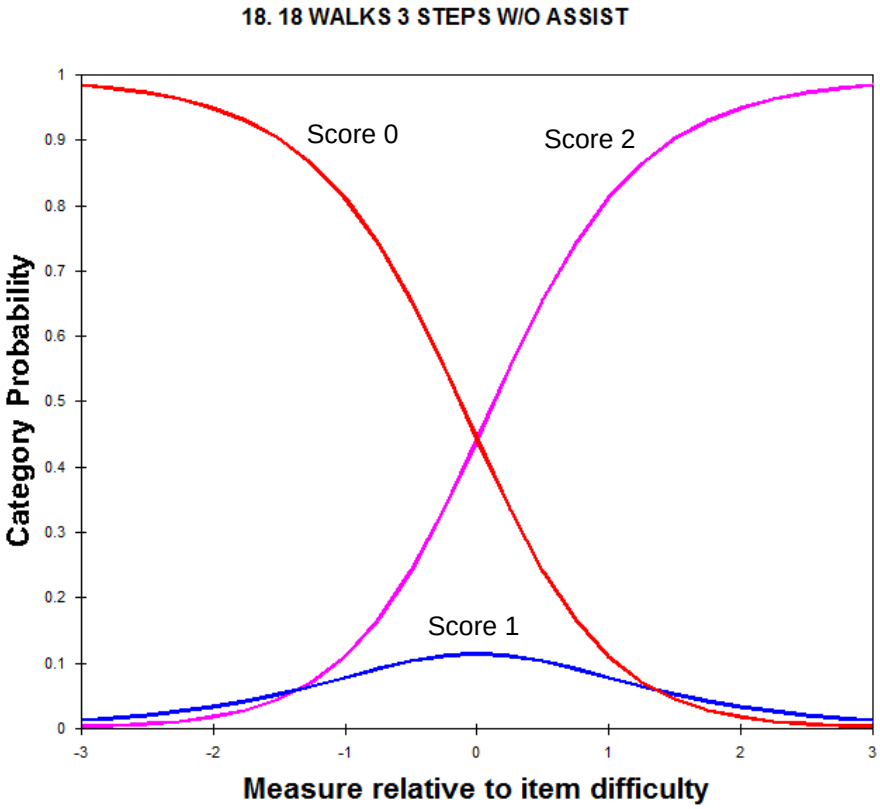

Figure 3. Score category probability curves for Item 18 ("Walks three steps without assistance").

Of the nine items that had ordered thresholds, the distance between the calibrations of the adjacent thresholds was smaller than the width of the standard error band around the threshold measures for Items 7, 24, 41, and 45. In other words, the threshold measures for these items were sufficiently similar—and the width of the error band around the threshold calibrations was sufficiently large—so as to prevent one from being certain that the thresholds were indeed ordered correctly. The extent to which category threshold disordering occurs for these items is further evidence that examiners did not use the score category of 1 as intended (i.e., to indicate

that a child displays an emerging skill); rather, examiners were more likely to assign a score of 2 (fully emerged skill) or 0 (skill not emerging).

Some of the previous findings (i.e., the relatively low usage for score category 1 and the disordering of category thresholds) suggest that many—or all—of the BDI-2 Gross Motor items may be more appropriately scored with a two-category scoring system. To determine whether this is the case, I first needed to devise a rationale for combining score categories. Figure 4 is a graphical representation of the observed category averages for each item in Run 3; the data points on the lines for 0, 1, and 2 represent the locations on the logit scale of the average ability of the examinees who received that score from examiners.

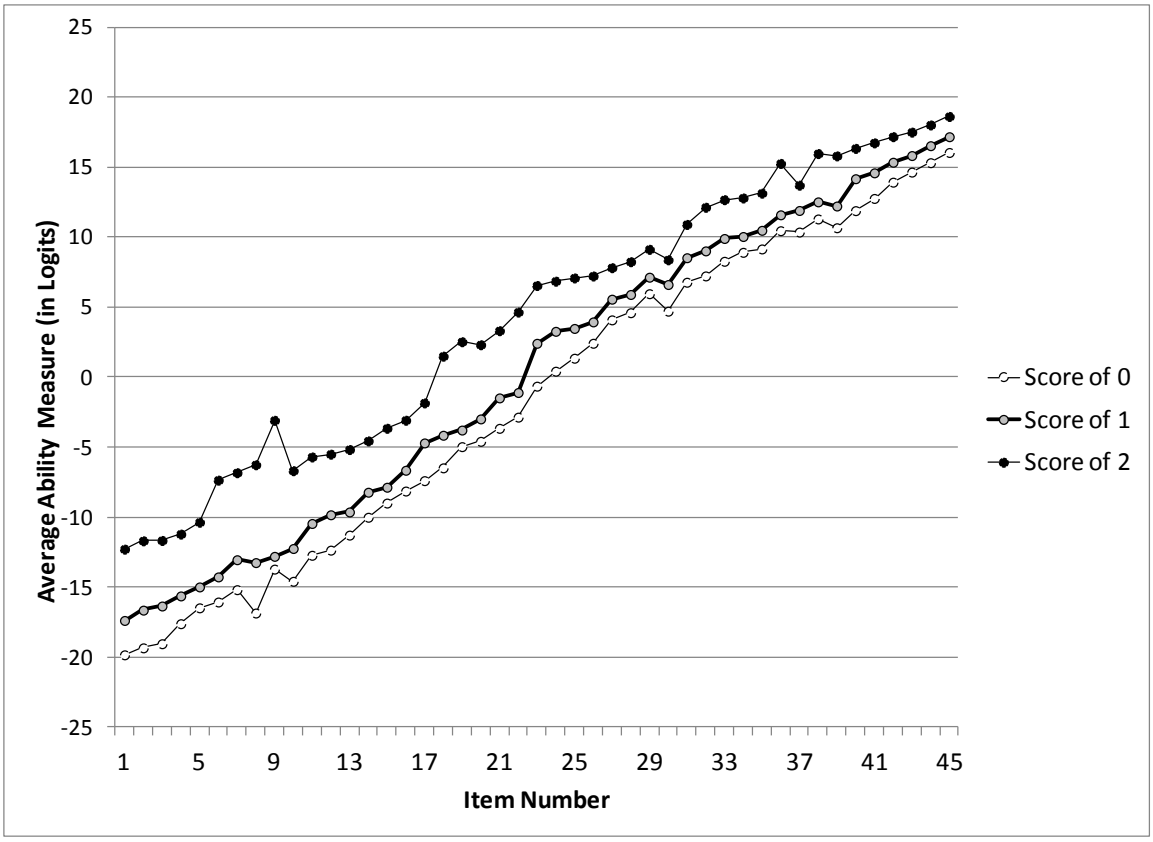

Figure 4. Graphical display of average ability measures for examinees in each score category of each item, Run 3.

As we can tell by examining the relative distances between the 0s, 1s, and 2s in Figure 4, when examiners did assign a score of 1, it tended to be for lower-ability examinees. This suggests that most items might be more interpretable if examiners used a two-category scoring system, whereby an examinee would receive full credit on an item *only* if the skill that the item measures had fully emerged. Although Table IX showed that five items did have threshold calibrations that were properly ordered outside the range of measurement error, it seemed to make both practical and pragmatic sense to collapse the 0 and 1 score categories for all 45 items. This solution also makes sense in the context of developmental assessment. Skills that have fully emerged are more reliably measured, because a child will more consistently display a fully emerged skill than one that is just beginning to emerge. As an example, a child who is just learning to walk will often teeter about for just a few steps before toppling over. At this point, the child's walking skill is not consistent or reliable; the child may successfully walk many steps on one try but have trouble taking more than one or two steps on the next try. However, once the child has mastered walking, the measurement of his or her walking skill will be more reliable because the child will be more stable on his or her feet, taking more consistent steps and falling less frequently.

Additionally, from a practical standpoint combining categories 0 and 1 would result in a much simpler, easier-to-interpret approach to scoring. This also lends support for collapsing all BDI-2 categories, so as to keep the scoring consistent from item to item. Many items have complex conditions or compound requirements to assign a score of 1, making the score category descriptions quite lengthy in some instances. This is even the case for some of the items with score categories that appear to be ordered correctly. For example, the category descriptions for the score category or 1 and score category of 0 for Item 41 ("Walks a 6-foot line on the floor, heel-to-toe, with eyes open") reference multiple-conditions:

> Child walks forward, heel-to-toe in any trial but steps off line up to 3 times in 6 feet *or* puts hands out momentarily for balance. (1 point)
>
> Child walks forward, heel-to-toe in any trial less than 6 feet *or* steps off line 4 or more times in 6 feet *or* puts hands out more than once *or* leaves hands out for balance *or* sways excessively. (0 points)

If the score category of 1 were eliminated, the description for the score category of 2 ("Child walks forward, heel-to-toe, at least 6 feet in any trial, keeping feet on line and not swaying to maintain balance") would remain the same, but the description for the score of 0 would effectively become "anything *less* than the requirements for score of 2." Several BDI-2 items have complex category descriptions; collapsing the categories of 0 and 1 would reduce both the examiner's cognitive load and the time required to score the items.

Finally, some BDI-2 items currently have score category descriptions that require examiners to distinguish with extreme precision between scores of 2, 1, or 0. For instance, Item 40 ("Stands on each foot alternately with eyes closed") requires that the examiner effectively distinguish between responses in which a child successfully completes the task for 3 or more seconds (score of 2), 1 or 2 seconds (score of 1), or less than 1 second (score of 0). Removing the score category of 1 for these types of items would not require the examiner to distinguish among such short time intervals; if this item were scored using two categories, a child who sustains a response for 3 or more seconds would receive full credit for the item while a child who does not sustain the response for at least 3 second would receive no credit.

To investigate the functioning of a two-category scoring system, I rescored all BDI-2 Gross Motor items by combining the score categories of 1 and 0 for all items, such that the combined score of 0/1 would indicate, in general, that a skill was "not fully emerged" while a score of 2 would indicate that a skill was "fully emerged." (This run is labeled as Run 4 in Table XVII.) After rescoring, scores of 1 or 0 became 0s, and scores of 2 became 1s. Category

frequencies increased when I combined the score categories of 1 and 0, whereby most items now had at least 50 scores in each category. (The exceptions were Items 1, 8, and 9, which contained 45, 34, and 11 observations, respectively.) With the combining of categories 0 and 1, an additional 26 examinees had extreme scores of 0. I dropped these examinees from the analysis. Overall item fit increased toward the expected value of 1, as evidenced by the average item mean-square outfit statistic (mean = 0.93, compared with 0.74 in the first analysis). Although the standard deviation of the overall item fit increased (0.80 compared with 0.39 in the prior analysis), the change in -2LL was greater than the critical χ^2 value and both AIC and BIC decreased, indicating overall improvement in model fit with the collapsed 0/0/1 scoring.

Table X presents measures, standard errors, and mean-square outfit statistics for all items before and after rescoring. Boldface type indicates misfit. From Table X we can see that the items maintained their relative order of difficulty after recoding (r_s = 0.9998). Item pairs 2-3 and 8-9 were reversed in difficulty order; however, in both analyses, the original and new item difficulty measures were nearly identical, so these rank-order shifts would have a negligible impact on examinee measurement. Items 2, 5, 7, 13, 17, 24, and 30 had mean-square outfit statistics greater than 1.5. Item 5 appeared to be very misfitting (mean-square outfit value = 4.4). When I examined item-by-examinee standardized residuals, I noticed that each of these items had at least one score with a very large (> 15.00) residual statistic, indicating that, after rescoring, at least one examinee's score became very unexpected for each of these items. I removed six additional unexpected examinee scores from the data (see Run 5 in Table XVII). After this edit, the overall item mean-square outfit statistic decreased (mean = 0.74; S.D. = 0.49); this result was expected

TABLE X

DIFFICULTY MEASURES AND MEAN-SQUARE OUTFIT STATISTICS FOR SCORE CATEGORIES 0 AND 1 FOR THE BDI-2 GROSS MOTOR ITEMS AFTER RESCORING, RUN 4

	0/1/2 Scoring (Before Rescoring)			0/0/1 Scoring (After Rescoring)		
Item	Difficulty Measure	S.E.	Item MNSQ Outfit	Difficulty Measure	S.E.	Item MNSQ Outfit
1	-18.25	0.18	0.68	-28.58	0.28	0.58
2	-17.12	0.15	1.07	-27.21	0.24	**2.03**
3	-16.77	0.14	0.55	-27.27	0.24	0.74
4	-16.05	0.14	0.44	-26.2	0.24	0.39
5	-14.92	0.14	1.79	-24.57	0.25	**4.40**
6	-14.23	0.15	0.22	-23.66	0.25	0.23
7	-13.11	0.14	0.91	-21.61	0.24	**1.66**
8	-13.77	0.26	0.11	-22.79	0.42	0.12
9	-13.11	0.30	0.08	-22.11	0.46	0.08
10	-12.61	0.14	0.58	-21.31	0.24	0.63
11	-10.50	0.13	0.61	-17.97	0.23	0.48
12	-10.11	0.13	0.32	-17.39	0.23	0.25
13	-9.34	0.12	0.55	-16.44	0.22	0.46
14	-8.41	0.11	1.07	-14.59	0.21	**1.68**
15	-7.38	0.12	0.52	-12.86	0.23	0.47
16	-6.39	0.14	0.52	-11.47	0.26	0.63
17	-5.34	0.14	0.53	-9.17	0.25	**2.49**
18	-4.63	0.13	0.28	-8.36	0.24	0.26
19	-3.15	0.13	0.21	-5.27	0.26	0.19
20	-3.27	0.13	0.68	-5.80	0.25	0.77
21	-1.61	0.14	0.16	-2.47	0.26	0.11
22	-1.09	0.14	0.25	-1.13	0.26	0.10
23	1.36	0.12	0.60	3.54	0.20	0.61
24	2.47	0.10	0.63	5.24	0.17	**2.13**
25	3.22	0.09	0.44	6.00	0.16	0.30
26	3.68	0.09	0.62	6.65	0.15	0.50
27	5.05	0.08	0.98	8.60	0.14	1.06
28	5.62	0.08	0.55	9.61	0.14	0.64
29	7.16	0.08	1.25	12.02	0.14	1.46
30	6.36	0.08	1.59	11.14	0.14	**2.11**
31	8.09	0.08	0.85	13.60	0.14	1.06
32	8.47	0.08	0.55	14.24	0.14	0.56
33	9.63	0.08	0.52	16.16	0.13	0.59
34	10.00	0.07	1.02	16.66	0.13	1.08
35	10.21	0.07	0.76	17.24	0.12	0.61
36	10.90	0.07	0.78	17.87	0.12	1.43
37	11.57	0.07	1.36	19.20	0.12	1.47
38	12.35	0.07	0.47	20.42	0.11	0.51
39	12.12	0.06	0.38	19.80	0.11	0.34
40	13.38	0.06	1.24	22.38	0.10	1.48

TABLE X
DIFFICULTY MEASURES AND MEAN-SQUARE OUTFIT STATISTICS FOR SCORE CATEGORIES 0 AND 1 FOR THE BDI-2 GROSS MOTOR ITEMS AFTER RESCORING, RUN 4 (CONTINUED)

	012 Scoring (Before Rescoring)			001 Scoring (After Rescoring)		
Item	Difficulty Measure	S.E.	Item MNSQ Outfit	Difficulty Measure	S.E.	Item MNSQ Outfit
41	14.14	0.06	0.78	23.19	0.10	0.70
42	15.36	0.06	1.35	24.52	0.09	1.04
43	15.57	0.06	0.91	25.23	0.10	0.83
44	16.69	0.06	1.00	26.62	0.11	1.39

because I had removed the very "noisy" scores from the data. The decrease in the standard deviation of the overall item outfit statistic after the removal of these six scores suggests that the relatively high standard deviation in the prior run was likely due to just a few extreme outlier scores. The change in -2LL was greater than the critical value of χ^2, and the values of AIC and BIC decreased. Individual item fit also improved.

