ABSTRACT

Content Validity Evidence for the Verbal Behavior Milestones Assessment and Placement Program

Kristen L. Padilla, Ph.D.

Mentor: Jessica Akers, Ph.D.

Autism spectrum disorder (ASD) affects one in 54 children in the United States and the prevalence has increased by exponentially in the last decade (Centers for Disease Control and Prevention [CDC]). With the rising prevalence, evidence-based treatment is critical for this population. Interventions based in applied behavior analysis (ABA) are the most effective research-based strategies for individuals with ASD (Axelrod, McElrath, & Wine, 2012; Foxx, 2008; Lovaas, 1987). In order to develop optimal treatment plans with accurately identified goals and intervention strategies, individuals must undergo a comprehensive assessment that includes the use of research-based instruments. The Verbal Behavior Milestones Assessment and Placement Program (VB-MAPP) is the most widely used instrument for curriculum development and treatment planning in the field of ABA. However, there is currently no validity or reliability studies to support its widespread use. The purpose of this study is to address this gap in the literature by providing content validity evidence for the VB-MAPP. A national panel of 13 subject matter experts (SMEs) provided an evaluation of the domain relevance, age

appropriateness, method of measurement appropriateness, and domain representation across the three levels of the Milestones Assessment, Early Echoic Skills Assessment (EESA), and the Barriers Assessment. Overall, the content validity evidence for the VB-MAPP Milestones, EESA, and Barriers Assessment was moderate to strong across the evaluated areas although there were areas with limited or conflicting support. The evidence suggests that the scores of the VB-MAPP provide information relevant to the target behaviors of interest but a few domains may not be fully represented by their specific items. When the VB-MAPP is used by itself, researchers and practitioners can have reasonable confidence in the results for many domains but should exercise caution for some domains across levels. That said, it is recommended that the VB-MAPP be used in conjunction with other sources of assessment information, which is recommended for assessment in general. The results of this study could inform revisions to future editions of the VB-MAPP. With some targeted revisions, the VB-MAPP could serve as a comprehensive assessment with strong validity evidence for this developmental age range.

Content Validity Evidence for the Verbal Behavior Milestones Assessment and Placement Program

by

Kristen L. Padilla, B.A., Ed.S.

A Dissertation

Approved by the Department of Educational Psychology

_____

Grant Morgan, Ph.D., Chairperson

Submitted to the Graduate Faculty of
Baylor University in Partial Fulfillment of the
Requirements for the Degree
of

Doctor of Philosophy

Approved by the Dissertation Committee

_____

Jessica S. Akers, Ph.D., Chairperson

_____

Nicholas F. Benson, Ph.D.

_____

Terrill Saxon, Ph.D.

_____

Sara L. Dolan, Ph.D.

_____

John Ferron, Ph.D.

Accepted by the Graduate School

August 2020

_____

J. Larry Lyon, Ph.D., Dean

Page bearing signatures is kept on file in the Graduate School.

TABLE OF CONTENTS

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to acknowledge my dissertation chair, Dr. Jessica Akers. Through such trying times in so many capacities throughout my dissertation, she stood strong and supportive. She is one of the most caring, authentic, and knowledgeable individuals I know. I cannot thank her enough for all she did to get me to the finish line. I would like to also acknowledge my peers – Dr. Regan Weston and soon to be "Dr." Supriya Radhakrishnan – who supported me during this journey. These individuals helped me pursue my research area, inspired me to work harder, and uplifted me when I needed it the most. Lastly, I would like to acknowledge Dr. Grant Morgan whose discussion on psychometric topics inspired me to bridge two academic fields that have not historically intersected as they should have. Because of our collaborative work and discussions, I continued to pursue this area of research despite what others may have advised. Every time we discussed my passion in working with individuals with developmental disabilities and the importance of assessment practices in the field, I knew this was what I truly wanted to do.

DEDICATION

To my children and family who made so many sacrifices to support me on my doctoral journey. They inspire me to always be a better person and make an impact in the world. I strive to make them proud and encourage them to be good people and do good work. I hope my journey - successes, obstacles, and failures - teach them to continue to aim high despite their past, current circumstances, and the expectations of others.

CHAPTER ONE

Introduction

Applied Behavior Analysis (ABA) is a science that seeks to address socially significant problems through the systematic manipulation of environmental variables. The principles of ABA have guided the development of research-based interventions and teaching techniques, which have shown to be very effective in treating individuals diagnosed with autism spectrum disorder (ASD; Axelrod, McElrath, & Wine, 2012; Foxx, 2008; Lovaas, 1987).

ASD is a condition that can affect several areas of a child's development, such as cognitive, social, and adaptive skills. ASD affects 1 in 54 children in the United States, which is a 104% increase over the last decade (Centers for Disease Control and Prevention [CDC], 2016). In light of the rise in prevalence and given the breadth of development areas affected, assessment processes should be comprehensive and address all major areas of human functioning, such as social, motor, language, daily living, play, and academic skills (Gould, Dixon, Najdowski, Smith, & Tarbox, 2011). In order to develop intervention plans that effectively target an individual's skill deficits, researchers and practitioners must utilize assessment practices and instruments that have strong evidence supporting their use. Assessment results should guide the development of a structured treatment program or curriculum that targets crucial skills that are functional across settings (Gould et al. 2011).

History of Assessment in ABA

Assessment of an individual's current level of performance provides information about the individual's present skills and deficits to determine what should be targeted next (Linehan, 1977). Hawkins (1979) specifies five phases of assessment, which are (a) screening, (b) defining and quantifying the problem or criteria for desired achievement, (c) pinpointing target behaviors, (d) monitoring progress, and (e) following up at a predetermined time to assess skill maintenance. Screening serves multiple purposes in the overall assessment phase, such as determining (a) if a client's case is appropriate for the agency, (b) what further assessment is needed, and (c) general information about the problems the individual may be experiencing (Hawkins, 1979). Screening helps evaluators focus subsequent evaluation procedures in order to identify the best formalized assessment that will increase the likelihood of successfully identifying target behaviors for skill acquisition. Screening can be accomplished with interview guides, clinical interviews, and rating scales (Hawkins, 1979).

Behavior analysts use a variety of assessments within their practice to assess an individual's strengths and weaknesses (Gould et al., 2011), to identify the function of an individual's challenging behavior (Iwata, Dorsey, Slifer, Bauman, & Richman, 1982), and to develop goals (Sundberg, 2014), all of which contribute to the remaining four phases outlined by Hawkins (1979). Each of these types of assessments provide information about the current status of the behavior of interest. This assessment information serves as the baseline from which further evaluations of the characteristic or behavior of interest will be compared against to determine the extent of the behavior change.