Table XI reports measures, standard errors, overall mean-square outfit statistics, and category mean-square outfit statistics for all 45 BDI-2 Gross Motor items after rescoring and data cleaning. After removing the six unexpected scores, the only item with a mean-square outfit value greater than 2.0 was Item 30 (outfit value = 2.16). Score category 0 exhibited misfit for Items 14 and 30. Score category 1 exhibited misfit for Items 5, 9, and 16. I examined the detailed item text and score category descriptions for these items to gain insight into why the score categories for these items were not fitting the model. For Items 14 and 30, the high misfit for score category 0 suggests that some higher-ability examinees received unexpectedly low scores on the items. Item 14 ("Child makes stepping movements when held in an upright position") is appropriate for infants and requires the examiner to hold the child upright under the arms and score the item according to the child's foot movements. If the child does not move his or her feet, the item is scored as 0. Unlike many of the other BDI-2 Gross Motor items, Item 14 does not have an Observation or Interview option for administration; therefore, if the examiner does not observe the desired response during the test session, the item must be scored as 0. I believe that allowing the examiner to use Observation and/or Interview procedures to assess the child on this item would increase the likelihood of a higher-ability child receiving a score of 2,, especially if he or she were tired or uncooperative during the testing session and failed to respond when the examiner used the Structured procedure.

TABLE XI

DIFFICULTY MEASURES AND MEAN-SQUARE OUTFIT STATISTICS FOR ITEMS AND SCORE CATEGORIES OF 0 AND 1 FOR THE BDI-2 GROSS MOTOR ITEMS AFTER RESCORING, RUN 5

Item	Difficulty Measure	S.E.	Item MNSQ Outfit	MNSQ Outfit Score of 0	MNSQ Outfit Score of 1
1	-29.48	0.28	0.59	0.68	1.10
2	-28.15	0.24	0.43	0.43	0.75
3	-28.16	0.24	0.81	0.98	0.93
4	-27.07	0.24	0.40	0.25	0.93
5	-25.46	0.25	1.29	0.57	**2.78**
6	-24.44	0.26	0.24	0.17	0.79
7	-22.33	0.25	0.48	0.76	0.39
8	-23.48	0.43	0.12	0.10	0.69
9	-22.74	0.47	0.10	0.05	**3.54**
10	-21.94	0.25	0.75	0.87	1.31
11	-18.41	0.24	0.50	0.51	0.68
12	-17.80	0.23	0.25	0.21	0.41
13	-16.81	0.22	0.49	0.71	0.26
14	-14.91	0.21	1.97	**3.09**	0.76
15	-13.13	0.24	0.47	0.34	0.65
16	-11.66	0.27	1.38	0.79	**2.35**
17	-9.20	0.25	0.38	0.58	0.24
18	-8.44	0.24	0.26	0.35	0.56
19	-5.33	0.26	0.19	0.11	0.48
20	-5.86	0.25	0.80	1.25	0.92
21	-2.49	0.27	0.11	0.09	0.17
22	-1.07	0.27	0.11	0.07	0.34
23	3.88	0.21	0.63	0.80	0.26
24	5.65	0.17	0.70	0.72	0.75
25	6.39	0.16	0.30	0.28	0.39
26	7.04	0.15	0.50	0.53	0.43
27	9.00	0.14	1.08	1.18	0.88
28	10.01	0.14	0.65	0.55	0.87
29	12.43	0.14	1.50	1.65	1.35
30	11.55	0.14	**2.16**	**2.91**	1.58
31	14.03	0.14	1.09	1.32	0.70
32	14.68	0.14	0.57	0.58	0.55
33	16.62	0.13	0.60	0.43	0.95
34	17.12	0.13	1.09	1.27	0.72
35	17.70	0.12	0.61	0.66	0.51
36	18.35	0.12	0.95	0.86	1.46
37	19.67	0.12	1.48	1.57	1.39
38	20.89	0.11	0.52	0.34	0.96
39	20.27	0.11	0.34	0.22	0.62
40	22.85	0.10	1.48	1.27	1.78

TABLE XI

DIFFICULTY MEAN-SQUARE OUTFIT STATISTICS FOR ITEMS AND SCORE CATEGORIES OF 0 AND 1 FOR THE BDI-2 GROSS MOTOR ITEMS AFTER RESCORING, RUN 5 (CONTINUED)

Item	Difficulty Measure	S.E.	Item MNSQ Outfit	MNSQ Outfit Score of 0	MNSQ Outfit Score of 1
41	23.66	0.10	0.70	0.63	0.79
42	24.99	0.09	1.04	1.17	0.94
43	25.7	0.10	0.83	0.72	0.93
44	27.09	0.11	1.39	0.99	1.55
45	28.78	0.15	1.15	1.05	1.16

Item 30 (“Child walks backward 5 feet”) is appropriate for older children and requires that the child walk backward, without support and with coordination, five or more feet for full credit. In the BDI-2 standardization, 15% of the children scored 1 on this item (the rubric stipulated awarding a score of 1 for “Child walks backward with coordination and balance, without support fewer than 5 feet”). The difference between a score of 1 and a score of 2 on the item was that a score of 2 required an examinee to walk backward at least 5 feet. If the examinee could walk backward some distance, but could not continue for at least 5 feet, the examiner would assign a score of 1. In the original calibration, the step thresholds for this item had virtually identical logit measures, indicating that it was not much harder to score a 2 on this item than a 1. The data file contained scores for two examinees who received a 1 on this item but who, given their overall ability measures, would have been expected to receive a score of 2. With the rescoring of the item to 0/0/1, these two high-ability examinees (1138 and 1166) had failing scores on the item. I suspected that removing these two unexpected examinee scores would improve the overall item fit and the fit of the categories.

Items 5, 9, and 16 exhibited misfit for the score category of 1, meaning that examiners unexpectedly assigned some lower-ability examinees scores of 2. When I reviewed the scoring rubric for Item 5 (“Child brings hands together at midline”), I found it to be confusing. This item is appropriate for a young infant. For a score of 2, the child must put his or her hands together during the time when the examiner is observing the child or must have done this at another time when a parent or caregiver was watching. There is no Structured procedure for administering this item. For a score of 1, the child must *attempt* to put his or her hands together at the midline. (The frequency with which the child attempts this action distinguishes a score of 1 from a score of 0, but the distinction between the two categories is not, in my opinion, very clear). The score

category descriptions for this item require that the examiner, parent, or caregiver be able to discern the child's *intention.* I would argue that it would be difficult for even a trained expert to determine whether an infant's arm movements are deliberate attempts to bring the hands together or are simply random movements. For this reason, I believe that examiners should have the option of using a Structured procedure for this item, with very specific category descriptions for scores of 1 or 0. For a score of 1, the examiner would need to observe the child performing this skill during the testing session or during observation.

Item 9 ("Child puts objects into his or her mouth") is appropriate for older infants and young toddlers. This item exhibited misfit for score category 1, indicating that some examiners awarded lower-ability examinees full credit on the item. Examinees can only use an Interview procedure when administering this item. I believe that allowing examiners to use a Structured and/or Observation procedure to administer this item would permit the examiner to validate the parent or caregiver's report about the child's ability to perform this task. Such validation may reduce the number of lower-ability examinees who receive full credit for the item.

Item 16 ("Child pulls himself or herself to standing position while holding onto a solid object without adult assistance") also had a mean-square outfit statistic greater than 2.0 for score category 1. This item is appropriate for a toddler-age child, and examiners can use a Structured, Observation, or Interview procedure to administer this item. When I examined the item-by-examinee residual statistics for this item, I found one very low-ability examinee (Examinee 444) who had received full credit on this item, even though this examinee's 0 scores on other items indicated that the skills of sitting (Item 13) and crawling (Item 15) had not yet emerged. I believe that this particular item score is either a scanning error or an examiner administration error and that removing it is warranted, given the evidence about this examinee's other gross motor skills.

Based on the detailed reviews previously described, I made three final adjustments to the dataset, removing the score for Item 30 for Examinees 1138 and 1136 and the score for Item 16 for Examinee 444 (see Run 6 in Table XVII for the results with these scores omitted). For Item 30, the overall mean-square outfit statistic improved from 2.16 in the prior analysis to 1.75. The fit of score category 0 improved from 2.91 in the prior analysis to 2.02. The mean-square outfit statistic for Item 16 improved very slightly (1.36 compared to 1.38 in the prior analysis), and the fit of score category 1 improved very slightly as well (2.30 compared to 2.35 in the prior analysis). Additionally, the change in -2LL was greater than the critical value of χ^2, and both AIC and BIC decreased, suggesting that the overall model fit was slightly better after removal of these three scores.

During the investigation of Research Question 1.1, I made several edits to the BDI-2 Gross Motor dataset and proposed a change to the scoring system to maximize data fit to the Rasch model. To determine how these proposed changes impacted the examinee ability measures, I plotted the examinee ability measures from Run 2 against those from Run 6, as shown in Figure 5. From Figure 5, I made the following observations:

1. Although the examinees were not identically rank-ordered in Run 2 and Run 6, all but four examinees had score differences that were within the standard error of measurement. In fact, the Spearman rank-order correlation for the two sets of ability measures was 0.997.

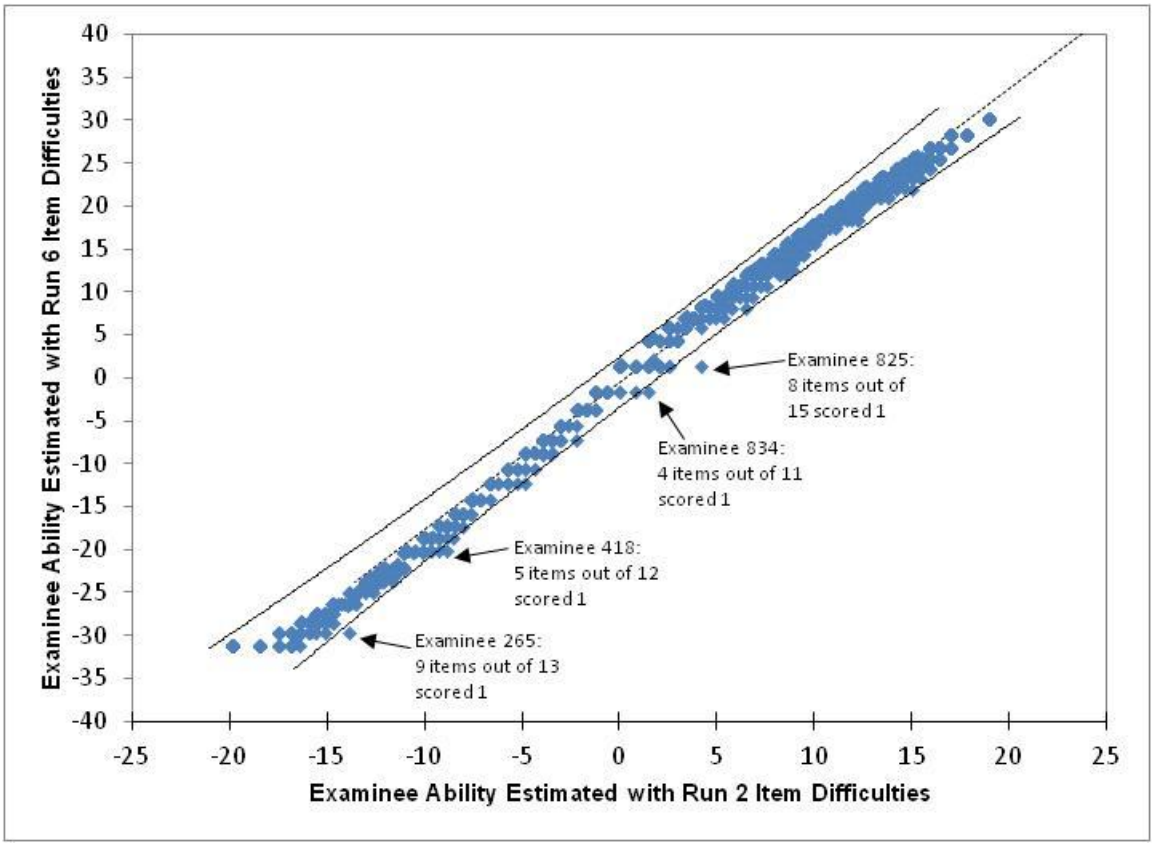

Figure 5. Ability measures for all examinees using Run 2 item difficulty measures plotted against ability measures for all examinees using Run 6 item difficulty measures.

2. The item difficulty measures from Run 6 tended to produce lower ability measures for examinees at the bottom end of the ability range; in other words, the two-category scoring system resulted in a slightly nonlinear relationship between measures from Run 2 and Run 6. Collapsing the 1 and 0 score categories penalizes the lowest-ability examinees.

The examinees for whom the ability measures differed most from Run 2 to Run 6 (i.e., the outliers appearing outside the standard error bands noted on the plot) were those for whom examiners assigned a score of 1 most frequently. For these examinees the rescoring had a greater impact on their relative standing in the standardization group of examinees.

B. Research Question 1.2

Research Question 1.2) In the standardization dataset, are there any anomalous examinee score strings that may have degraded the quality of the item calibrations?

An examinee mean-square outfit value greater than 1.5 indicate that the examiner assigned one or more item scores that were highly unexpected, given the examinee's ability. Prior to beginning any data analysis in this chapter, I first cleaned the dataset by identifying and removing 170 scores that I believed were the result of scanning errors or examiner scoring errors. After these removals in Run 2, the overall examinee fit to the model improved (mean-square examinee outfit value = 0.71); however, the variability in the examinee mean-square outfit statistics as indicated by a large standard deviation (1.13) suggested that the dataset still contained some highly misfitting examinees. Indeed, there were 284 examinees with mean-square outfit values greater than 1.5.

After I collapsed the score categories of 0 and 1 and removed some additional unexpected scores in Runs 3 through 6, the overall examinee mean-square outfit value for Run 6 decreased to 0.69 with a standard deviation of 1.36. This finding suggests that the collapsing of the scoring categories (and the subsequent removal of highly misfitting scores) had the effect of producing more overfit in the data, but the large standard deviation indicates that some examinee scores became very unexpected after the collapsing of categories. In fact, there were still 258 examinees with mean-square outfit values greater than 1.5. From an examination of the individual examinee scores, I identified 724 scores (out of the total of 25,241 individual scores associated with all examinees in the dataset) with standardized residuals outside the range $z = \pm 2.0$. When I removed these scores from the dataset and attempted to rerun the item analysis, the data did not contain enough connectivity to calibrate the items together. Instead, the items split into two

separate sets, Items 1 through 22 and Items 23 to 45. This disconnection is a result of the administration procedures for the BDI-2, because examiners do not administer every item to every examinee; rather, examiners follow suggested starting points to administer only those items that are reasonably targeted to an examinee's ability. When I removed scores from the already sparse item-examinee data matrix, the overlap between examinees and items became even more sparse, preventing the calibration of the items onto one common scale. Because Item 23 was a starting point for many examinees, very few examinees who had scores for Item 23 also had scores for Items 21 and 22.

To deal with this problem, I ran several successive analyses, removing fewer unexpected scores in each run. Starting with a tolerance level of $z = 2.0$, I gradually increased the tolerance for unexpected scores by $z = \pm 0.5$. The analysis failed to produce a connected item set calibration until I reached a tolerance of $z = \pm 9.0$. At this tolerance level of ±9.0, I flagged the 26 most unexpected scores in the dataset. When I removed these scores, the overall mean-square examinee outfit value decreased to 0.63, with a standard deviation of 1.14 (labeled as Run 7 in Table XVII). The highly unexpected scores that still remained in the dataset were likely the cause of the relatively high standard deviation; however, the connectivity issues that I encountered prevented me from investigating how the removal of those items might impact the standard deviation of the outfit statistic. The change in -2LL was greater than the critical value of χ^2, and both AIC and BIC decreased, suggesting that the overall model fit was slightly better after removal of the 26 most unexpected scores.

To determine if the removal of the unexpected examinee scores in Runs 2 through 7 impacted the individual item calibrations, I compared the item difficulty measures from Run 2 to those from Run 7. Figure 6 shows the relationship between the item difficulty measures from

Run 2 and those from Run 7. Datasets that exhibit more Guttman-like patterns of scores produce larger spread in examinee and ability measures. Therefore, the larger range of item difficulties in Run 7 was expected because excessive "noise" was removed from the measures. Despite the differences in minimum and maximum item difficulty measures from the two runs, there was a nearly linear relationship between the item difficulty measures from Run 2 and Run 7 (Pearson correlation = .999; Spearman rank-order correlation = .992). With the exception of very minor item rank order changes[6] for Items 2 and Item 3 and for Item 8 and Item 9, the original BDI-2 item hierarchy was preserved, even after the removal of significantly misfitting scores and the collapsing of score categories 0 and 1. The item difficulty measures from Run 7 were slightly higher for Items 22 through 26 and slightly lower for Items 42 to 45.

C. Research Question 2.1

Research Question 2.1) Do the data from the BDI-2 Gross Motor Subdomain represent one dominant underlying domain?

If the items on a test measure one dominant underlying trait, then this provides structural validity evidence for inferences drawn from the test scores. If, however, the items appear to be measuring several different traits, then inferences made from the test scores may be called into question. To examine the dimensionality of the BDI-2 Gross Motor items, I used three different approaches. First, I examined the point-measure correlations for all items. Next, I evaluated the item mean-square outfit statistics for values greater than 1.5. Finally, I conducted a principal components analysis of the Rasch residuals and compared the results to those of a PCA from a simulated dataset in order to determine if a significant second factor existed in the data.

[6] Items 2 and 3 had difficulty values that differed by less than one-half of a logit in both Run 2 and Run 7. Because their difficulty values were nearly identical, this slight shift in rank order likely had minimal impact on the examinee ability measures derived from these item difficulty calibrations. The same explanation applies to Items 8 and 9, because their difficulty values were also nearly identical in both runs.

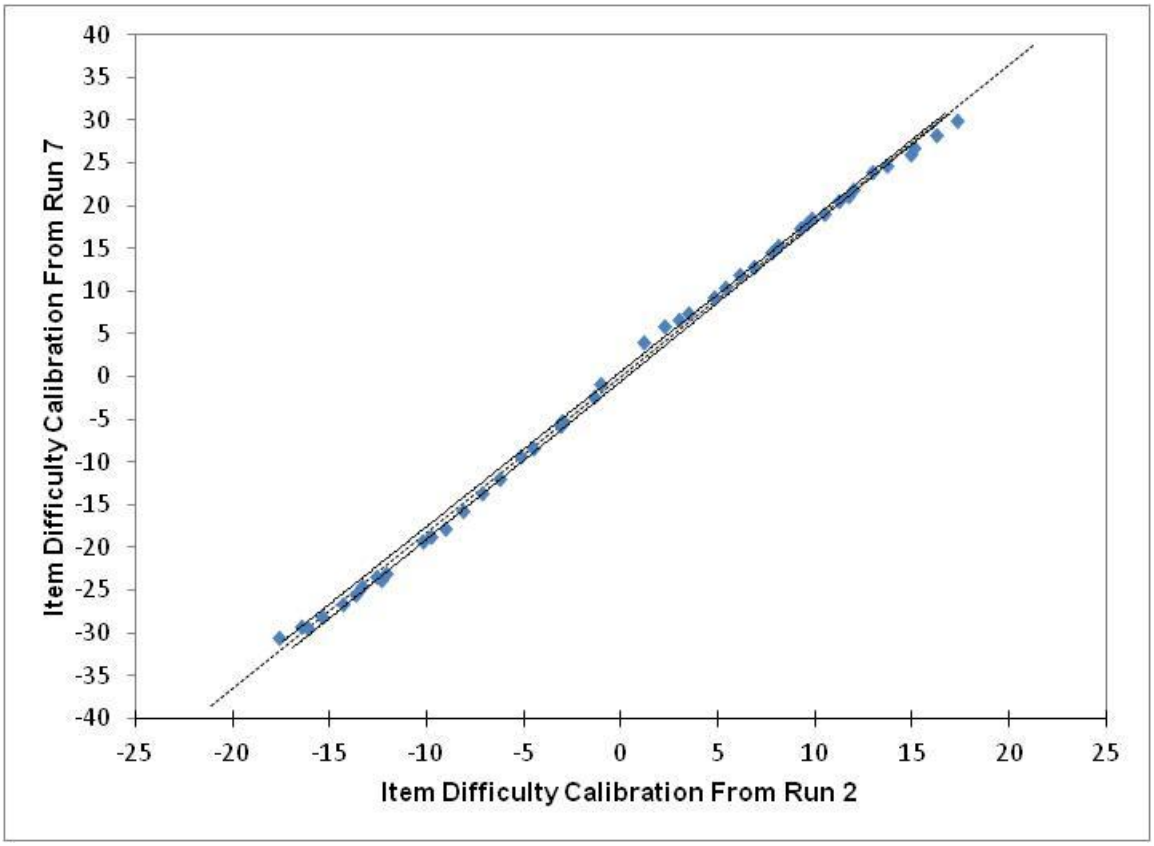

Figure 6. Item difficulty calibrations from Run 2 and Run 7.

Because the validity evidence that I gathered in this study will be used to support or refute the interpretation of scores on the published BDI-2, I used the standardization dataset (with the 170 unexpected scores that reflected scanning or scoring errors removed; i.e., Run 2) for this investigation. However, I was also interested in whether the assumption of unidimensionality held for the two-category scoring system that I proposed in the prior section; therefore I also investigated the item point-measure correlations, item outfit statistics, and the results from the PCA of the residuals for the dataset from Run 7, in which I collapsed the score categories.

Point-measure correlations are correlations between an examinee's score on an item (either a 0 or a 2) and the examinee's overall ability measure (a continuous variable). Point-

measure correlations indicate how well the examinee's performance on an item correlates with his or her overall test performance. A high point-measure correlation indicates that the item generally appears to be measuring the same underlying trait as the other items; a low or negative point-measure correlation indicates that the item may be measuring a different construct from what the other the items on the test measure. Table XII displays the point-measure correlations for all BDI-2 Gross Motor items from Run 2.

All BDI-2 Gross Motor subdomain item point-measure correlations using the three-category (Run 2) and two-category (Run 7) scoring systems were positive and were in the moderate to high range. Item 9 ("Puts object into his or her mouth") had the lowest correlations. This is not surprising, as Item 9 was originally part of the BDI-2 Perceptual Motor Subdomain; the publisher moved the item to the Gross Motor Subdomain after the standardization and prior to final publication of the test. The other item point-measure correlations ranged from 0.47 for Item 1 ("Maintains upright posture at adult's shoulder without assistance for at least 2 minutes") and Item 45 ("The child throws a ball and hits a target with the nondominant hand") in the original, three-category scoring system to 0.81 for Item 21 ("Walks without support for 10 feet without falling") in the two-category scoring system, with most correlations falling in the 0.60 to 0.80 range.

A review of the item mean-square outfit statistics can identify items that are functioning in an unexpected manner. If an item has a high mean-square outfit statistic, this could indicate that the item is measuring a different ability from the ability that the other items are measuring. Table XIII displays item difficulty measures and mean-square outfit statistics for all items in Runs 2 and 7. High mean-square outfit statistics are shown in bold.

TABLE XII
BDI-2 GROSS MOTOR ITEM POINT-MEASURE CORRELATION COEFFICIENTS

	Point-Measure Correlation	
BDI-2 Item	Run 2	Run 7
1	.47	.57
2	.58	.63
3	.59	.63
4	.66	.62
5	.68	.65
6	.68	.66
7	.70	.69
8	.62	.61
9	.29	.30
10	.66	.65
11	.70	.70
12	.71	.70
13	.70	.70
14	.68	.69
15	.73	.74
16	.75	.76
17	.69	.69
18	.68	.70
19	.76	.78
20	.69	.71
21	.80	.81
22	.74	.78
23	.71	.72
24	.73	.73
25	.75	.74
26	.70	.70
27	.68	.67
28	.70	.69
29	.71	.71
30	.70	.69
31	.74	.75
32	.71	.73
33	.75	.75
34	.70	.71
35	.71	.71
36	.59	.60
37	.68	.67
38	.76	.75
39	.79	.78
40	.75	.71

TABLE XII
BDI-2 GROSS MOTOR ITEM POINT-MEASURE CORRELATION COEFFICIENTS (CONTINUED)

	Point-Measure Correlation	
BDI-2 Item	Run 2	Run 7
41	.79	.75
42	.73	.71
43	.67	.68
44	.62	.65
45	.47	.65

TABLE XIII
DIFFICULTY MEASURES AND MEAN-SQUARE OUTFIT STATISTICS FROM THREE CATEGORY SCORING (RUN 2) AND TWO-CATEGORY SCORING (RUN 7)

	Difficulty Measure		Mean-Square Outfit	
Item	Run 2	Run 7	Run 2	Run 7
1	-17.61	-30.65	0.71	0.60
2	-16.48	-29.32	0.95	0.44
3	-16.13	-29.38	0.56	0.33
4	-15.42	-28.22	0.45	0.40
5	-14.31	-26.58	**1.51**	1.35
6	-13.64	-25.52	0.19	0.25
7	-12.59	-23.35	0.77	0.51
8	-13.38	-24.54	0.21	0.13
9	-12.35	-23.77	0.79	0.10
10	-12.12	-23.02	0.47	0.47
11	-10.18	-19.34	1.10	0.52
12	-9.79	-18.70	0.30	0.27
13	-9.07	-17.72	0.50	0.24
14	-8.17	-15.72	1.01	0.71
15	-7.18	-13.75	0.49	0.55
16	-6.22	-12.05	0.52	0.14
17	-5.19	-9.27	0.51	0.40
18	-4.49	-8.46	0.27	0.28
19	-3.02	-5.26	0.21	0.20
20	-3.14	-5.80	0.64	0.84
21	-1.43	-2.38	0.69	0.11
22	-1.03	-0.94	0.22	0.11
23	1.19	4.10	0.59	0.64
24	2.24	5.86	0.61	0.48
25	2.98	6.63	0.44	0.31
26	3.43	7.30	0.61	0.52
27	4.79	9.28	0.88	0.81
28	5.35	10.32	0.53	0.52
29	6.86	12.83	1.18	1.07
30	6.12	11.91	**2.25**	**1.78**
31	7.76	14.56	0.81	0.94
32	8.12	15.28	0.53	0.52
33	9.25	17.38	0.50	0.49
34	9.60	17.85	1.00	0.77
35	9.81	18.48	0.75	0.64
36	10.49	19.11	0.77	0.57
37	11.19	20.48	**1.65**	1.26
38	11.92	21.80	0.46	0.48

TABLE XIII
DIFFICULTY MEASURES AND MEAN-SQUARE OUTFIT STATISTICS FROM THREE CATEGORY SCORING (RUN 2) AND TWO-CATEGORY SCORING (RUN 7) (CONTINUED)

	Difficulty Measure		Mean-Square Outfit	
Item	Run 2	Run 7	Run 2	Run 7
39	11.7	21.15	0.37	0.34
40	12.94	23.87	1.22	1.39
41	13.69	24.69	0.77	0.72
42	14.9	26.04	1.33	1.09
43	15.10	26.76	0.90	0.80
44	16.22	28.17	1.00	1.16
45	17.28	29.88	1.14	1.01

In Run 2, three items exhibited significant misfit to the Rasch model: Items 5, 30, and 37. During my investigation of Research Question 1.1 I identified several unexpected scores that contributed to these items' misfit. Additionally, my investigation of the function of the rating scales revealed that the BDI-2 standardization data did not appear to support the use of the three-category scoring structure. After I removed several unexpected scores and collapsed the score categories 0 and 1 in Runs 2 through 7, only one item (Item 30) still had a mean-square outfit statistic greater than 1.5. For this item, I hypothesized that the misfit was more likely due to issues with unclear and/or non-discrete score category descriptions than due to multidimensionality. In the prior section I discussed some possible approaches to changing the rubric for Item 30 in a future revision of the BDI so as to clarify the scoring for examiners. The overall pattern of fit for the items in Run 7 indicates that the BDI-2 Gross Motor items conform to the Rasch model expectations, suggesting a unidimensional measure of gross motor development; however, Linacre (2012) suggested that outfit statistics are influenced by "accidents" in the data and generally cannot detect additional dimensions in the data as well as other methods such as principal components analysis (PCA).

A PCA of Rasch residuals can help to identify how many factors a test measures. Using the datasets from Run 2 and Run 7, I ran PCAs using WINSTEPS. In a Rasch PCA of residuals, the first factor extracted from the data is often termed the "Rasch dimension" or "Rasch factor" (Linacre, 2012) and is analogous to the first principal component. Any variance that the contrasts explain can potentially be interpreted as additional dimensions in the data, depending on the size of the contrasts. In Run 2 the Rasch factor accounted for 72.7% of the variance in the data, leaving 27.3% of variance unexplained. Of this unexplained variance, the first contrast accounted for 6.2% of the variance, or 2.8 of the total of 45 units of unexplained variance. This first factor

accounted for a relatively small percentage of the unexplained variance, just above the absolute eigenvalue size of 2.0 that Linacre (2012) recommended for significance. Items 5, 6, 8, 19, 20, 23, 24, and 28 had positive loadings on the contrast, while Items 1, 12, 21, 22, 27, 28, and 30 had negative loadings. Figure 7 displays a plot of the standardized residuals for the first contrast in Run 2.

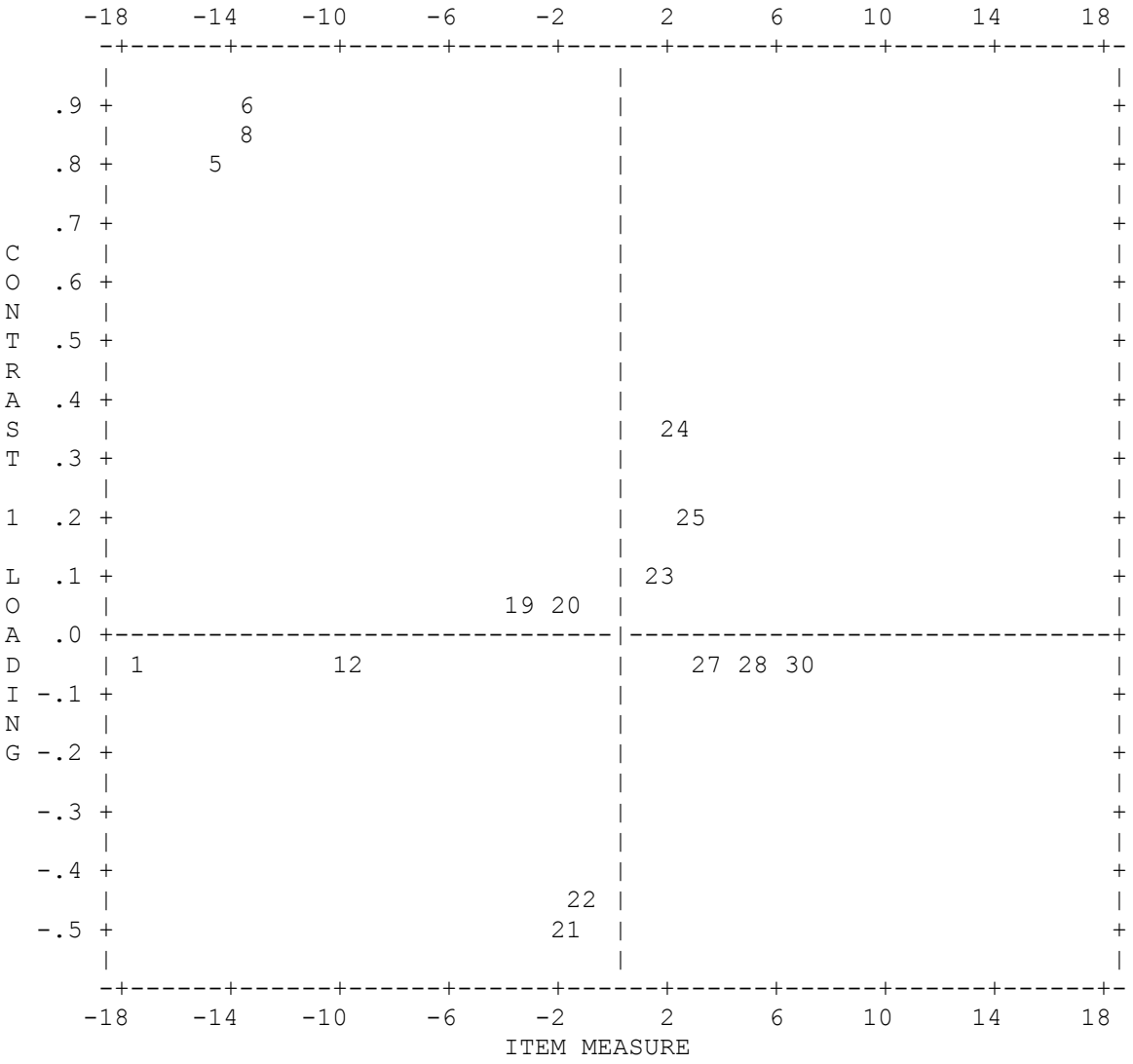

Figure 7. First contrast plot of standardized residuals, Run 2.