Several forms of assessment have been used in the field to obtain crucial information to develop treatment plans for individuals with ASD. There are also different purposes of assessment in the field of ABA depending on the problem areas identified during the screening process. Function-based behavioral assessments provide vital information used for addressing challenging behaviors (e.g., experimental functional analyses) (Cooper, Heron, & Heward, 2007) and skill acquisition assessments which aide in the identification of an individual's current skill level and potential areas of growth. Regardless of the purpose, assessments based within the framework of ABA heavily rely on direct observation of an individual's behavior.

Function-based Behavioral Assessment

For purposes of reducing challenging behavior, various types of assessments can be used to identify the function of behavior. Semi-structured and structured interviews have been developed, such as the Motivation Assessment Scale (MAS) (Durand, 1989) and the Functional Assessment Screening Tool (FAST) (Florida Center on Self-Injury, 2000) to determine hypothesized functions of behavior. These can be viewed as screening measures that help focus further function-based behavioral assessments. These indirect measures contain a variety of items that address the client's schedule, topography of the behavior, environmental factors associated with the behavior of interest (e.g., antecedents, consequences, setting events), client preferences, and forms of communication. Scales often use multiple Likert-type items designed to provide information about a variety and/or frequency of behaviors including social, adaptive, communication, and academic skills. Checklists can be used to indicate whether or not target behavior(s) was observed during an observation period. Behavior observation

assessments may also include antecedent-behavior-consequence recording or the use of scatterplots (Cooper, Heron, & Howard, 2007).

In 1982, Iwata and colleagues developed a standardized assessment process that included systematic procedures to determine functions of behavior, termed the experimental functional analysis. Using these procedures, the researchers established functional relationships between self-injury and specific environmental events. Functional analysis has been repeatedly used in research and practice to determine behavior functions in order to develop effective treatment plans to reduce challenging behaviors. Until the late 1990s, experimental functional analysis was the most widely used standardized form of assessment throughout the field of behavior analysis that focused on identifying functions of behavior. Since then, there are a variety of ABA-based assessments that have been developed to identify behavioral deficits to target for skill acquisition. ABA-based skill acquisition assessments, which are typically described as criterion-referenced instruments, are a relatively new phenomenon when compared to other fields, such as psychology. Several ABA-based criterion-referenced assessments will be presented in Chapter Two. The following section will address the assessment characteristics that distinguish norm- and criterion-referenced assessments.

General Types of Assessment

There are two major types of assessments that are generally used to determine an individual's level of performance that are not specific to ABA: norm-referenced and criterion-referenced assessments. Both are considered standardized forms of assessments because they have standardized administration procedures, scoring criteria, and score

interpretation that allows for comparison against normed data and/or predetermined specific criteria (AERA et al., 2014).

Norm-referenced assessments. Norm-referenced assessments are based on normed data and compare an individual's skill set to the performance of others within a relevant population (i.e., norm group) (Anastasi, 1988). The normed data is typically a nationally representative sample of the target population of similar age and/or grade. These types of assessments have standardized administration procedures, scoring criteria, and score interpretation guidelines. Raw scores from the assessment are commonly converted to standard scores, national percentile ranks, or age equivalents. These scores provide information that can be used to determine the level of functioning of an individual compared with his or her peers. Many norm-referenced assessments are scaled to have a normative population mean of 100 with a standard deviation of ± 15; the mean is also located at the 50^{th} percentile (Cicchetti, 1994). Norm-referenced assessments include those that are commonly used to measure constructs, such as cognitive ability, using multiple items that provide a general overview of the individual's current level of performance. When using norm-referenced assessments, the obtained score is an estimate of the individual's *relative* position on an underlying continuum, or amount of the construct being assessed. Confidence-like intervals, also referred to as error bands, can be constructed around the observed score to gain insights into the range within which the score might be if the instrument were to be administered again (Cichetti, 1994; Shultz, Whitney, & Zickar, 2014). Confidence-like intervals account for variability in scores for various reasons, such as error.

When carefully developed and validated, norm-referenced assessments provide a great deal of information of an individual's global performance, ability, proficiency or some similar type of outcome relative to other people in the target population (i.e. norm group). Although, norm-referenced assessments provide information regarding the relative strengths and weaknesses of individuals in comparison to their peers, they are often insensitive to instruction or other types of specific interventions. That said, they do not provide an estimate of the absolute level of performance achieved (Bond, 1996). This differs from assessments based in ABA, which rely on direct observation of specific behaviors of individuals.

Criterion-referenced assessments. Criterion-referenced instruments are used to determine an individual's performance by comparing it to a predetermined criterion or standard for the purpose of making a decision or classification (e.g., skill level, mastery, proficiency, certification). These types of assessments make no direct reference to the performance of other examinees. Criterion-referenced instruments either indicate the likely proportion of correct responses that would be obtained on some larger domain of similar items or indicate that an examinee's level of tested skill is adequate to perform successfully in some other setting (AERA et al., 2014). There have been a variety of criterion-referenced assessments developed within the ABA framework to identify an individual's strengths and weaknesses to develop skill acquisition programs. According to Crocker and Algina (2006), the commonly used term criterion-referenced measure is actually a substitute for the more cumbersome term criterion-behavior-referenced measurement, which implies that measurements are to be interpreted in terms of the criterion behaviors an individual can exhibit. This type of measurement is well-suited to

meet the needs of behavior analysts. Criterion-referenced assessments have made a huge impact on how behavior analysts identify target behaviors, develop treatment plans, and monitor progress to enhance an individual's skill repertoire (Padilla, 2019). Over the last 30 years, several instruments have been developed within the ABA framework that specifically target skill acquisition for individuals diagnosed with ASD. Assessments commonly used include the Verbal Behavior Milestones Assessment and Placement Program (VB-MAPP; Sundberg 2014), Assessment of Basic Language Learning Skills-Revised (ABLLS-R; Partington, 2006), and the Promoting the Emergence of Advanced Knowledge Relational Training System (PEAK; Dixon 2014) with the VB-MAPP being the most common (Austin & Thomas, 2017; Padilla, 2019). The VB-MAPP and PEAK were both initially developed in 2008 whereas the ABLLS was initially published in 1998. According to its manual, the VB-MAPP is described as a criterion-referenced assessment, curriculum guide, and progress-monitoring tool designed for parents and professionals to gain information regarding their child's language and social skills for individuals aged 0-48 months. The VB-MAPP and other ABA-based skill acquisition assessments will be further discussed in Chapter Two.