This contrast does not appear to show any content-related patterns. Both the positive and negative loadings for the first contrast contain items that are related to head movements, hand movements, and whole body movements and are spread across the entire range of difficulty.

In Run 7, the first principal component (i.e., the "Rasch" factor; Linacre, 2012) explained 70.6% of the variance in the data, leaving 29.4% of the variance unexplained. Of this unexplained variance, the first contrast accounted for 4.3% of the variance, or 2.0 of 45 units of unexplained variance. This amount represents an even smaller percentage of the total unexplained variance than the amount that the first contrast explained in Run 2. Items 1, 2, 8, 9, 10, 14, 25, and 26 had positive loadings on the contrast, and Items 3, 4, 5, 6, 7, 11, and 13 had negative loadings. Figure 8 displays a plot of the standardized residuals for the first contrast in Run 7.

Although the amount of variance that the first contrast accounts for by the first contrast in both Run 2 and Run 7 appears to be small, interpretation of the PCA results depends upon the choice of the critical value for the eigenvalue of the first factor (Smith, 2004). To determine whether the percentage of variance that the first PCA factor accounted for is meaningful in each run, I compared the values to the baseline values generated from five simulated datasets from Runs 2 and 7. Table XIV reports the results of the PCA analysis of residuals for Runs 2 and 7 for the empirical and simulated datasets.

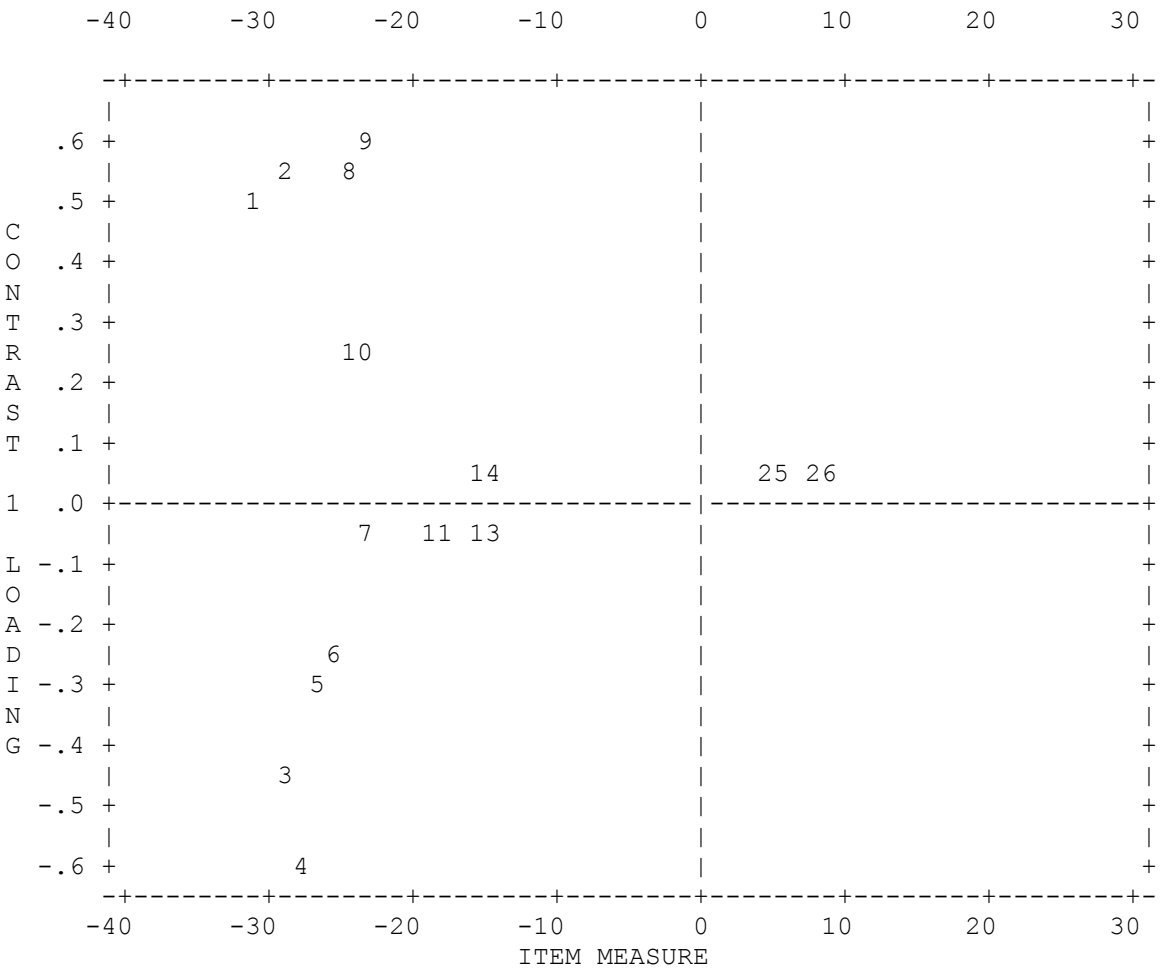

Figure 8. First contrast plot of standardized residuals, Run 7.

TABLE XIV

RESULTS OF PCA ANALYSIS OF RESIDUALS FOR RUN 2 AND RUN 7

	Run 2		Run 7	
Dataset	Eigenvalue	Percentage of unexplained[a] variance accounted for	Eigenvalue	Percentage of unexplained[a] variance accounted for
Empirical	2.8	6.2	2.0	4.3
Simulation 1	2.0	4.4	1.6	3.5
Simulation 2	2.0	4.4	1.5	3.4
Simulation 3	1.9	4.3	1.6	3.5
Simulation 4	1.9	4.3	1.5	3.4
Simulation 5	1.7	3.9	1.8	4.0

[a]The percentage of variance explained by the observations (i.e., the "Rasch factor"; see Linacre, 2012) in the empirical data was approximately equal to that explained by the observations in the simulated datasets, in both Run 2 and Run 7.

The percentage of variance accounted for in the first contrast of residuals from these simulated datasets represents what we would expect through chance alone. The difference between the percentage of residual variance that the first contrasts accounted for in the empirical data and the simulated datasets is relatively small for both Runs 2 and 7, suggesting that the BDI-2 Gross Motor items measure a unidimensional construct in both the two-category and three-category scoring systems.

D. Research Question 2.2

Research Question 2.2) Do the data from the BDI-2 Gross Motor Subdomain satisfy the Rasch model requirement of local independence?

To test the assumption of local independence under both the three-category and two-category scoring systems, I compared the standardized residuals for all item pairs from Runs 2 and 7. There were 990 unique item pairs in the BDI-2 Gross Motor subdomain. Using the ICORFILE command in WINSTEPS, I obtained correlations of the standardized residuals for all

possible item pairs in both Run 2 and Run 7. Using these correlations I computed a Fisher's *Z* statistic (Shen, 1997) for each pair.

Next, I identified whether each item in the test was (a) *independent,* meaning that the item measured a skill that was functionally unrelated to any other item, or (b) *clustered,* meaning that the item measured a similar skill to another item but at a different ability level. Table XV provides descriptions and counts of item pairs for each skill item cluster. Items 1, 13, 15, 19, 32, 35, 36, 40, and 44 are independent; that is, they measure specific skills that are not measured by any other item on the test. By pairing each of these independent items with every other item on the test, I obtained 890 independent item pairs.[7]

TABLE XV
BDI-2 GROSS MOTOR DOMAIN ITEM SKILL CLUSTERS

Skill Cluster	Cluster Contains Items That Measure:	Item Numbers	Number of Item Pairs in Cluster
A: Stairs	child's skill in ascending and descending stairs	20, 25, 26, 29, 34	10
B: Walking, running, skipping	child's skill in locomotion	14, 18, 21, 27, 30, 33, 38, 39, 41, 42	45
C: Head posture	child's skill in maintaining head posture	2, 3, 4, 6, 7	10
D: Transitioning	child's skill in transitioning from one position to another	11, 16, 17, 22, 23, 24	15
E: Catching/ throwing	child's skill in catching and throwing a ball	28, 31, 37, 43, 45	10
F: Perceptual motor	child's skill in using his or her hands to move objects	5, 8, 9, 10, 12	10

[7] Because examiners use basal and ceiling rules to administer only the items that are appropriate for an examinee's ability level, there are some item pairs for which it is not possible to compute correlations because the items were never taken by the same examinee. The 890 item pairs in this analysis were those for which correlations were computable.

Next, I calculated the means (-0.00502, -0.01002) and standard deviations (.031104, 0.04425) of the Fisher's *Z* statistic for the 890 independent item pairs for Runs 2 and 7, respectively. I compared the correlation of the standardized residuals for each clustered item pair with the cutoff values of ±2 standard deviations from the distribution of independent item pair correlations to determine whether each clustered item pair was significantly correlated. Significant correlations represent possible violations of the Rasch model assumption of local independence. Of the 100 clustered item pairs in each run, the standardized residuals from six pairs had significant Fisher's *Z* statistics in Run 2, and four pairs of items had significant Fisher's Z statistics in Run 7. This represented 6% and 4% of the clustered item pairs, respectively. Both of these values were similar to the Type I error rate (i.e., one would expect that 5% of the item pairs would show significant Fisher's *Z* statistics due to chance alone). This finding suggests that there may be some possible local dependence among the BDI-2 Gross Motor items, although it is minimal and may not be greater than the amount that would occur due to chance alone.

E. Research Question 3.1

Research Question 3.1) Are the BDI-2 Gross Motor subdomain scores sufficiently reliable for making inferences about children's development?

I evaluated the reliability of the BDI-2 Gross Motor domain measures using the examinee and item separation reliability indices from the WINSTEPS analysis of Runs 2 and 7. In Run 2, the examinee separation index was 11.85 and the examinee separation reliability was 0.99. In Run 7, the examinee separation index was 12.46 and the examinee separation reliability was 0.99. These large values for the separation index indicate that the BDI-2 Gross Motor subdomain items in their published form and under the collapsed scoring rubric are sensitive enough to separate the examinees in the standardization sample into many distinct ability levels. The high

separation reliability suggests that the items can reliably reproduce these examinee ability measures with this sample, whether the three-category or two-category scoring system is employed. The item separation index in Run 2 was 86.17 and the item separation reliability was 1.00. In Run 7, the item separation index was 88.02 and the item separation reliability was 1.00. The high item separation values in both runs indicate that the examinees in this sample were able to precisely locate the items' locations on the Gross Motor ability continuum; in other words, we can be confident that item calibrations very similar to those would be obtained from this same sample of people measured at another time, regardless of whether the examiners used a two-category or three-category scoring system to assign the scores..

Figures 9 and 10 display standardized conditional reliabilities for the BDI-2 examinees when they are scored with a three-category scoring system (Run 2) and a two-category scoring system (Run 7). Conditional, or *examinee-level*, reliabilities are useful to practitioners for evaluating the precision of examinee scores across the entire range of ability (Raju, Price, Oshima, & Nering, 2006). Conditional reliabilities for examinees in both the three-category and two-category scoring systems are high across the entire range of ability.

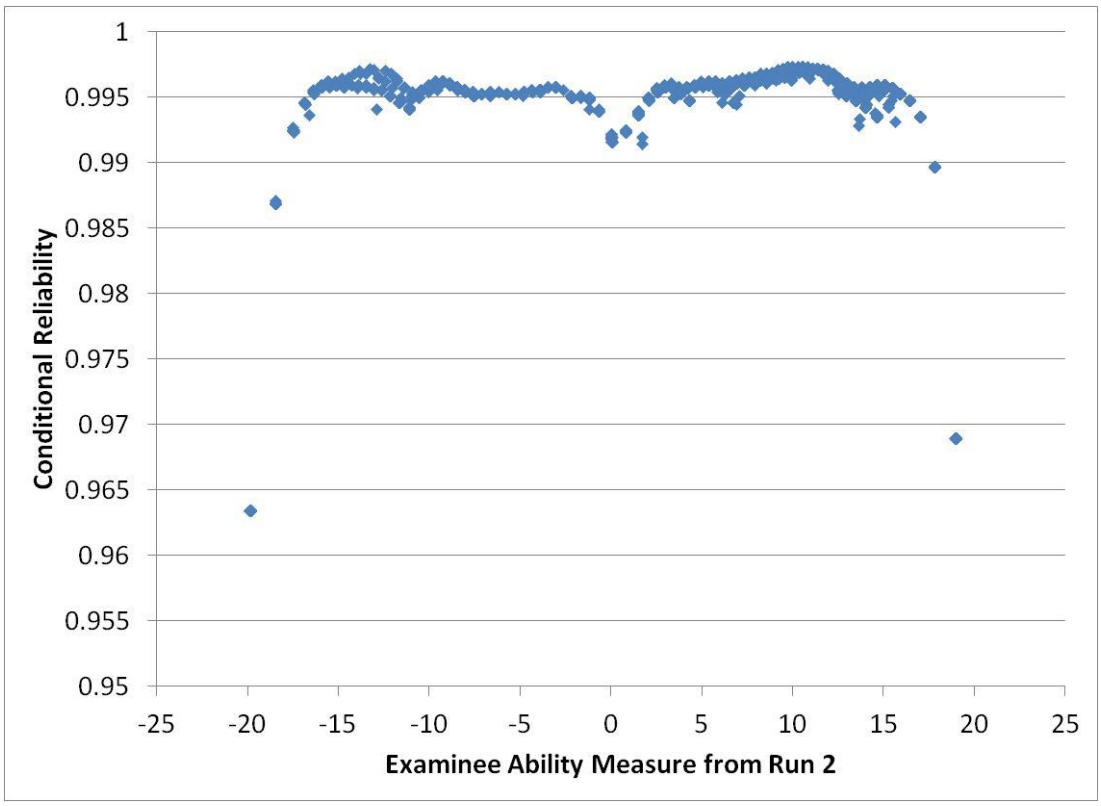

Figure 9. Conditional reliabilities for all BDI-2 examinees scored with a three-point scoring system.

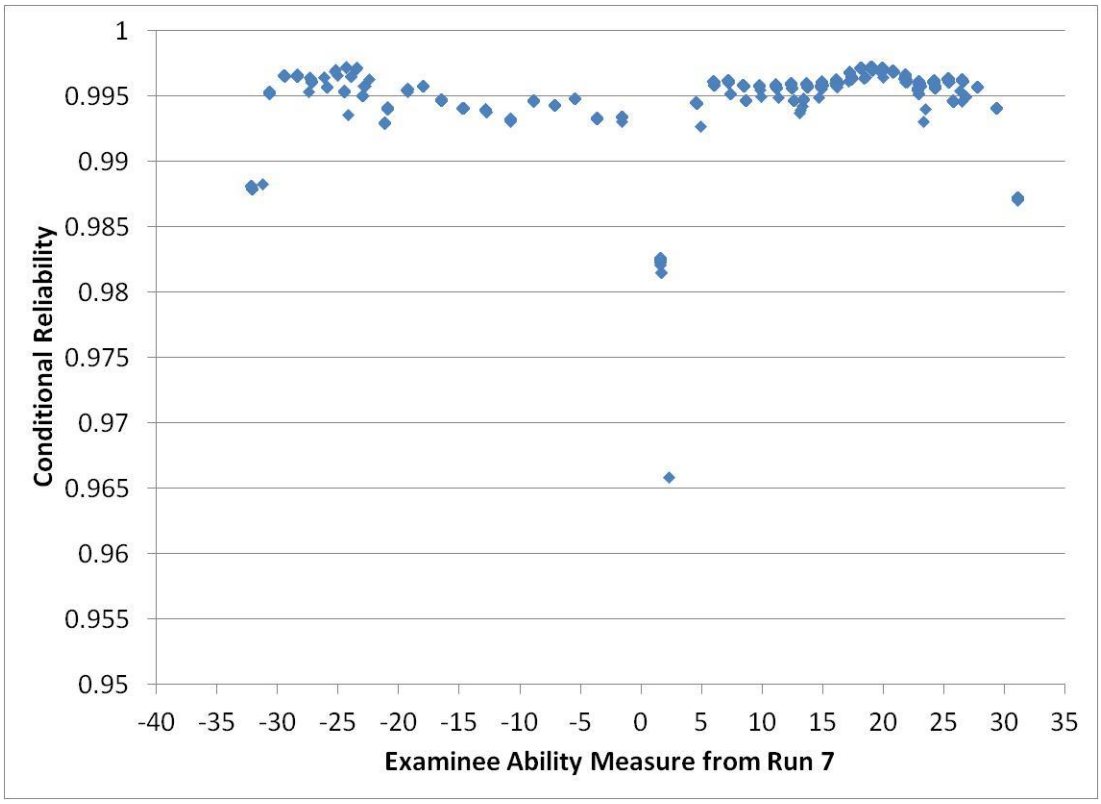

Figure 10. Conditional reliabilities for all BDI-2 examinees scored with a two-point scoring system.

V. DISCUSSION

In this chapter I review the results of the analyses reported in Chapter IV. I discuss how these results might support or refute the use of scores from the BDI-2 as a tool for identification of developmental delay, progress monitoring, and reporting. Based on the results of my analyses in Chapter IV, I suggest some modifications that might be incorporated into future revisions to increase the utility of the test's scores for practitioners. I will also present some ideas for future research using the BDI-2 that might add to the existing body of validity evidence and provide additional score interpretations for this assessment.

A. Review of the Validity Evidence

In this study, my goal was to collect evidence that would either support or refute the use of the BDI-2 Gross Motor domain scores for identifying young children with developmental delay, planning interventions, and reporting on progress. In Chapter I, I presented several challenges, resulting from the use of norm-referenced test scores, that face practitioners tasked with assessing children for identification of developmental delay, planning interventions, and reporting progress for children served by early childhood intervention programs. I introduced the BDI-2 and its Rasch-based Change-Sensitive Score (CSS) metric, and I explained how the BDI-2 CSSs could solve some of these measurement challenges. I introduced several propositions that could provide validity evidence to support the use of BDI-2 scores for the stated purposes. For each proposition, I set forth one or more research questions to guide my analysis of the BDI-2 data. Specifically, I used the guidelines proposed by Wolfe and Smith (2007b) to investigate sources of validity evidence that one could obtain by using the Rasch measurement model to analyze the data. The validity evidence that I gathered adds to the Classical Test Theory–based

validity evidence documented in the BDI-2 Examiner's Manual (Newborg, 2005b) and in several recent studies appearing in peer-reviewed journals.

This chapter summarizes the validity evidence that I gathered in Chapter IV. For each research question set forth in Chapter II, I describe to what extent the validity evidence either supports or refutes the use of BDI-2 scores. Table XVI shows the propositions that I presented in Chapter II, the research questions that guided my investigation, the specific analyses performed to investigate each research question, and the analysis run numbers that correspond to each question.

In Run 1, I identified 170 examinee scores that appeared to be scoring, scanning, or administration errors in the BDI-2 Gross Motor dataset. Prior to performing any analyses, I first cleaned up the dataset by removing these 170 scores, as well as 126 examinees with extreme high or low total scores on the test.

1. Evidence Related to the Substantive Aspect of Validity

The BDI-2 is based on a *developmental model*, a concept that Wolfe and Smith (2007a) defined as one that "make[s] explicit theoretical assumptions about what constitutes higher levels of proficiency" (p. 105). The BDI-2 assesses mastery of a set of generally accepted milestone skills and abilities in early childhood. If BDI-2 scores are to provide valid indicators of delay for use in identifying children who need intervention and for reporting on the progress of children receiving intervention, then these scores must accurately measure a child's mastery of the underlying developmental milestones. Proposition 1 hypothesized that the BDI-2 scores accurately represent children's level of mastery or non-mastery of the Gross Motor developmental milestones. For this proposition to be supported,

TABLE XVI
STUDY PROPOSITIONS, RESEARCH QUESTIONS, METHODS OF INVESTIGATION, AND ANALYSIS RUN NUMBERS

Proposition	Related Research Question(s)	Method(s) of investigation	Analysis Run(s)[a]
1) BDI-2 Gross Motor subdomain scores are useful for accurately describing mastery or non-mastery of gross motor developmental milestones in childhood.	1) Does substantive validity evidence support the current uses of BDI-2 Gross Motor subdomain scores to make inferences about children's development?		
1.1) BDI-2 Gross Motor subdomain score categories are appropriate for obtaining information about a child's current level of gross motor development.	1.1) Do the examiners use the BDI-2 score categories of 2, 1, and 0 as expected to label behaviors that are fully emerged, emerging, and not yet emerged?	• Examination of rating scale functioning using Linacre's (2004) guidelines • Examination of item mean-square fit statistics	2,3,4,5, and 6
1.2) The BDI-2Gross Motor subdomain item hierarchy accurately represents the underlying milestone theory of development.	1.2) In the standardization dataset, are there any anomalous examinee score strings that may have degraded the quality of the item calibrations?	• Examination of examinee mean-square fit statistics	6,7
2) The BDI-2 Gross Motor subdomain scores provide meaningful measures of the distinct abilities that contribute to a child's development.	2) Does structural validity evidence support the use of BDI-2 Gross Motor subdomain scores to make inferences about children's development?		
	2.1) Do the data from the BDI-2 Gross Motor subdomain represent one dominant underlying dimension?	• Examination of point-measure correlations • PCA analysis of Rasch residuals • Examination of item mean-square fit statistics	2,7
	2.2) Do the data from the BDI-2 Gross Motor subdomain satisfy the Rasch model requirement of local independence?	• Examination of residual correlations (standardized with Fisher's *Z*) for all clustered item pairs	2,7

TABLE XVI
STUDY PROPOSITIONS, RESEARCH QUESTIONS, METHODS OF INVESTIGATION, AND ANALYSIS RUN NUMBERS (CONTINUED)

Proposition	Related Research Question(s)	Method(s) of investigation	Analysis Run(s)*
3) BDI-2 Gross Motor subdomain scores are precise enough to locate an examinee's ability on the scale and reveal changes in ability over time.	3) Does evidence relevant to the generalizability aspect of validity support the use of BDI-2 Gross Motor subdomain scores to make inferences about children's development?		
	3.1) Are the BDI-2 Gross Motor subdomain scores sufficiently reliable for making inferences about children's development?	• Examination of Rasch item and examinee separation reliability and separation indexes	2,7

[a]Run 1 included all items and people. Prior to the investigation of Research Question 1.1, I discovered in Run 1 that the original standardization dataset contained 170 scores that were likely scanning or administration errors. I removed these scores and then conducted Run 2 with the clean dataset that I used to begin the investigation of Research Question 1.1.

both the score category descriptions (Proposition 1.1) and the item task requirements (Proposition 1.2) would need to support the developmental milestone theory underlying the BDI-2. Wolfe and Smith (2007b) termed this type of evidence “substantive” and suggested that the results from a Rasch measurement analysis can provide important information relevant to the substantive aspect of validity by assessing whether examinees respond to the items in an expected manner, and (in the case of examiner scoring) whether examiners use the score categories in a way that is consistent with the intentions of the test developer. To investigate whether the substantive validity evidence obtained from the Rasch measurement analysis supports the use of the BDI-2 Gross Motor subdomain scores for making inferences about children's development, I posed two questions.

Research Question 1.1) Do the BDI-2 examiners use the BDI-2 scores of 2, 1, and 0 as expected to label behaviors that are fully emerged, emerging, and not yet emerged?

The BDI-2 test developer designed the score category descriptions so that a child's level of development on any item could be described as "not yet emerging," "emerging," or "fully emerged," using score categories of 0, 1, and 2, consistent with the underlying developmental milestone theory. To determine whether examiners in the standardization study used the Gross Motor score category descriptions in the intended manner, I used Linacre's (2004) guidelines for investigating rating scale functioning. Guideline 1 requires that each score category of each item must contain a minimum of 10 observations. I found that of the 135 possible score categories for the BDI-2 Gross Motor items, four score categories had fewer than 10 scores in my dataset. Of these, three were for the score category of 1. Guideline 2 requires approximately equal score category usage across items. I noted that score category 1 was the least used score for all but two items, and that the frequency of its category usage ranged from 2% to 26%. Guideline 3 suggests

analysis of the average measures for the categories within each item (i.e., the average ability of all examinees obtaining each score on each item), to ensure that these increase with ascending score categories. I found that the average examinee ability measures for ascending score categories of all items increased, indicating that for this sample of children, each successive score category did represent a higher level of gross motor development. Guideline 4 suggests that mean-square outfit statistics greater than 2.0 for score categories might indicate unsystematic variance in the data, which in turn may negatively affect the item and category calibrations. I noted that the score category of 0 was misfitting for five items, and that the score category of 2 was misfitting for three items. I hypothesized that these misfits in the BDI-2 data were likely the result of either unusual examinee skill profiles or errors in data entry or scanning. When I identified and removed 19 additional examinee scores with residual *z*-scores outside the range of +/-2.0 in Run 3, score category fit improved markedly. Finally, guideline 5 requires that successive category thresholds advance for each item. I examined the step calibrations for all 45 items and determined that 36 items had step calibrations that were disordered. This is compelling evidence that the examiners in this study often did not assign scores of 1 to examinees who displayed emerging skills on the BDI-2 items.

The information that I collected during my analysis of how examiners used the score categories suggested that the examiners tended to score most of the BDI-2 Gross Motor items using a two-category scoring system, rather than the three-category system the test developer provided. I suggested several possible reasons for this tendency. On many items, the criteria for receiving a score of 1 are so specific as to minimize the likelihood that a child will exhibit the exact behavior necessary for that score. Additionally, some of the items require examiners to distinguish between very short, precise time intervals to determine whether a response should be

scored as 1 or 2. For instance, one item requires the examiner to distinguish between a response that is 3 seconds long (for a score of 2) and a response that is 1 or 2 seconds long (for a score of 1). I hypothesized that collapsing the score categories 1 and 0 for each item to create a two-category scoring system might result in better data fit to the model and more efficient administration for examiners, without losing any information about examinee ability.

To investigate this hypothesis, in Run 4 I collapsed score categories 1 and 0 so that anything less than a full-credit (score of 2) response would represent a skill that is still emerging or not yet emerged. In this rescoring, I changed scores of 1 to 0 and scores of 2 to 1. Overall item fit increased, indicating that the two-category scoring system yielded examinee score strings that were more expected, under the Rasch model, than the three-category scoring system.

After addressing issues with the several additional unexpected scores after category collapsing, I reviewed the score category descriptions for several items.

The original intent of the test developer in providing score category 1 for the BDI-2 Gross Motor items was to allow examiners to award partial credit for a skill that appeared to be emerging at the time of testing. However, the analyses that I performed suggested that most of the BDI-2 Gross Motor items function essentially as two-category items, because examiners tend to award either no credit or full credit to the great majority of examinees. Although the item difficulties formed a hierarchy that was consistent with theoretical expectations, the score categories for most items did not function as expected. In many cases, score category 1 was never the most probable score at any point across the range of ability for some items. Based on my item analysis and review of the test content, I do not believe that this category disordering is due to a lack of adherence to the underlying developmental milestone theory; rather, I believe that it more likely due to the design of the score category descriptions. For many items, the

requirements for awarding a score of 1 were either overly vague, or were so specific and narrow that examiners seldom observed examinees responding at that level of ability in the standardization study. When I collapsed score categories 1 and 0 to create a two-category scoring system, the overall fit of the data to the model was comparable to the fit of the data using the three-category scoring system. The item difficulty rank order did not change, and the examinee ability measures correlated highly with the original measures (r = .997). Additionally, the resulting scoring system became much simpler to use; for most items, a score of 0 would now be "anything less than a score of 2."

A cross-plotting of the examinee ability measures obtained from the three-category scoring system and the two-category scoring system revealed that the two-category scoring system tended to produce lower examinee ability measures for examinees at the extreme low end of the ability distribution, but very similar ability measures for examinees in the middle- and at the high end of the ability distribution. Additionally, the only examinees whose ability measures differed by an amount greater than the standard error of measurement were those for whom the standardization examiners had assigned the score of 1 most often in the original dataset.

In summary, the evidence gathered for Research Question 1.1 suggests that examiners did not use the three-category scoring system for the Gross Motor subdomain as the test developer intended to differentiate between children whose skills are not emerged, emerging, and fully emerged. This evidence does not support the three-category scoring system in the current BDI-2. Linacre (2010) describes score categories such as the BDI-2 score of 1 as *transitional categories*. According to Linacre (2010),

> Transitional categories correspond to narrow intervals on the latent variable. They may indicate growth states of short duration. They are usually less frequently observed than the neighboring dominant categories. A result is that the probability of observing transitional categories tends to

> be low, but...users of the measures certainly need to be aware of where on the latent variable there is some probability of them occurring. (p. 8)

Linacre (2010) goes on to suggest that although Rasch practitioners have traditionally perceived disordered Rasch-Andrich thresholds as an indication of some failure in the scoring system (and therefore a threat to valid measurement), perhaps these disordered thresholds, which may indicate transitional categories, should be considered as an "integral and increasingly important part of the advance of social science" (p. 10). Linacre suggests that practitioners evaluate the Rasch-Thurstone thresholds, rather than the Rasch-Andrich thresholds, for transition score categories. Rasch-Thurstone thresholds describe the points along the ability continuum where the cumulative probability of receiving a score in the category or the next-higher category is equal to the probability of receiving a score in any lower category. These Rasch-Thurstone thresholds demonstrate the presence of a transition category that is real, but is perhaps "squeezed between two dominant categories" (Linacre, 2012, p. 7) on the latent trait. In the case of the BDI-2 three-category scoring system, the category 1 may be useful for describing emerging skills, and perhaps the "transitional category" approach using the Rasch-Thurstone thresholds would allow practitioners to investigate this. However, my review of the score category descriptions suggests that many of the current descriptions for scores of 1 are either vague or so overly specific that they may have limited utility in practice. If BDI-2 scores of 1 were to be useful, the score category descriptions would need to be operationalized in such a way to capture the true essence of the emerging skill, rather than simply describing a level of the skill or ability that is rarely observed in practice.