Educational and psychological testing and assessment are among the most important contributions of cognitive and behavioral sciences to our society because they provide fundamental and significant sources of information about individuals and groups. Instruments can provide information that results in better decisions about type or level of skill, knowledge, or behavior (AERA et al., 2014), which can in turn lead to more specific and sensitive interventions to align with the assessment results. An instrument can be thought of as a standardized procedure for sampling behavior (Hubley, & Zumbo,

2013) similar to what Iwata developed for the use of functional analyses. That is, to assess the behavior of interest both forms of assessment have procedures that are systematically and consistently implemented to obtain a measure that best reflects the individual's performance. Use of standardized instruments for educational and psychological measurement has been common practice for over a century and has led to the development of a variety of intelligence, achievement, social, emotional, and behavioral measurements. To ensure professionals in the field of ABA are being good test consumers and users, they need to understand fundamental psychometric concepts as they relate to test development, administration, and interpretation.

Validity

According to the *Standards for Educational and Psychological Testing* (AERA et al., 2014), there are certain criteria that must be met when developing, administering, and interpreting results from instruments. Developers of standardized instruments in ABA may consider adopting these standards to ensure instruments meet all requirements and have validity and reliability evidence to support an instrument's use. Developing an instrument involves defining a target construct or behavior (e.g., fifth grade math achievement, spelling, receptive language) for measurement and designing a measure related to knowledge, skills, abilities, attitudes, or characteristics of the target behavior. In doing so, the developer must first consider the intended use and expected interpretations of the scores, which will help specify, if not determine, the content and format of the instrument. An instrument's validity and reliability evidence are critically important issues to address during the phases of test design and development (AERA et al., 2014).

Validity Theory

The current, unitary conceptualization holds that validity is the degree to which evidence and theory support the interpretations of test scores and inferences/decisions made based on those scores within the context of an instrument's intended use (Benson, 1998; Loevinger, 1957; Messick, 1989b). Such inferences should be based on theoretical and empirical data from multiple sources that align with the conclusions (Shultz, Whitney, & Zickar, 2014). Assessment guidelines related to validity provided in the *Standards for Educational and Psychological Testing* (AERA et al., 2014) address the need for each instrument's development to include clear expectations for its intended use, the target population for which the instrument was designed, and a rationale for the intended interpretation of the instrument's scores.

Several dimensions of ABA align with the principles of validity. Both frameworks have a basis in scientific theory that guide the provision of research and practice in the respective fields. Validity, again, refers to the accuracy of inferences made about a construct based on data collected using an instrument. As such, validity is a property of inferences, not an instrument. The accuracy of inferences is rarely, if ever, known by the researcher so it is imperative to collect evidence that supports the use of an instrument in a particular way or context. There are several types of evidence that are useful in supporting the validity of the proposed interpretation of test scores for a particular use (AERA et al., 2014). They are construct-, criterion-, and content-related validity (Benson, 1998; Cronbach & Meehl, 1995), each of which is discussed in more detail next.

Construct-related validity. A construct is the concept, attribute, or variable that is the target of measurement (Haynes, Richard, & Kubany, 1995). The construct is guided by a theoretical framework and requires consensus on the operational definition of the construct in order for it to be measured more effectively (Shultz, Whitney, & Zickar, 2014). Construct-related validity refers to degree to which a score can be interpreted as representing the intended underlying construct. Many forms of validity have been conceptualized under the overarching framework of "construct validity" (Benson, 1998; Cronbach & Meehl, 1995; Loevinger, 1957). This approach signifies the argument that an instrument's scores are only useful if they reflect the construct of interest and evidence is collected to support this relationship (Cook & Beckman, 2006). For example, many skills-based criterion-referenced assessments used in the field of ABA are based on Skinner's analysis of verbal behavior, which include constructs such as manding and tacting. The meaning and definition of manding, for instance, are based on Skinner's (1957) theoretical framework governing verbal behavior, which is based on the premise that language is a learned behavior that is reinforced through the mediation of other persons (Sundberg, & Michael, 2001). In keeping with Skinner's theoretical framework, an instrument designed to reflect one's manding skills should have strong evidence that the results/scores support decisions about manding and that the instrument's use is also supported by evidence from other studies based on the same theoretical framework. This process includes empirically testing the relationships that involve a particular construct and interpreting the results to determine if the instrument validates the construct as defined by theory (Higgins & Straub, 2006). Assessments should include information about the validation process for the instrument and the accompanying literature to support

the validity of the assessment's intended use. This information can usually be found in the technical manual for the instrument or in sections that discuss the standardization of the instrument.

Criterion-related validity. To support criterion-related validity for an assessment, an expected relationship between the assessment results and some external, pre-validated criterion must exist (Shultz, Whitney, & Zickar, 2014). Evidence for criterion-related validity is typically the degree to which the correlation between the instrument being validated and an external, validated criterion aligns with theoretical expectation (Kane, 2006). The external criterion could be another instrument, a series of tasks, a future outcome, or some other theoretically supported criterion.

Criterion-related validity evidence can be divided along two dimensions: convergent versus discriminant, which differ based on the theoretically expected relationship with the focal instrument and the criterion, and concurrent versus predictive, which differ based on when the external criterion is observed relative to the focal instrument (Campbell & Fiske, 1959). For convergent validity evidence, at least a moderately strong positive correlation between the focal instrument and the external criterion is expected whereas discriminant validity indicates that a zero or negative correlation between the focal instrument and the external criterion is expected. Concurrent validity refers to the extent to which an instrument correlates with another instrument measuring the same, or similar construct (i.e., external criterion) administered at or near the same time (Cicchetti, 1994). Here, assessment results are compared against some other preexisting validated instrument that measures the same construct to provide evidence that the assessment measures the specific construct. For example, to determine

the use of a newly developed instrument that assesses adaptive behavior skills, a test user must carefully review the studies used to compare the new instrument to another instrument that measures the same construct and has validity and reliability evidence. For example, the Autism Diagnostic Observation Schedule, 2nd edition (ADOS-2) (Lord et al., 2012) is considered to be the gold standard in the diagnosis of ASD. Many instruments, such as the Childhood Autism Rating Scales, 2nd edition (Schopler, Van Bourgondien, Wellman, & Love, 2010), are measured against the ADOS-2 through empirical validation to determine if there is a positive correlation between the two instruments. If a positive correlation is observed, there is evidence to support the valid use and interpretation of scores to diagnose ASD using the instrument. Predictive validity evidence provides information about whether a score predicts a future criterion, such as behaviors, outcomes, or instrument scores (Johnson & Morgan, 2014). For example, in education, standardized cognitive measures are often used to predict performance in other academic areas.