Although the two-category scoring results in better model fit, the examinee ability measures obtained from the two-category and three-category scoring systems are very highly correlated. The scatterplot of examinee ability measures from Run 2 and Run 6 suggests that,

except in the case of a few examinees who received scores of 1 most frequently, the choice between these two scoring systems produce very similar examinee ability measures. Additionally, many of the descriptions for score category 1 were long and detailed; for this reason, implementation of the two-category scoring system in future versions of the BDI would greatly reduce the examiner's cognitive load and would decrease testing time.

Research Question 1.2) In the standardization dataset, are there any anomalous examinee score strings that may have degraded the quality of the item calibrations?

In a developmental model, the extent to which the item hierarchy reflects the underlying model provides evidence to support or refute the claim that the test scores function as accurate measures of examinees' levels of development. For this reason, it is important that the dataset used for item calibration is free from administration or data entry errors. For Research Question 1.2, I was interested in determining whether there were misfitting or unexpected scores that adversely affected the calculation of the difficulty measures for the Gross Motor subdomain items. Rasch examinee fit statistics are useful for identifying examinees whose score strings do not contribute to the useful calibration of the items. Examinee mean-square outfit statistics greater than 1.5 indicate that an examiner has assigned at least one score that is unexpected, given the examinee's ability. In the original dataset, the overall examinee mean-square outfit statistic was 0.83 with a standard deviation of 1.60. The outfit statistic is well within the acceptable range for examinee fit (i.e., close to 1.0), but the relatively high standard deviation indicates a high degree of variability in these statistics across all examinees.

While investigating Research Question 1.1, I identified and removed individual scores that were extremely unexpected. In most cases I attributed these highly unexpected scores to examiner administration errors, scanning errors, or ambiguity in the score category descriptions.

To determine whether the examinee misfit I identified during the investigation of Research Questions 1.1 and 1.2 impacted the item difficulty measures, I compared the item difficulty measures from the original run (Run 2) to those from the last run (Run 7). The item rank-order correlation was 0.999, and the correlation of the two sets of item difficulty measures was 0.992. With the exception of two very minor shifts in rank ordering, the scatterplot of the item difficulty measures from these two runs showed a nearly linear relationship. This suggests that, although the item difficulty measures showed greater spread with the removal of the unexpected scores, the inclusion of the unexpected scores in the original data did not adversely affect the calibration of the items. This finding supports the use of the current BDI-2 scores to make inferences about children's gross motor development for identification and progress reporting; although there were some unexpected scores in the standardization data that resulted in misfit to the Rasch model, removing these unexpected scores did not significantly change the calibration of the items.

2. Evidence Related to the Structural Aspect of Validity

In addition to noting the benefits of using Rasch measurement procedures to gather substantive validity evidence, Wolfe and Smith (2007b) also suggested that the Rasch model assumptions of unidimensionality and local independence, if met, can provide structural validity evidence to support a test's score interpretations. The assumption of unidimensionality requires that the test items measure one and only one latent trait. The assumption of local independence requires that an examinee's performance on one item does not impact his or her performance on other items, after controlling for overall examinee ability. If one or both of these assumptions are not met, then test users cannot be certain that they are effectively measuring the abilities that the test claims to measure, and score interpretations become unclear.

The BDI-2 domain and subdomain structure allows examiners to assess children's developmental abilities in broad areas (such as motor or cognitive areas) using domain-level scores, and developmental abilities in narrow, specific areas (such as gross motor or attention and memory), using subdomain scores. Practitioners who use the BDI-2 subdomain scores for identification of delay, planning interventions, and reporting progress to OSEP need to be certain that the scores do not contain construct-irrelevant variance; in other words, it is important that each subdomain measures what it is supposed to be measuring. Proposition 2 asserts that the BDI-2 Gross Motor subdomain scores provide meaningful measures of distinct developmental abilities. If this proposition is true, then this structural validity evidence would support the use of the BDI-2 Gross Motor scores for identification, planning, and reporting. To determine whether the evidence gleaned from my analysis supports or refutes the use of the BDI-2 Gross Motor subdomain scores to make inferences about children's development, I posed two questions. For each of these questions, I was interested in whether the validity evidence supported or refuted the use of BDI-2 scores from the current test, which uses a three-category scoring system, as well as the scores from using a two-category system. For this reason, I performed the analyses for Research Questions 2.1 and 2.2 using the datasets from Run 2 and Run 7.

Research Question 2.1) Do the standardization data from the BDI-2 Gross Motor subdomain represent one and only one underlying dimension?

I used three different approaches to examine the dimensionality of the BDI-2 Gross Motor data. First, I examined the item point-measure correlations for Runs 2 and 7, finding that they were all positive and moderate to large, with most of them between .60 and .80. This finding suggests that each BDI-2 Gross Motor item is correlated with the total subdomain score. Next I examined the item mean-square outfit statistics for evidence of misfit. The results for the

analysis of the original dataset (Run 1, prior to any collapsing or removal of unexpected scores) revealed that 11 items out of 45 had mean-square outfit statistics greater than 1.5. Removal of the 170 scores that suggest apparent administration or scanning errors in Run 2 decreased the number of misfitting items from 11 to 3. After removing the unexpected scores and collapsing score categories 1 and 0 during the investigation of Research Question 1.1, the overall item fit improved dramatically, and the number of misfitting items decreased to three. After removing several additional misfitting scores for Run 7, the number of misfitting items decreased to one.

Finally, I performed a principal components analysis of the Rasch residuals to determine whether the additional factors in the data contributed to multidimensionality. In Run 2, the first principal component accounted for 72.7% of the variance in the data; in Run 7, the first principal component accounted for 70.6% of the variance in the data. Of the unexplained variance in Runs 2 and 7, the first contrast accounted for 6.2% and 4.3%, respectively. The patterns of positive and negative loadings on the first contrast did not suggest that there were content-specific relationships between these items in either dataset. Simulated datasets showed first-contrast eigenvalues very similar to those in the empirical data for Run 2 and Run 7, lending support for the assumption of unidimensionality in both cases. The results from this analysis support use of the scores to make inferences about children's development. In both the standardization dataset and the revised (Run 7) data, the items appear to be measuring one distinct construct—gross motor ability—without an apparent second dimension. Test users can be confident that the scores for this subdomain are not confounded by construct-irrelevant variance or the unintentional measurement of some other ability.

Research Question 2.2) Do the standardization data from the BDI-2 Gross Motor subdomain satisfy the Rasch model requirement of local independence?

To test the assumption of local independence, I compared the standardized residuals for all item pairs in both Run 2 and Run 7. In each case, I obtained correlations of the standardized residuals and then computed a Fisher's *Z* statistic for each pair. Next I determined whether the test developer intended each item in the test to be independent or clustered, based on the gross motor skill the item measured. Using the mean and standard deviation of the distribution of the Fisher's *Z* statistics for the independent item pairs, I determined the critical value of *Z* equal to ±2 standard deviations from the mean of the independent item pairs. I considered clustered item pairs with *Z* values outside this range to be significantly correlated. Of the 100 clustered item pairs in the Gross Motor subdomain, 4% had significant *Z* statistics in Run 2, and 6% had significant *Z* statistics in Run 7. These percentages are both similar to the Type I error rate of 5%, indicating that in any set of independent items, we expect that approximately 5% would appear to be significantly correlated due to chance alone. While there may be one or more instances of item dependence among the clustered BDI-2 Gross Motor items, my findings generally suggest that the local item independence assumption was met under both the current BDI-2 scoring system and the dichotomous scoring system.

These results provide structural validity evidence to support the use of the BDI-2 Gross Motor subdomain scores. Practitioners using the BDI-2 to identify developmental delay can be confident that the items from the Gross Motor subdomain are indeed measuring gross motor development and not any other skills or traits. This evidence is especially useful in planning interventions, as practitioners can use the information gleaned from the Change-Sensitive Scores to develop a tailored intervention plan for each child.

3. Evidence Related to the Generalizability Aspect of Validity

The results from a Rasch measurement analysis can provide estimates of internal consistency, or reliability, for both items and examinees. High item separation reliability suggests that the items are well spread across the difficulty continuum; high examinee separation reliability suggests that the set of test items is able to separate the examinees into several different levels of ability. If practitioners are to use the Gross Motor scores from the BDI-2 to identify developmental delay, the scores must be sufficiently precise to measure a child's ability accurately. Additionally, if practitioners use the BDI-2 to monitor progress, the items must produce a large enough spread of examinee ability to permit detection of meaningful improvement between test administrations. Proposition 3 asserts that the BDI-2 Gross Motor subdomain scores are sufficiently precise to locate examinees on the ability scale and to reveal changes in ability over time. To determine whether generalizability validity evidence obtained from my analysis supports or refutes the use of the BDI-2 Gross Motor subdomain scores to make inferences about children's development, I posed the following question.

Research Question 3.1) Are the BDI-2 Gross Motor subdomain scores sufficiently reliable for making inferences about children's development?

I obtained reliability and separation information from an analysis of the original BDI-2 Gross Motor standardization data (Run 2) and from an analysis of the dataset after optimizing the scoring system and removing unexpected scores (Run 7). In Run 2, the examinee separation index was 11.85 and the examinee separation reliability was 0.99. In Run 7, the examinee separation index was 12.46, and the examinee separation reliability was 0.99. These large values for the separation indices indicate that, for this sample, the BDI-2 Gross Motor subdomain items were sensitive enough to separate the sample of examinees into many distinct ability levels,

while the high reliabilities of the separation suggest that the items can reliably reproduce these examinee ability measures with this sample. In Run 2, the item separation index was 86.17 and the item separation reliability was 1.00. In Run 7, the item separation index was 88.02 and the item separation reliability was 1.00. These findings indicate that the items in this sample were able to precisely locate the items' locations on the Gross Motor ability continuum, and that these locations would be reproducible with this sample of examinees or another sample that is of similar ability. These reliability data provide evidence related to the generalizability aspect of validity, because we can be confident that the range of examinee ability in the standardization sample was appropriate for obtaining accurate item difficulty measures; in other words, the items were well targeted to the abilities of the examinees in the sample. Because the test publisher carefully chose the standardization sample to represent the entire age range from birth to 8 years in terms of gender, region, community type, race, and ethnicity, we can be confident that the high examinee and item separation reliabilities obtained in this study would generalize to other similar samples of examinees. Additionally, the high level of examinee separation reliability suggests that the BDI-2 is able to distinguish among many distinct levels of ability, providing evidence that the BDI-2 Gross Motor scores are suitable for detecting change in examinee ability over time. This evidence supports the use of the BDI-2 as a tool for tracking child progress under the current OSEP system.

B. Suggestions for Changes to Future Editions of the BDI

In Chapter IV, I reported results from my analyses of the BDI-2 Gross Motor standardization data. These analyses led me to conclude that most of the validity evidence gathered supports the use of BDI-2 scores to make inferences about children's gross motor

development. However, some findings in the analyses suggest that changes to future editions of the BDI-2 might be appropriate.

Perhaps the most substantive finding is that the majority of the BDI-2 Gross Motor items could be diagnostically useful with a two-category scoring system. In my analyses, I collapsed the score categories of 1 and 0. This change resulted in better category and overall model fit. In addition to the psychometric advantages for item and score interpretation, perhaps an even bigger advantage to collapsing the score categories 1 and 0 is the increased efficiency of test administration for examiners and examinees. If the score categories of 1 and 0 were collapsed, the multi-part or ambiguous category 1 descriptions for many items would be eliminated, reducing the examiners' cognitive load and simplifying the scoring process.

In Chapter IV I also suggested some minor changes to specific item administration procedures that might improve data fit to the model. For Item 14 ("Child makes stepping movements when held in an upright position"), I suggested that allowing examiners to use an Interview or an Observation administration procedure might increase the likelihood that an examiner would assign full credit for this item to children who demonstrate that skill. For Item 5 ("Child brings hands together at midline"), I suggested that allowing examiners to use a Structured administration procedure would eliminate the need for examiners, parents, or caregivers to interpret the intention of the child's random hand movements. Finally, for Item 9 ("Child puts objects into his or her mouth"), I believe that allowing examiners to use a Structured or Observation administration procedure would permit them to validate the response that a parent or caregiver provides during the Interview administration procedure.

C. Limitations of the Current Study

1. Findings Are Not Generalizable to Other BDI-2 Subdomains and BDI-2 Composite Scores

The data I used for this study came from the BDI-2 standardization study. Independent examiners that the test developer employed gathered the data between 2002 and 2003. The purpose of the standardization study was to create norms for the complete BDI-2, which contains 13 subdomains. The main purpose of this study was to show how using a Rasch measurement approach to data analysis might contribute to the body of validity evidence to support the use and interpretation of the BDI-2 scores for making inferences about children's development. Although it would be useful to gather this type of validity evidence for each subdomain of the BDI-2, conducting similar analyses for 13 subdomains (i.e., a total of 450 items) was beyond the scope of this dissertation. Therefore, I chose to analyze only the items included in the Gross Motor subdomain.

I chose the Gross Motor subdomain for several reasons. First, it relies heavily on the use of a Structured administration procedure, which requires examiners to elicit an actual response to a stimulus; this procedure minimizes the introduction of construct-irrelevant variance that may be present when examiners employ Observation or Interview procedures. Also, the approach that test developers use to measure gross motor development has not changed significantly over time, and other measures of early childhood development assess similar skills , indicating that the test items are not susceptible to revisions over time due to changes in early childhood development theory.

The disadvantage of choosing one subdomain for my analyses is that I cannot make general inferences about the validity of the current uses of other BDI-2 subdomain scores or the BDI-2 composite scores from the results of my study. Researchers interested in gathering

validity evidence for that purpose would need to perform additional, similar analyses on the test data for the remaining 12 subdomains.

2. Analyses Do Not Take Administration Procedures Into Account

Another limitation of the current study stems from the inclusion of one, two, or three different administration procedures for each BDI-2 Gross Motor item. Recall that each item allows examiners to use one or more of the following administration procedures: Structured, Observation, and/or Interview. Using the Structured procedure, the examiner assigns a score based on the examinee's response to a prescribed stimulus. Using the Observation procedure, the examiner assigns a score based on whether the examiner saw the child perform a specific action or behavior. Using the Interview procedure, the examiner asks the child's parent or caregiver questions about the frequency and duration of the child's behaviors, and assigns a score based on the feedback that the parent or caregiver provides.

According to the BDI-2 Examiner's Manual (Newborg, 2005b), the test developer prefers the Structured administration procedure since it requires the examiner to witness the child performing a specific action or behavior. However, the assessment of young children frequently does not lend itself to the use of a structured testing approach; children might be tired, scared, bored, or—in the case of very young children—even sleeping during the testing session. For this reason the test developer allowed examiners to use the Observation and Interview procedures for many items.

I did not take these differences into account when I analyzed the data. In other words, I treated a score of 2 on a given item the same as any other score of 2 on that item, regardless of the administration procedure that the examiner used to assign the score. There are two reasons why I did not account for differences in administration procedures in my analyses. First, for

many items the standardization dataset did not contain enough scores for some administration procedures to provide stable item calibration. Second, the BDI-2 does not distinguish between these administration procedures in the assignment of an examinee score. Because the purpose of my study was to gather validity evidence for the current uses of the BDI-2, I treated the administration procedures as essentially equivalent for the purposes of my analyses.

3. Item Calibrations Do Not Take into Account Possible Differences in Examiner Severity

Similar to many individually administered clinical tests, an examiner administers the BDI-2 and assigns scores based on an examinee's responses or behaviors. Because examiners must observe and score the examinee's responses and behaviors, one could posit that some degree of examiner subjectivity may be present in the BDI-2 scores. The test developer took great care to remove as much ambiguity as possible from the BDI-2 scoring. For example, the score category descriptions for each item describe the specific actions required for a score of 2, 1, or 0. The test developer also provided extensive training for all examiners participating in the standardization study prior to the start of data collection.