Content-related validity. In light of the focus of the study to be described in Chapter Three, content validity is described in more detail. Generally, content-related validity is the extent to which assessment items sufficiently define and represent the construct they are designed to reflect (Shultz et al., 2014). Linehan (1980) stated that the types of inferences that are made in behavioral assessments "necessitate attention to content validity" (p. 152). In fact, she argued the need for content validity in most instances of behavior assessment. That said, the concept of content validity has been controversial over the past 100 years with respect to its formulation and necessity to validity argumentation (Sireci, 1998). Early views of validity focused primarily on what

was described above as criterion-related validity; correlation coefficients between an assessment and an external criterion were used as indexes of whether the assessment measured what it purported to measure (Sireci, 1998). Many scholars acknowledged limitations or shortcomings of the early conceptualization of validity, which led to academic debate over fundamental components of assessment, such as face validity. Rulon (1946), Mosier (1947), and Gullikson (1950) provided the first writings that laid the groundwork for content validity, but Gullikson (1950) was the first to emphasize evaluating test content using subject matter experts (SMEs) as a source of empirical support. Within his overview of the history of the controversy and conceptualization of content validity, Sireci (1998) noted three common components of content validity originated with the writings of Rulon (1946), Mosier (1947), and Gullikson (1950), which were domain definition, domain representation, and domain relevance. It should be noted that other prominent scholars have also contributed to these areas, such as Messick's (1975, 1980, 1989a) work related to domain relevance and definition and Nunnally's (1967) work related to domain representation. A fourth component was identified from the work of Loevinger (1957), Ebel (1956), Nunnally (1967), Cronbach (1971) and Fitzpatrick (1983) that related to the appropriateness of the test development process. The combined efforts of these and other scholars over the decades informed the standards that are used in educational and psychological testing and measurement. Each of the four components are discussed below.

Domain definition refers to how the construct being measured is operationally defined (Sireci & Faulkner-Bond, 2014). In other words, the domain definition provides details about the specific aspects of the construct measured by an instrument and

transforms the construct from theory into a concrete content domain (Sireci & Faulkner-Bond, 2014). Providing evidence for domain definition involves SMEs evaluating the congruence between the definition and the SMEs common understanding of the construct (Sireci, 1998)).

Domain representation refers to the degree to which the items on an instrument adequately represent and reflect the target domain. In providing support for domain representation, SMEs typically review and rate how adequately and/or fully items represent the target domain (Sireci, 1998; Sireci & Faulkner-Bond, 2014). For example, in the context of verbal behavior, instruments should contain items that measure the aspects of verbal behavior, such as mands, tacts, interverbals, and echoics. To illustrate this example further, items used to determine if an individual has appropriate manding skills should ideally be operationally defined and related to the construct of verbal behavior. As Linehan (1980) noted, behavioral assessors typically make inferences based on responses observed during a testing situation. Moreover, assessors who observe a child's behavior in, say, a structured setting for 30 minutes in one day will likely use the observations or scores to infer the child's behavior at another point in time when the assessor is not present or under different stimulus conditions (Coleman, Whitman, & Johnson, 1979; Linehan, 1980).

Domain relevance refers to the degree of importance or relevance each item has to its target domain. It should be noted that an item with high domain representation may or may not have high domain relevance depending on the content (Sireci & Faulkner-Bond, 2014). Therefore, domain representation and domain relevance are different but ideally related components of content validity. For domain relevance, SMEs commonly review

and rate the relevance of each item as it relates to the target domain. Domain relevance is important at the item-level, and domain representation can be aggregate up from the item-level to provide a more robust evidence of domain representation as a whole. When used together, evidence of domain relevance and representation informs whether an instrument includes items that address important and relevant aspects of a domain (Sireci & Faulkner-Bond, 2014).

The appropriateness of the test development process refers to how faithfully and fully processes were in creating instruments for measuring intended constructs (Sireci & Faulkner-Bond, 2014). Evidence for the appropriateness of the test development process can take multiple forms, such as SMEs reviewing items for technical accuracy, high quality item-writing, sensitivity review, pilot testing and statistical item analysis, and statistical evidence of differential item functioning (Haladyna, 2004; Sireci, & Faulkner-Bond, 2014; Waugh & Gronlund, 2012).

Methods of Content Validity Studies

As indicated for of the four areas of content validity, SMEs clearly play a critical role in providing evidence supporting an instrument's content validity. The methods for providing content validity evidence can be divided into traditional and alignment methods. The traditional content validity studies require SMEs to (a) match items to their intended construct, (b) rate the domain representation of items, and/or (c) rate the domain relevance of items. Using the matching approach, SMEs are provided a list of items along with defined areas of a test (e.g., cognitive levels, content areas). SMEs then assign (i.e., match) each item to areas of the test based on the perceived congruence between the two. Rating methods tend to be used to collect information related to how well the items

measure the targeted domain. For the rating method, SMEs review and rate each item on its relevance and/or representation using a Likert-type response scale. Taken together, data collected using the rating method can provide support for how well items, individually and collectively, reflect the intended construct. Data analyses for traditional methods can include descriptive or inferential statistical analysis. Descriptive analysis includes determining the relative frequency of SME ratings and comparing against published criteria. For example, Popham (1992) suggested a criterion of 70% of SMEs rating an item congruent with its standard. Multiple indexes have also been proposed that can be used descriptively or inferentially to evaluate item-construct congruence. These include Lawsche's (1975) content validity ratio (CVR), Rovinelli's and Hambleton's (1977) item-objective congruence index (*I*), and Aiken's (1980) content validity index (*V*). Hypothesis tests are available for each of these indexes, which allows researchers to evaluate chance observations.

There are also multiple methods for conducting alignment content validity studies. In general, the alignment approach involves several levels of test specifications and, at times, instruction. Alignment methods tend to evaluate item content at a more granular level by focusing on depth, breadth, or cognitive complexity of items with respect to content standards in specific subject areas. Sireci and Faulkner-Bond (2014) noted that alignment methods emerged from state-level educational achievement testing in the United States. The common alignment-based methods include Webb (Webb, 1997), Achieve (Rothman, Slattery, Vranek, & Resnick, 2002), and Surveys of Enacted Curriculum (SEC; CCSSO SEC Collaborative Project, 2005; Porter & Smithson, 2001). Each of the alignment methods requires (a) a clearly articulated set of content standards

against which to evaluate a set of test items and (b) convening a panel of SMEs with expertise in the area(s) relevant to the testing purpose. The tasks in which the SMEs engage is determined by the specific alignment method, but all begin with training session to familiarize the SMEs with both the content standards and the test (Sireci & Faulkner-Bond, 2014). In line with its development, the alignment method is typically used for academic achievement tests, such as end-of-grade or standardized state testing programs.

Conclusion

In conclusion, validity and related types of evidence are critical to the development, use, and interpretation of assessments across all scientific fields. Validity evidence supporting instruments' use is necessary to ensure researchers and practitioners are basing their decisions on data collected from high quality assessments. In light of the increasing prevalence of developmental disabilities, in particular ASD, coupled with the rising popularity of applied behavior analysis and related assessments, there is a growing need to collect and evaluate the validity evidence of these assessments (Gould et al., 2011).