Even with the most rigorous training and detailed score category descriptions, however, there may have been instances during the standardization study when an examiner needed to use subjective judgment to score an examinee's response. It is possible that, in some cases, a different examiner observing the same response or behavior might have assigned a higher or lower score for that item. If test developers predict that inter-examiner differences in judgment or severity might result in the assignment of different scores to the same behavior, it is often useful to treat the effects of the individual examiners as a “facet” in a many-facets Rasch measurement (MRFM) approach to data analysis. This approach is often used in licensure testing, for example,

when candidates are required to perform a series of tasks while two or more examiners, or raters, observe and score each candidate's performance.

The question of whether differences in the levels of severity that individual examiners exercised might have impacted the BDI-2 item calibrations is an intriguing one. Unfortunately, a MFRM analysis is not possible using the BDI-2 standardization dataset, because such an analysis requires that multiple examiners observe and score at least some of the examinees. These multiple examiners' scores provide the connectivity the model requires in order to compare the effects of differences in examiner severity on the estimates of item difficulty and examinee ability. Although a small study conducted with BDI-2 tryout data provided strong evidence that the administration procedure that the examiner chose did not introduce construct-irrelevant variance by changing the difficulty of the item (Pomplun & Custer, 2004), the BDI-2 standardization study design did not allow for multiple examiners to score children's performances on the test items. Therefore, it was not possible to investigate the impact of differences in examiner severity on examinees' scores.

D. Directions for Future Research

In the prior section I described some limitations of this study that were primarily due to restrictions on time and resources, or lack of appropriate data. Each of these limitations presents an opportunity for future research that might add to the body of validity evidence to support the current uses of the BDI-2. For example, a researcher might want to replicate the analyses I conducted using data for one or more of the other 12 BDI-2 subdomains to determine if the results support or refute the use of the individual subdomain scores to make inferences about children's levels of development in each of those areas. MFRM analyses might provide interesting insight into the equivalency of the Structured, Observation, and Interview procedures

for each item and the equivalency of examiner's scores. These last two types of analyses would require the collection of additional data. To study the effects of administration procedures, each item would have to be scored using at least two different administration procedures for the same child. To study the effects of differences in examiner severity, at least two different examiners would have to score each examinee. While these types of studies would provide valuable evidence to support or refute the current practice of treating administration procedures and examiners as equivalent, they are generally quite expensive and logistically complex; thus, test developers do not often conduct these types of studies.

The purpose of this study was to gather and present additional validity evidence to support or refute the current uses of the BDI-2 Gross Motor subdomain scores. The use of the BDI-2 for determining delay and reporting progress would make interpretation of the results from early childhood testing programs comparable across states and municipalities. For the most part, the existing body of validity evidence for the BDI-2 supports the use of BDI-2 scores for identifying children with developmental delays. This study added some additional types of validity evidence to support the BDI-2. If the test were adopted for use as a universal measure, researchers could carry out studies to determine cut scores representing delay for each age and each developmental subdomain.

There are many documented procedures for setting cut scores (generally referred to as standard setting); the choice of procedure depends on such factors as the type of test data available, the purpose for the standard setting, and the restraints on time and budgets. One example of a possible method for determining BDI-2 cut scores for developmental delay might be to use the Rasch-based Change-Sensitive Score (CSS) metric for the BDI-2 along with adaptations of the Bookmark Standard Setting procedure (Lewis, Mitzel, & Green, 1996).

The traditional Bookmark Standard Setting procedure is an item-based method in which an expert panelist examines the test items in ascending order of difficulty and then places a "bookmark" on the item that represents the point on the item difficulty continuum where a proficient examinee should respond correctly to all previous items. The Bookmark procedure is an appropriate method for setting standards on a test for which measures of item difficulty and examinee ability are on the same scale. As Cizek and Bunch (2007) stated, "Once participants [i.e., expert panelists] provide page numbers, the associated theta values have a built-in relationship to scores, and results can be interpreted in the same manner as other procedures carried out with these tests" (p. 160). Additionally, the Bookmark procedure accommodates test items with multiple score points, such as the BDI-2 items.

The Bookmark procedure and its modifications have been used extensively in state achievement testing programs and in the National Assessment of Educational Progress (NAEP) testing program (Mitzel, Lewis, Patz, & Green, 2001). Currently there is no published literature on the use of a Bookmark procedure or a similar method to set cut scores on an early childhood developmental assessment.

In a large-scale achievement test, one or more cut scores are set to represent the boundaries between performance categories—for example, between "limited knowledge," "proficient," and "advanced." These same cut scores apply to all examinees who take that test. On a test such as the BDI-2, where examiners administer different sets of items to examinees of different ages, establishing only one cut score would not be appropriate (i.e., it would not make sense to impose the same cut score on both 3-year-olds and 5-year-olds). For this reason, researchers would need to establish different cut scores for each age range covered by the test.

As the norms associated with the BDI-2 distinguish 48 age groups, establishment of 48 separate cut scores would be required.

Finally, although I have presented the BDI-2 as an instrument that is appropriate for identifying developmental delay, monitoring the progress of children served by early childhood special education programs, and reporting progress to OSEP, it is only one of many instruments currently available to early childhood diagnosticians and practitioners. As I discussed in Chapter I, the currently available instruments vary widely with respect to the domains measured, types of scores provided, and their technical quality. For this reason, it would be helpful for early childhood practitioners to know how children's scores on the BDI-2 would compare with scores on other, similar measures. Researchers interested in this question could develop a system for establishing concordance between scores from the BDI-2 and from other early childhood assessments. In the context of the OSEP reporting requirements, a concordance study between the BDI-2 and the COSF would be particularly useful.

E. Chapter Summary

The purpose of this study was to gather validity evidence to support or refute the use of the BDI-2 Gross Motor subdomain scores for identifying developmental delay, designing interventions, and monitoring progress. I posed several research questions to guide the collection of structural, substantive, and generalizability evidence relevant to the validity of the BDI-2 scores for these purposes. While the substantive validity evidence I gathered did not support the three-category scoring system that the BDI-2 employs, the alternative two-category scoring system I proposed did not rank order the examinees in a significantly different way. I found that the BDI-2 standardization dataset contained some very unexpected examinee scores. While these anomalous scores contributed to score category and overall item misfit, they did not appear to have a significant effect on the item

measures. This evidence supports the use of the BDI-2 scores, especially the Rasch-based Change-Sensitive Scores, as the item difficulty hierarchy appears to be robust to unexpected item scores in the dataset that was used to construct the CSS scale. In my investigation of the substantive validity evidence, I found that the BDI-2 standardization dataset met the Rasch model assumptions for unidimensionality and local independence. The evidence relevant to the generalizability aspect of validity that I gathered suggests that the items in the BDI-2 Gross Motor subdomain were able to separate the examinees into many distinct levels of ability, and that these levels would be reproducible with another similar sample of examinees. Additionally, the examinees in the sample were able to precisely locate the items on the difficulty continuum and that another, similar sample of examinees would produce the same item difficulty measures. I concluded the discussion of my study by describing some of the limitations for generalizing my study results, and suggesting some areas for future research.

IV. REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest, Hungary: Academaia Kiado.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing.* Washington, DC: AERA.

Bailey, D. B., Vandiviere, P., Dellinger, J., & Munn, D. (1987). The Battelle Developmental Inventory: Teacher perceptions and implementation data. *Journal of Psychoeducational Assessment, 3,* 217–226.

Bayley, N. (2006). *Bayley Scales of Infant and Toddler Development* (3rd ed.). San Antonio, TX: Harcourt Assessment.

Beelman, A., & Bambring, M. (1998). Implementation and effectiveness of a home-based early intervention program for blind infants and preschoolers. *Research in Developmental Disabilities, 19*(3), 225–244.

Binet, A. (1916). New methods for the diagnosis of the intellectual level of subnormals. In E. Kite (Ed. & Trans.), *The development of intelligence in children.* Baltimore, MD: Williams and Wilkins (Original work published 1905). Retrieved January 3, 2007, from http://psychclassics.yorku.ca/Binet/binet1.htm

Boyd, R. D. (1989). What a difference a day makes: Age-related discontinuities and the Battelle Developmental Inventory. *Journal of Early Intervention, 13*(2), 114–119.

Bracken, B. (1987). Limitations of preschool instruments and standards for minimal levels of technical adequacy. *Journal of Psychoeducational Assessment, 4*, 313–326.

Bracken, B. (1991). *The psychoeducational assessment of preschool children.* Boston, MA: Allyn and Bacon.

Campbell, F. A., & Pungello, E. (2000, June). *High quality child care has long-term educational benefits for poor children.* Paper presented at the 5th Head Start National Research Conference, Washington, DC.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests.* Thousand Oaks, CA: Sage.

Cohen, J. A., & Mannarino, A. P. (1996). Factors that mediate treatment outcomes of sexually abused preschool children. *Journal of the American Academy of Child & Adolescent Psychiatry, 35*(10), 1402–1410.

Connolly, A. J., Nachtman, W., & Pritchett, E. M. (1971). Keymath Diagnostic Arithmetic Test. Circle Pines, MN: American Guidance Service.

Early Childhood Outcomes Center. (2012). Overview of the Child Outcomes Summary Process. Retrieved from http://projects.fpg.unc.edu/~eco/assets/pdfs/COSF_overview.pdf

Early Childhood Technical Assistance Center. (2006). *Overview of the Child Outcomes Summary Form.* Retrieved from http://www.fpg.unc.edu/~eco/pdfs/COSF_overview_9-29-06.pdf

Early Childhood Technical Assistance Center. (2007). *BDI-2 Crosswalk.* Retrieved from http://ectacenter.org/~pdfs/eco/BDI-2_crosswalk_10-10-07.pdf

Elbaum, B., Gattamorta, K. A., & Penfield, R. D. (2010). Evaluation of the Battelle Developmental Inventory, Second Edition, screening test for use in states' Child Outcomes Measurement Systems under the Individuals with Disabilities Education Act. *Journal of Early Intervention, 32*(4), 255-274.

Elliott, C. D. (1990). Differential Ability Scales. San Antonio, TX: Pearson.

Elliott, C. D. (2007). *Differential* Ability Scales-II. San Antonio, TX: Pearson.

Elliott, C. D., Murray, D. J., & Pearson, L. S. (1979). British Ability Scales. Windsor, England: NFER-Nelson.

Enhance. (2012, October). *Quality of Child Outcomes Data: District Experiences, State Support, and National Findings.* Presentation at the Division for Early Childhood Annual International Conference on Young Children with Special Needs and Their Families, Minneapolis, MN. Retrieved from http://enhance.sri.com/downloads/POL_235_QualityOutcomesDataSlides.pdf

Enhance. (2013). Project overview [Web page]. Retrieved from Enhance website, http://enhance.sri.com/project/overview.html

Fallon, M. A. (1994). The effectiveness of sensory integration activities on language processing in preschoolers who are sensory and language impaired. *Infant-Toddler Intervention: The Transdisciplinary Journal, 4*(3), 235–243.

Farmer-Dougan, V., & Kaszuba, T. (1999). Reliability and validity of play-based observations: Relationship between PLAY behaviour observation system and standardised measures of cognitive and social skills. *Educational Psychology*, *19*(4), 429–440.

Frisbie, D. (2005). Measurement 101: Some fundamentals revisited. *Educational Measurement: Issues and Practice, 24*(3), 21–28.

Geers, A. E. (2002). Factors affecting the development of speech, language, and literacy in children with early cochlear implantation. *Language, Speech, and Hearing Services in Schools, 33*(3), 172–183.

Gesell, A. (1928). *Infancy and human growth*. New York, NY: Macmillan.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes. *American Psychologist, 18,* 519–521.

Grantham-McGregor, S., Powell, C., Walker, S., Chang, S., & Fletcher, P. (1994). The long-term follow-up of severely malnourished children who participated in an intervention program. *Child Development, 65*(2), 428–439.

He, W., & Wolfe, E. W. (2012). Treatment of non-administered items on individually administered intelligence tests. *Educational and Psychological Measurement, 72*, 808–826.

Hernandez-Reif, M., Field, T., Largie, S., Mora, D., Bornstein, J., & Waldman, R. (2006). Children with Down syndrome improved motor functioning and muscle tone following massage therapy. *Early Child Development and Care, 176*, 395–410.

Hill, J. L., Brooks-Gunn, J., & Waldfogel, J. (2003). Sustained effects of high participation in an early intervention for low-birth-weight premature infants. *Developmental Psychology, 39*(4), 730–744.

Hundert, J., Mahoney, B., Mundy, F., & Vernon, M. L. (1998). A descriptive analysis of developmental and social gains of children with severe disabilities in segregated and inclusive preschools in Southern Ontario. *Early Childhood Research Quarterly, 13*(1), 49–65.

Individuals with Disabilities Education Improvement Act of 2004, Pub. L. No. 108-446, § 631, 118 Stat. 2744 (2004).

Johnson, L. J., Cook, M. J., & Kullman, A. J. (1992). An examination of the concurrent validity of the Battelle Developmental Inventory as compared with the Vineland Adaptive Scales

and the Bayley Scales of Infant Development. *Journal of Early Intervention, 16*(4), 353–359.

Kaufman, A. S., & Kaufman, N. L. (1983). Kaufman Assessment Battery for Children. San Antonio, TX: Pearson.

Kaufman, A. S., & Kaufman, N. L. (2004). Kaufman Assessment Battery for Children-II. San Antonio, TX: Pearson.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*, 773–795.

Kouri, T. A. (2005). Lexical training through modeling and elicitation procedures with late talkers who have specific language impairment and developmental delays. *Journal of Speech, Language, and Hearing Research, 48*(1), 157–171.

Kumin, L., Von Hagel, K. C., & Bahr, D. C. (2001). An effective oral motor intervention protocol for infants and toddlers with low muscle tone. *Infant-Toddler Intervention: The Transdiciplinary Journal, 11*, 181–200.

Ledbetter, M., Betts, J., Boney, T., & Custer, M. (2013, February). *Evaluating growth trajectories and measurement invariance for early developmental domains.* Paper presented at the annual meeting of the National Association of School Psychologists, Seattle, Washington.

Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996). Standard setting: A bookmark approach. In D. R. Green (Chair), *IRT-based standard-setting procedures utilizing behavioral anchoring*. Symposium presented at the Council of Chief State School Officers National Conference on Large Scale Assessment, Phoenix, AZ.

Linacre, J. M. (2004). Optimizing rating scale category effectiveness. In E. V. Smith, Jr. & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory, models and applications* (pp. 258–278). Maple Grove, MN: JAM Press.

Linacre, J. M. (2010). Transitional categories and usefully disordered thresholds. Online Educational Research Journal. Retrieved from http://www.oerj.org/

Linacre, J. M. (2012). WINSTEPS (Version 3.74) [Computer software and manual]. Retrieved from www.winsteps.com

Mahoney, G., & Perales, F. (2003). Using relationship-focused intervention to enhance the social-emotional functioning of young children with autism spectrum disorders. *Topics in Early Childhood Special Education, 23*(2), 77–89.

Mardell, C., & Goldenberg, D. S. (2011). Developmental Indicators for the Assessment of Learning (4th ed.): *Examiner's Manual*. Circle Pines, MN: Pearson Assessments.

Markowitz, J., & Larson, J. C. (1988). *A longitudinal study of children in preschool special education programs* (Rep. No. BBB16912). Rockville, MD: Montgomery County Public Schools.

Masse, L. N., & Barnett, W. S. (2002). *A benefit cost analysis of the Abecedarian Early Childhood Intervention*. Brunswick, NJ: National Institute for Early Education Research. (ERIC Document Reproduction Service No. ED479989).

Matson, J. L., Hess, J., Sipes, M., & Horovitz, M. (2010). Developmental profiles from the Battelle Developmental Inventory: A comparison of toddlers diagnosed with Down syndrome, global developmental delay, and premature birth. *Developmental Neurorehabilitation, 13*(4), 234–238.

McGrew, K. S., Schrank, F. A., & Woodcock, R. W. (2007). Woodcock-Johnson III Normative Update: *Technical Manual.* Rolling Meadows, IL: Riverside Publishing.

McGrew, K. S., Werder, J. K., & Woodcock, R. W. (1991). Woodcock-Johnson–Revised*: Technical manual.* Rolling Meadows, IL: Riverside Publishing.

McLean, M., McCormick, K., & Baird, S. M. (1991). Concurrent validity of the Griffiths Mental Development Scales with a population of children under 24 months. *Journal of Early Intervention, 15*(4), 338–344.

McLinden, S. E. (1989). An evaluation of the Battelle Developmental Inventory for determining special education eligibility. *Journal of Psychoeducational Assessment, 7*, 66–73.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13–103). New York, NY: American Council on Education and Macmillan.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*(4), 5–8.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249–281). Mahwah, NJ: Lawrence Erlbaum.

Mullen, E. M. (1995). *Mullen Scales of Early Learning: AGS Edition, Manual*. Circle Pines, MN: American Guidance Service.

Newborg, J. (2005a). *Battelle Developmental Inventory* (2nd ed.). Rolling Meadows, IL: Riverside Publishing.

Newborg, J. (2005b). *Battelle Developmental Inventory* (2nd ed.): *Examiner's manual.* Rolling Meadows, IL: Riverside Publishing.

Oser, C., & Cohen, J. (2003). *Improving early intervention: Using what we know about infants and toddlers with disabilities to reauthorize Part C of IDEA.* Washington, DC: Zero to Three Policy Center.

Pomplun, M., & Custer, M. (2004). The equivalence of three data collection methods with field test data: A FACETS application. *Journal of Applied Measurement, 5*(3), 120–132.

Raju, N. S., Price, L. R., Oshima, T. C., & Nering, M. L. (2006). Standardized conditional *SEM*: A case for conditional reliability. *Applied Psychological Measurement, 30*, 1-12)

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology, 25*, 111–164.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Chicago, IL: University of Chicago Press.

Reynolds, A. J., Temple, J. A., Robertson, D. L., & Mann, E. A. (2002). *Age 21 cost-benefit analysis of the Title I Chicago Child-Parent Centers* (Discussion paper). Madison, WI: University of Wisconsin Institute for Research on Poverty.

Roid, G. H., & Miller, L. J. (1997). Leiter International Performance Test-Revised. Wood Dale, IL: Stoelting.

Rynders, J. E., & Horrobin, J. M. (1990). Always trainable? Never educable? Updating educational expectations concerning children with Down syndrome. *American Journal on Mental Retardation, 95*(1), 77–83.

Sayers, L. K., Cowden, J. E., Newton, M., & Warren, B. (1996). Qualitative analysis of a pediatric strength intervention on the developmental stepping movements of infants with Down syndrome. *Adapted Physical Activity Quarterly, 13*(3), 247–268.

Saylor, C. F., Boyce, G. C., Peagler, S. M., & Callahan, S. A. (2000). Brief report: Cautions against using the Stanford-Binet IV to classify high-risk preschoolers. *Journal of Pediatric Psychology, 25*(3), 179–183.

Scarborough, A. A., Spiker, D., Mallik, S., Hebbeler, K. M., Bailey, D. B., & Simeonsson, R. J. (2004). A national look at children and families entering early intervention. *Exceptional Children, 70*(4), 469–483.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461–464.

Scientific Advisory Committee (1995, September). Instrument review criteria. *Medical Outcomes Trust Bulletin*.

Shackelford, J. (2006). *State and jurisdictional eligibility definitions for infants and toddlers with disabilities under IDEA* (NECTAC Notes No. 21). Chapel Hill, NC: University of North Carolina, FPG Child Development Institute, National Early Childhood Technical Assistance Center.

Sheinkopf, S. J., & Siegel, B. (1998). Home-based behavioral treatment of young children with autism. *Journal of Autism and Developmental Disorders, 28*(8), 15–23.

Shen, L. (1997, March). *Quantifying item dependency using Fisher's Z.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Sipes, M., Matson, J. L., & Turygin, N. (2011), The use of the Battelle Developmental Inventory, Second Edition (BDI-2) as an early screener for autism spectrum disorders. *Developmental Neurorehabilitation, 14*(5), 310–314.

Smith, E. V., Jr. (2004). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. In E. V. Smith, Jr. & R. M.

Smith (Eds.), *Introduction to Rasch measurement: Theory, models and applications* (pp. 575–600). Maple Grove, MN: JAM Press.

Smith, R. (1996). Polytomous mean-square fit statistics. *Rasch Measurement Transactions, 10*(3), 516–517.

Snyder, P., & Lawson, S. (1993). Evaluating the psychometric integrity of instruments used in early intervention research: The Battelle Developmental Inventory. *Topics in Early Childhood Special Education, 13*(2), 216–232.

Sparling, J., & Lewis, I. (1991). Partners: A curriculum to help premature, low birthweight infants get off to a good start. *Topics in Early Childhood Special Education, 11*(1), 36–55.

Thomas, R. M. (1979). *Comparing theories of child development.* Belmont, CA: Wadsworth.

Tingey, C. (1991). Developmental attainment of infants and young children with Down syndrome. *International Journal of Disability, Development and Education, 38*(1), 15–26.

Van Den Wymelenberg, K., Deitz, J. C., Wendel, S., & Kartin, D. (2006). Early intervention service eligibility: Implications of using the Peabody Developmental Motor Scales. *Journal of Occupational Therapy, 60*(3), 327–332.

Wasik, B. H., Ramey, C. T., Bryant, D. M., & Sparling, J. J. (1990). A longitudinal study of two early intervention strategies: Project CARE. *Child Development, 61*(6), 1682–1696.

Whitmore, E., Ford, M., & Sack, W. H. (2003). Effectiveness of day treatment with proctor care for young children: A four-year follow-up. *Journal of Community Psychology, 31*(5), 459–468.

Wiersma, W., & Jurs, S. G. (1990). *Educational measurement and testing* (2nd ed.). Boston, MA: Allyn and Bacon.

Woodcock, R. W. (1973). Woodcock Reading Mastery Tests. Circle Pines, MN: American Guidance Service.

Woodcock, R. W. (1978). *Development and standardization of the Woodcock-Johnson Psycho-educational Battery.* Rolling Meadows, IL: Riverside Publishing.

Woodcock, R. W. (1999). What can Rasch-based scores convey about a person's test performance? In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 105–127). Mahwah, NJ: Lawrence Erlbaum Associates.

Woodcock, R.W., & Dahl, M. N. (1971). *A common scale for the measurement of person ability and test item difficulty* (AGS Paper No. 10). Circle Pines, MN: American Guidance Service.

Woodcock, R. W., & Johnson, M. B. (1977). Woodcock-Johnson Psycho-educational Battery. Rolling Meadows, IL: Riverside Publishing.

Woodcock, R. W., & Johnson, M. B. (1989). Woodcock-Johnson Psycho-educational Battery-Revised. Rolling Meadows, IL: Riverside Publishing.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001, 2007). Woodcock-Johnson III. Rolling Meadows, IL: Riverside Publishing.

Wolfe, E. W., & Smith, E. V., Jr. (2007a). Instrument development tools and activities for measure validation using Rasch models: Part I—Instrument development tools. *Journal of Applied Measurement, 8,* 97–123.

Wolfe, E. W., & Smith, E. V., Jr. (2007b). Instrument development tools and activities for measure validation using Rasch models: Part II—Validation activities. *Journal of Applied Measurement, 8,* 204–233.

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

Wright, B. D., Mead, R. J., & Ludlow, R. H. (1980). *Kidmap: Person-by-item interaction mapping* (Memo #29). Chicago: MESA Press.

Wright, B. D., & Stone, M. S. (1979). *Best test design: Rasch measurement.* Chicago: MESA Press.

Zeece, P. D., & Wang, A. (1998). Effects of the family empowerment and transitioning program on child and family outcomes. *Child Study Journal, 28*(3), 161–178.

APPENDIX A

TABLE XVII
DESCRIPTION OF ANALYSIS RUNS WITH EXAMINEE, ITEM, AND MODEL FIT STATISTICS

Run	Description	Overall Examinee MNSQ Outfit (*SD*)	Overall Item MNSQ Outfit (*SD*)	Δ in -2 LL value (Δ in df)	Critical value of Chi-Square at $p = .01$	AIC	BIC
1	All items, all examinees, 0/1/2 scoring	0.83 (1.60)	2.47 (3.31)			26,785	46,861
2	170 unexpected scores (likely scanning errors) omitted	0.71 (1.13)	0.74 (0.41)	1970.47 (202)	251.7	24,807	44,831
3	19 additional unexpected scores omitted	0.69 (1.07)	0.71 (0.39)	255.72 (19)	36.19	24,551	44,573
4	All items rescored from 0/1/2 to 0/0/1	0.71 (1.42)	0.93 (0.80)	8291.13 (112)	149.7	16,118	35,548
5	6 additional unexpected scores omitted	0.70 (1.39)	0.74 (0.49)	98.65 (6)	16.81	16,019	35,448
6	3 additional unexpected scores omitted	0.69 (1.36)	0.73 (0.48)	36.94 (3)	11.34	15,982	35,411
7	26 additional unexpected scores omitted	0.63 (1.14)	0.61 (0.38)	304.7 (26)	45.64	15,677	35,103

APPENDIX B

STANDARDIZATION DATA AGREEMENT

THIS AGREEMENT entered into as of January 15, 2014, by and between The Riverside Publishing Company, with its principle place of business at 3800 Golf Road, Suite 200, Rolling Meadows, Illinois 60008-4015 (hereinafter referred to as "Publisher") and Erica LaForte, 369 May Avenue, Glen Ellyn, Illinois 60137 (hereinafter referred to as "Licensee").

WHEREAS, the Publisher is the owner of standardization and validity study data sets for the *Battelle Developmental Inventory—Second Edition ("BDI-2")* (hereinafter referred to as the "Works"); and

WHEREAS, the Licensee wishes to use data sets from the Works (hereinafter referred to as "Data") in a Ph.D. dissertation study to develop Rasch-based interpretations and to use Table 1 and Table 2 (attached) which the Publisher has granted (hereinafter referred to as "Licensed Study").

NOW, THEREFORE, the Publisher and Licensee agree as follows:

1. The Publisher hereby grants the Licensee a nonexclusive, nontransferable license to use the Data solely to perform the Licensed Study ("Licensed Use"). Under no circumstances shall any third party be granted any access whatsoever to the Data without the prior written consent of the Publisher.

2. The License granted herein shall be for a period commencing with the date first stated above and shall terminate on June 30, 2014, whereupon all of the Licensed Uses shall cease.

3. This Agreement shall only become effective if it is executed by the Licensee within thirty (30) days of receipt. Permission for any extension of the Agreement must be secured in writing from the Publisher.

4. It is understood and agreed that any fee is hereby waived.

5. The License granted herein is non-exclusive, and non-transferable to any third party without prior written permission from the Publisher. The Data may not be used for any additional studies or publications without seeking prior written authorization. The Publisher expressly reserves all rights in the Data not herein granted to the Licensee. The Licensee represents and warrants that the Licensed Study and all publications resulting therefrom will not infringe the intellectual property rights or other proprietary rights of any party.

6. Licensee shall provide Publisher with advance and final copies of Licensee's analyses and Licensee's interpretation thereof, published or unpublished; any reproduction of the raw Data requires the prior written consent of the Publisher. Licensee agrees that Publisher shall have the right to use Licensee's analyses with proper references to Licensee, as part of the technical information provided in technical, educational, or marketing materials to support the Works. This does not constitute permission for Publisher to reprint any such studies in their entirety.

7. Any publications resulting from Licensee's analyses shall contain the following notice:

"Standardization data from the *Battelle Developmental Inventory, Second Edition* ("BDI-2"). Copyright © 2004 by The Riverside Publishing Company. All rights reserved. Used with permission of the publisher."

8. Upon expiration of this agreement, the Licensee agrees to return to Publisher any original copies of the Data, which have been provided to Licensee by Publisher or by anyone for any purpose, and to destroy any copies of such Data which Licensee made or had made for any purpose.

9. This Agreement constitutes the entire agreement between the parties. No amendment or modification of this Agreement shall be valid unless executed in writing by both parties. This Agreement shall be governed and interpreted under the laws of the State of Illinois, without reference to its principles of conflicts of law.

IN WITNESS WHEREOF, the parties hereto have caused this Agreement to be executed as of the date first above written.

THE RIVERSIDE PUBLISHING COMPANY

Mark Ledbetter
Vice President, Publisher

Erica M. LaForte

Erica M LaForte

TABLE 1

BDI-2 CHANGE-SENSITIVE SCORE DIFFERENCES AND CORRESPONDING INTERPRETATIONS FOR EXAMINEE LEVEL OF DEVELOPMENT

CSS Difference Score	Relative Developmental Index (RDI) Range	Examinee's Level of Development	Examinee Will Find Test Tasks That Average Same-Age Peers Perform with 90% Success:
+31 and above	100/90	Very Advanced	Extremely Easy
+14 to +30	98/90 to 100/90	Advanced	Very Easy
+7 to +13	95/90 to 98/90	Age- Appropriate to Advanced	Easy
−6 to +6	82/90 to 95/90	Age- Appropriate	Manageable
−13 to −7	67/90 to 82/90	Mildly Delayed to Age- Appropriate	Difficult
−30 to −14	24/90 to 67/90	Mildly Delayed	Very Difficult
−50 to −31	3/90 to 24/90	Moderately Delayed	Extremely Difficult
−51 and below	0/90 to 3/90	Severely Delayed	Virtually Impossible

Note: From *Manual and Checklist. Report Writer for the WJ III* (p. 10), by F.A. Schrank and R.W. Woodcock, 2002, Rolling Meadows, IL: Riverside Publishing. Copyright 2002 by Riverside Publishing. Adapted with permission.

TABLE II
BDI-2 EXPRESSIVE COMMUNICATION SCORES FOR A CHILD REFERRED FOR EVALUATION

Score Metric	24 Months	36 Months	48 Months
Raw Score	20	31	41
Scaled Score (Mean = 10, SD = 3)	3	3	3
Percentile Rank	1	1	1
Change Sensitive Score (CSS)	417	461	491
Relative Developmental Index (RDI)	5/90	11/90	27/90
Level of Development	Moderately Delayed	Moderately Delayed	Mildly Delayed

Note: Adapted from *Evaluating Growth Trajectories and Measurement Invariance for Early Developmental Domains*, by M. Ledbetter, J. Betts, T. Boney, and M. Custer (2013). Presentation at the National Association of School Psychologists conference, April.

VITA

Name:	Erica M. LaForte
Education:	B.S., Psychology and Sociology, University of Wisconsin–Madison, 1995 M.A., Research Methodology, Loyola University of Chicago, 2001 Ph.D., Educational Psychology, University of Illinois at Chicago, 2014
Professional Membership:	National Association of School Psychologists Illinois School Psychologists Association (Former member) American Psychological Association (Former member)
Professional Experience:	Woodcock-Muñoz Foundation (2012 to present) Measurement Learning Consultants, LLC (2012 to present) Senior Manager, Psychometric Services, PSI (2010 to 2012) Senior Psychometrician, Riverside Publishing (2008 to 2010) Psychometrician, Riverside Publishing (2004 to 2008) Consultant, Riverside Publishing (2003 to 2004) Senior Program Manager, Riverside Publishing (2002 to 2003) Project Director, Riverside Publishing (2000 to 2002) Senior Production Editor, Riverside Publishing (1997 to 2000) Production Editor, Riverside Publishing (1996 to 1997)
Publications:	McGrew, K. S., LaForte, E. M., & Schrank, F. A. (In preparation). Technical Manual. *Woodcock-Johnson IV*. Rolling Meadows, IL: Riverside Publishing.