CHAPTER TWO

Literature Review

Introduction

Sundberg (2014) noted that the primary purpose of assessment is to identify the baseline level of a child's skills and to compare it to those of their typically developing peers. Assessment is the process of compiling information about an individual's present levels of functioning that will determine the trajectory of their treatment plan. As previously discussed, there have been a variety of forms of assessment to examine and predict human behavior for the purposes of behavior reduction or skill acquisition. The primary focus of this paper will be on skill acquisition assessments that have been developed within the ABA framework.

Use of Standardized Instruments in ABA

Despite the critical role assessment plays in the diagnosis, treatment planning, and progress monitoring for individuals with ASD, there is minimal research about the types of instruments used with this population for these purposes (Luiselli, et al., 2001). Luiselli et al. (2001) surveyed 113 treatment centers in the United States that served children with ASD regarding their use of standardized instruments and purposes of assessment practices. The majority of identified assessments were used to evaluate intelligence, motor skills, and language/communication and were primarily used for diagnostic and curriculum design. The most commonly reported instrument being used was the Vineland Adaptive Behavior Scales (Sparrow, 1994) with 60.6% respondents

reporting its use for screening (15.6%), diagnosis (22.8%), curriculum design (16.3%), and semiannual/annual evaluations (23.2%). Forty-five percent of respondents reported using the Family Needs Survey (FNS) for screening (42.23%), diagnosis (40%), curriculum design (37.5%), and semiannual/annual evaluations (72.1%). The third most commonly reported assessment was the Transdisciplinary Play Assessment with 42.8% of respondents reporting its use for purposes of screening (24.1%), diagnosis (52.7%), curriculum design (28.5%), and semiannual evaluations (20%). Between 75-80% of treatment centers used instruments within the adaptive behavior, curriculum/education, intellectual, and language/communication domains (Luiselli et al., 2001). Although these types of assessments provide a great deal of information about an individual's overall skill set, the results, however, reflect a more generalized behavior repertoire (Hawkins, 1979). Typically, those instruments include very few items to assess specific skills that might change in a desired direction as the result of a specifically-developed curriculum. Instruments that are based on operationally defined target behaviors are most useful for behavior analysts because these instruments measure changes in specific behaviors of interest as opposed to overall change in variables assumed to be proxies for hypothetical constructs (Gould, et al., 2011).

Austin and Thomas (2017) conducted a small-scale survey with 99 participants in Washington regarding their clinical assessment practices for diagnostic and educational programming. This was the first study to focus on practices specific to professionals practicing in behavior analysis. The participants had the following credentials: Board Certified Behavior Analyst (BCBA; 55%), Board Certified Behavior Analysts with a doctoral designation (BCBA-D; 26%), Licensed Behavior Analyst (LBA; 6%), Board

Certified Assistant Behavior Analyst (BCaBA; 6%), Counselor (6%), Psychologist (5%), None (4%), or Other (20%). About 19% of the participants were faculty members, 37% were administrators or supervisors, 32% were practitioners and 8% were trainers/coaches. Slightly more than half (56%) of respondents reported using the Verbal Behavior Milestones Assessment and Placement and Program (VB-MAPP; Sundberg, 2014) and 41% reported using the Assessment of Basic Language Learning Skills-Revised (ABLLS-R; Partington, 2006) for programming. Seven percent reported using the Vineland Adaptive Behavior Scales, Second Edition (Vineland-II; Sparrow, Cicchetti, & Balla, 2005) for diagnostic purposes. In general, 24% of respondents reported selecting instruments based on client need whereas 10% reported using the assessments as required by their organization (Austin & Thomas, 2017). Other studies have also shown that the Promoting the Emergence of Advanced Knowledge Relational Training System (PEAK), a criterion-referenced assessment, is used for curriculum planning for individuals with ASD (Dixon, Belisle, Stanley, Rowsey, Daar, & Szekely, 2015).

Padilla (2019) expanded this area of research by surveying 1,428 individuals who primarily practice in ABA throughout the world. According to the survey results, the most widely used assessment for educational and curriculum programming is the VB-MAPP with 76% (n = 1,086) of the respondents reporting its use by itself or in addition to another assessment. The prevalent use of the VB-MAPP was reported across practitioners in varying professional positions and certification levels. Approximately 45% (n = 638) of the sample reported using the ABLLS-R, 34% (n = 485) reported using the Vineland Adaptive Behavior Scales, and roughly 14% (n = 197) reported using the PEAK. It should be noted that respondents could select multiple assessments and provide

written responses for assessments not included in the survey. The three most commonly reported assessment practices were using VB-MAPP *and* ABLLS-R (*n* = 228, 16%), VB-MAPP *only* (*n* = 199, 14%), and VB-MAPP, ABLLS-R *and* Vineland (*n* = 148, 10%). Thirty-five percent of respondents indicated that they used assessments other than or in addition to the four included on the survey via written comments. Respondents provided 115 different forms of assessments used in the written comments. Among written responses, the most commonly reported assessments were the Assessment of Functional Living Skills (AFLS) (*n* = 172, 12%) and the Essential for Living (EFL) (*n* = 72, 5%).

Assessment Descriptions

The VB-MAPP is based on Skinner's analysis of verbal behavior and the science of ABA. The instrument includes five components: (a) the Milestones assessment, which is designed to provide a representative sample of a child's existing verbal and related skills across three development age levels; (b) the Barriers Assessment, which considers both common learning and language acquisition barriers faced by children with ASD or other developmental disabilities; (c) the Transition assessment, which provides a measurable way for an individualized education program (IEP) team to make decisions regarding the child's placement in a less restrictive educational environment; (d) the Task Analysis and Supporting Skills, which provides an even further breakdown of the skills mentioned previously; and (e) the Curriculum Placement and IEP Goals, which helps the individual designing the program develop an all-inclusive intervention plan.

The ABLLS-R is described as a criterion-referenced assessment that provides a comprehensive review of 544 skills from 25 skill areas (i.e., repertoires) including language, social interaction, self-help, academic, and motor skills. The ABLLS-R is

based on Skinner's analysis of *Verbal Behavior* (1957). According to the assessment developers, the ABLLS-R contains limited empirical support for its psychometric properties despite its widespread use (Partington, Bailey, & Partington, 2018).

The PEAK consists of four modules, each of which contains 184 programs that are hierarchically ordered by complexity. The modules include direct training, generalization, equivalence and transformation focusing on verbal relations ranging from simple vocalizations to complex language in accordance with Skinner's analysis of verbal behavior. The modules use different methodologies including discrete-trial training, stimulus equivalence, and relational frame theory (Dixon, Rowsey, Gunnarsson, Belisle, Stanley, & Daar, 2017).

In light of the prevalence with which skill acquisition assessments are used with ABA (Padilla, 2019), understanding psychometric properties of assessments is crucial for the scientific advancement of ABA assessment within research and practice. Furthermore, the use of psychometrics in the social sciences has a certain social validity (Baer, Wolf, & Risley, 1987). In the field of ABA, psychometric research on criterion-referenced assessments for skill acquisition is an important, emerging area of research with a growing number of studies examining the psychometric properties of assessments, such as the ABLLS-R, PEAK, among others. Given the value and widespread use of these skill acquisition assessments based in ABA, continued psychometric research is highly needed.

Psychometric Evidence for ABA Assessments

In order to determine the current reliability and validity evidence for ABA assessments used for curriculum development and educational programming, Padilla and

colleagues (2019) conducted a systematic literature review on the assessments that have been developed using the ABA framework. To the best of our knowledge, there have been no other systematic literature reviews conducted focusing on these types of assessments within the field of ABA. The search was conducted using the keywords "reliability," "validity," "factor analysis," "correlation," "item response theory," "structural equation modeling," "regression," and "assessment." These keywords were selected from common terms found in psychometric textbooks and literature. These keywords were paired with the following terms using Boolean operators and truncation: "verbal behavior" and "applied behavior analysis." This initial search produced additional instruments not included in Padilla's (2019) survey results; thus we decided to include the titles of those assessments to ensure a more comprehensive search. The following keyword search terms were included: "verbal behavior milestones assessment and placement program" OR "VB-MAPP," "assessment of basic language and learning skills" or "ABLLS," "assessment of basic language and learning skills" OR "ABLA," "promoting the emergence of advanced knowledge" OR "PEAK," "training and assessment of relational precursors and abilities" OR "TARPA," and "Verbal Behavior Assessment Scale" OR "VerBAS." The VB-MAPP, PEAK, and ABLLS-R were included as search terms based on survey results from the Austin and Thomas (2017) and Padilla (2019) studies. All assessment titles were paired with the psychometric terms as well in order to return available articles evaluating psychometric properties of the assessments.

Inclusion and Exclusion Criteria

In order for a study to be included in this review, a study must have included an ABA-based assessment that directly evaluated participants' present level of functioning

regarding their verbal repertoire (e.g., skills acquisition assessments). Additionally, the study must have assessed the validity or reliability of the instrument being utilized. Studies that were not published in a peer-reviewed journal (e.g., dissertations) or did not report utilizing instruments for clinical practice were not included.

A total of 1,268 articles were identified in the search, of which 1,182 were unique. Twelve articles were identified through other sources for a total of 1,194 articles to be reviewed. The titles and abstracts of the resulting 1,194 articles were reviewed for inclusion, after which 1,149 articles were excluded. A full-text search was conducted on the remaining 45 articles, during which nine articles were excluded. Once the full-text search was completed, an ancillary search was conducted in which the reference lists in the included studies were reviewed to identify additional articles for possible inclusion. An author search for the author with the highest number of publications (i.e., M. R. Dixon, J. Belisle, & K. E. Rowsey) and a hand search in the journal in which the most studies were published (i.e., *Journal of Developmental and Physical Disabilities, Research in Autism Spectrum Disorders)* were conducted to identify any potential studies to be included in the current review. This review process resulted in 35 articles to be included in the review with 39 total studies overall (four articles presented findings from two studies).

Data Extraction and Analysis

After the systematic search was completed, data related to each of the following categories were extracted from each article: (a) instrument-specific characteristics, (b) participant characteristics, and (c) general study information. For instrument-specific characteristics, data were extracted on the specific instrument used (e.g., VB-MAPP,

ABLLS), reliability estimates, content-, criterion-, and/or construct-related validity evidence, and the statistical analyses utilized. For participant characteristics, data were extracted on the number of participants, age, gender, race, ethnicity, IQ reporting, diagnoses, and any other measures collected (e.g., verbal repertoire, physical abilities). For general study information, data were extracted on the geographic location of the study, instrument assessor, and the setting in which the instrument was used.

The extracted data were analyzed and reported globally to summarize the total numbers and percentages across all included articles. The data were also analyzed by each assessment; for example, all articles evaluating the reliability or validity of the PEAK were analyzed and reported in terms of the total number and percentages for the PEAK.

Inter-rater Agreement

Search reliability. Two authors independently conducted the electronic database search, abstract and title search, and full text search to assess the reliability of the search terms and inclusion criteria. After independently applying the inclusion and exclusion criteria to the resulting articles, inter-rater agreement was calculated by dividing the total number of agreements by agreements plus disagreements and multiplying by 100%. Inter-rater agreement for the search was 99%. Discussion among all authors was used to resolve the resulting discrepancies.

Data extraction reliability. Two authors independently summarized 40% (n = 14) of the included studies included in the review to assess reliability of data extraction. Each study was summarized based on 17 items related to characteristics described above.

Interrater agreement on data extraction was calculated by taking the total number of agreements by the number of agreements plus disagreements and multiplying by 100%. Mean interrater agreement was 89%. Given the complexity of the studies, the researchers conducted consensus coding (e.g., Johnson, Penny, & Gordon 2000) with a psychometrician until interrater agreement reached 100%. Six instruments were identified in the literature that had reliability and/or validity studies. The global results are provided below followed by instrument-specific results specific to validity evidence for skill-based assessments developed within the ABA framework (i.e., ABLLS-R, PEAK, and VB-MAPP). Full descriptions for reliability and validity evidence for individual assessments can be found in Padilla and colleagues (2019).

Global Findings

Thirty-five articles were found, some of which presented multiple studies. Therefore, the total number of psychometric studies returned using the search procedures described above was 39. Some articles indicated separate studies within the same article which were denoted by "Experiment 1" or "Study 1" because they included different samples and thus were considered separate studies for the purposes of these analyses. These 39 studies provided reliability and/or validity evidence across five instruments identified: ABLLS-R (n = 2), PEAK (n = 11), TARPA (n = 5), ABLA (n = 20), and VerBAS (n = 1).

Validity. Of the 39 studies identified, 35 presented evidence for an instrument's validity. Two of the validity studies reported content-, 26 reported criterion-, and seven reported construct-related validity evidence. In the two studies reporting evidence for

content-related validity, one (i.e., Usry, Partington, & Partington, 2018) used an expert panel to rate how "essential" each skill was on the instrument under investigation, and the second The authors reported the content validity ratio (CVR) as content-related validity evidence. For those reporting criterion-related validity evidence, the criteria used for comparisons were intelligence tests (n = 11, 41%; 8 full scale, 3 abbreviated scale), adaptive measures (n = 9, 33%), achievement/academic instruments (n = 3, 11%), and language and communication instruments (n = 4, 15%). Twelve studies used task completion or skill acquisition as the external criterion. The Gilliam Autism Rating Scale (GARS), VB-MAPP, and ABLLS-R were each used in one study each as the external criterion for validity evidence. In some cases, multiple instruments were used for comparisons. As for the statistical analyses used for support claims of criterion-related validity, 14 (52%) estimated a correlation coefficient (e.g., Pearson r, Spearman ρ, φ), three (11%) used a type of regression analysis (e.g., linear, quadratic, cubic, logistic), eight (30%) used descriptive statistics (e.g., percentages of patterns of performance on assessment tasks), one (4%) used a t test, and one (4%) employed a Fisher exact test. With respect to construct-validity evidence, three (43%) reported using principal components analysis, two estimated correlation and regression models (29%), and two (29%) reportedly used order analysis.

Participant demographics. A total of 1,776 participants were represented across all studies where sample size was reported. The mean sample size was approximately 47 (SD = 48), and ranged from 5 to 206. Of the 39 studies, the sex of the participants was reported in 32. Overall about 68% (n =991) of the participants were male and 32% were female (n = 474). The percentage of male participants ranged from 29% to 92%.

Participants ranged in age from 6 months to 66 years across all studies. Seventeen of the studies included participants with ASD, 15 included participants with intellectual disability, 14 included participants with developmental disability, six included typically developing participants, six studies included those classified as "other", and five studies did not include enough information to discern disability category. Only three studies provided the IQ scores of its participants.

Instrument-specific Findings

Informed by the assessment use findings of Padilla (2019), the following are the available results from the systematic literature review related to validity evidence for instruments developed within the ABA framework.

ABLLS-R

One study provided evidence for the instrument's content-related validity (i.e., Usry, Partington, & Partington, 2018). In this study, an expert panel rated how "essential" each skill was on the ABLLS-R, and the authors reported the CVR. The sample was comprised of 6 "experts," who were asked to evaluate the 544 items of the ABLLS-R. Three experts had a master's degree and were certified behavior analysts, (i.e., BCBA), one expert had a PhD and was a BCBA, one expert was an applied behavior therapist with a bachelor's degree, and one expert had no degree and worked as an applied behavior therapist. Three of the experts were female and three were male. Four experts were from Georgia and two from California. The average years of experience across experts 10.2 years (range = 9 to 19). The results provided content-related validity evidence as they reported that at least five out of six expert panel members rated 441 out

of the 544 items as "Essential." They calculated the CVR for all items and found that 441 items either met or exceeded their predetermined cutoff of .67 for containing evidence of content-related validity.

PEAK

Of the 13 studies identified, 12 (92.3%) presented evidence for validity of the use of PEAK. Of these studies, six (50%) reported criterion-validity and six (50%) reported construct-validity. . For the six criterion-related validity studies, the PEAK was evaluated against existing instruments such as the ABLLS-R, VB-MAPP, Peabody Picture Vocabulary Test, the Wechsler Intelligence Scale for Children (4th edition; WISC-IV), Vineland Adaptive Behavior Scales (2nd edition, VABS-2), and other intelligence, adaptive behavior, and language assessments. Four of the construct validity studies were regression-based using a normative sample for different modules of the PEAK. The remaining two studies used principal components analysis to examine the dimensional structure of the PEAK.

VB-MAPP

Based on the findings of the Padilla's (2019) survey, the VB-MAPP was the most commonly reported instrument used, however, within the systematic literature review no studies met the criteria related to assessing its validity or reliability evidence. It should be noted that several studies are referenced on the assessment's website but, again, these studies were not identified in the systematic search process for the VB-MAPP.

Conclusions

Within the systematic literature review, the PEAK had the highest number of psychometric studies (n = 11) of the three instruments reviewed here despite being the least reportedly used instrument in the field of behavior analysis based on Padilla (2019). In contrast, the VB-MAPP did not have any identified studies within the systematic literature review despite it being the most widely used instrument in the field. Although the VB-MAPP appears to be a promising assessment for children with language and language-related skill deficits to assist in identifying and developing interventions, currently there appears to be no evidence supporting reliability or validity of this instrument and its scores (Meadows, 2017). The findings from the systematic literature suggest that there is a mismatch between use of specific assessments and available reliability and validity evidence for those assessments. Additional research is needed to address this important gap in the literature and, ultimately, practice.

Purpose & Research Questions

Within the field of ABA, decisions regarding assessment need to be based on data and research. In the systematic literature review, only the presence of psychometric evidence was recorded – not the quality of the evidence. There is a strong need for both an evaluative tool for examining psychometric studies of ABA instruments as well as high quality validity studies involving instruments in ABA. Given that the VB-MAPP is the most widely used instrument, validity evidence should be examined in order to have confidence in the decisions made based on the results of this instrument. The purpose of the current study was to provide evidence for the content-related validity of the VB-MAPP. As noted previously, content validity has four components: (a) domain definition,

(b) domain relevance, (c) domain representation, and (d) appropriateness of test development procedures. The VB-MAPP is now in its second edition so domain definition and appropriateness of test development procedures are beyond the scope of this study. These two components typically occur within the initial stages of test development and are presumed to have already been completed. Therefore, domain relevance and domain representation of the VB-MAPP are the foci for this content validity investigation.

The guiding research questions are:

1) To what extent are items relevant for their target domains in the VB-MAPP?

2) To what extent are items within each domain of the VB-MAPP appropriate for the corresponding developmental age?

3) To what extent are the methods of measurement used for evaluating skills appropriate within each domain of the VB-MAPP?

4) To what extent do the specific items for each domain measured in the VB-MAPP collectively represent the target domain?

The next chapter will present the methods and procedures to address these research questions.

CHAPTER THREE

Method

Participants

There is variability regarding the recommended number of individuals to serve on expert panels for the purposes of evaluating content validity; recommendations range from two to 20 (Rubio, Berg-Weger, Tebb, Lee, & Rauch, 2003). For setting performance standards in which panel experts are asked to determine a cut-off score that demonstrates proficiency in a specific area, Raymond and Reid (2001) suggest a panel of 10-15 members. For the current study, I invited 25 subject matter experts (SMEs) to evaluate the assessment items of the VB-MAPP. Fifteen SMEs agreed to participate but two did not complete the survey due to effects of the COVID-19 pandemic.

Inclusion Criteria

Eligibility criteria for SMEs were based on recommendations from the *Standards for Educational and Psychological Testing* (AERA et al., 2014), additional inclusion criteria adapted from a similar validity study conducted by Usry, Partington, and Partington (2018) for the ABLLS-R, and essential qualifications stipulated in the VB-MAPP manual. According to the *Standards for Educational and Psychological Testing* (AERA et al. 2014), SMEs should have relevant training, experience, and qualifications, in addition to a history of publications in refereed journals, national presentations, and research on the phenomenon of interest. Because SMEs are expected to have a specific area of expertise, I recruited SMEs with certain credentials, an abundance of experience

working with individuals with autism or other developmental disabilities, as well as years of experience in applied practice.

The VB-MAPP manual states that evaluators must have the necessary prerequisite skills to conduct the VB-MAPP, which includes: (a) a basic understanding of Skinner's (1957) analysis of verbal, (b) knowledge of basic behavior analysis, (c) familiarity with basic linguistic structure, (d) familiarity of linguistic development of typically developing children, and (e) a good understanding of autism and other developmental disabilities (Sundberg, 2014).

In order to qualify as an SME, participants were required to (a) be a certified practicing behavior analyst (i.e., BCBA or BCBA-D), (b) have at least seven years of applied experience, five of which were specifically with individuals diagnosed with autism or other developmental disabilities, (c) have received training on how to administer the VB-MAPP from a qualified professional (i.e., behavior analyst, test author), (d) have used the VB-MAPP in some capacity (e.g., administering the VB-MAPP, using the VB-MAPP to select treatment goals, progress monitoring) for at least five years, (e) have independently administered the VB-MAPP at least once prior to the study (adapted from Usry, Partington, & Partington, 2018), and (f) meet the necessary prerequisite skills stated in the VB-MAPP manual. Participants had to review and confirm via email that they met the qualifications to participate in the study.

Participant Background and Demographics

Thirteen participants were recruited to serve as SMEs for this study, who represented five different regions of the United States. Four SMEs (31%) were from the Midwest region, four SMEs (31%) were from the Southwest region, three SMEs (23%)

were from the West region, and one SME (8%) each from the Northeast and Southeast regions. Twelve SMEs (92%) were female and one SME (8%) was male. Ninety-two percent of SMEs (n = 12) were White and one SME (8%) was Hispanic. Six SMEs (46%) held the BCBA credential and seven SMEs (54%) had the BCBA-D credential. Regarding highest degree attained, eight SMEs (62%) had a PhD, four SMEs (31%) had a master's degree (e.g., MA, MS), and one SME (8%) had an EdS degree. The average years of experience was 12.5 years (SD = 4.6). These demographic variables and current position for each of the SMEs are provided in Table 3.1.

Table 3.1.

Participant Demographics

SME	Highest Degree	Cert.	Current Position	Region	Race / Ethnicity	Gender	Exp.
1	PhD	BCBA-D	Asst. Professor	Northeast	White	F	14
2	PhD	BCBA	Clinical Professor	Midwest	White	F	15
3	PhD	BCBA-D	Clinical Professor	Southwest	White	F	8
4	MA	BCBA	Doctoral Candidate	West	White	F	25
5	PhD	BCBA-D	Lecturer	Southeast	White	F	8
6	PhD	BCBA-D	Asst. Professor	Midwest	White	M	14
7	MEd	BCBA	Behavior Analyst	Southwest	White	F	10
8	EdS	BCBA	Executive Director	Southwest	White	F	12
9	MA	BCBA	BCBA Supervisor	West	White	F	7
10	PhD	BCBA-D	Assoc. Professor	West	White	F	14
11	PhD	BCBA-D	Asst. Professor	Midwest	Hispanic	F	13
12	MS	BCBA	Behavior Analyst	Southwest	White	F	11
13	PhD	BCBA-D	Post-Doctoral Fellow	Midwest	White	F	12

Note. SME = subject matter expert; Cert. = Certification; Exp. = Years of Experience

Instrumentation

Participants were provided with the following: (a) cover letter, (b) conceptual information about the instrument, (c) instructions on how to rate items, and (d) a link to the survey. The cover letter included information on why the individual was selected to serve as an SME and the significance of evaluating a highly used instrument in the field

(Sudman & Bradburn, 1982). The letter included information regarding the purpose of the study and protocol for content review (i.e., rating each instrument item for representativeness, relevant domain dimensions). Participants were informed that they would be asked to provide recommendations for item revisions, additions, or deletions (Grant & Davis, 1997). Participants were also provided with materials that have background information on the VB-MAPP. Grant and Davis (1997) suggest that panel experts receive definitions for the content domain of the instrument, target population, setting descriptions, and scoring criteria.

Each participant serving as an SME received a unique survey link developed in Qualtrics that included the 170 items of the VB-MAPP Milestones, five groups of items of the EESA, and the 24 categories of the VB-MAPP Barriers Assessment. The survey format followed the copyright guidelines set forth in the VB-MAPP Guide. To ensure further copyright protection, the Qualtrics source code was edited to disallow copy and paste functionality. SMEs were also provided with instructions on how to complete the review questionnaire (c.f., Grant & Davis, 1997). For each of the five items in a milestone domain, each SME assigned a rating indicating the (a) domain relevance (i.e., items are relevant to target domain), (b) developmental age appropriateness, and (c) method of measurement appropriateness. For domain relevance, SMEs used a three-point Likert-type response scale developed by Lawshe (1975) with the following categories: "Not Necessary," "Useful, but Not Essential," and "Essential." For developmental age appropriateness and method of measurement appropriateness, SMEs used a three-point Likert-type response scale with the following categories: "Not Appropriate," "Somewhat Appropriate," and "Very Appropriate." For each domain, SMEs rated the domain

representation based on the number and content of the five items within each domain using a three-point Likert-type response scale with the following categories: "Inadequate," "Somewhat Adequate," and "Adequate."

For the EESA, SMEs followed a similar procedure as described above. That is, they assigned a rating for each group of prompts indicating the (a) domain relevance (i.e., items are relevant to overall EESA), (b) developmental age appropriateness, and (c) prompt appropriateness (e.g., the specific words used to assess "simple and reduplicated syllable"). For domain relevance, SMEs used a three-point Likert-type response scale with the following categories: "Not Necessary," "Useful, but Not Essential," and "Essential." For developmental age appropriateness and prompt appropriateness, SMEs used a three-point Likert-type response scale with the following categories: "Not Appropriate," "Somewhat Appropriate," and "Very Appropriate." For the EESA, SMEs rated the domain representation based on the number and content of the groups of items using a three-point Likert-type response scale with the following categories: "Inadequate," "Somewhat Adequate," and "Adequate." Although the EESA contains five groups with a varying number of items for each groups, ratings were provided on the groups of items as a whole. This was done because of the way the EESA is scored per group which is similar to how the Milestones Assessment is scored. Additionally, the EESA is considered a separate subtest and is the basis for the Echoic domain scores in the Milestones Assessment.

For the VB-MAPP Barriers Assessment, SMEs assigned a rating for each barrier category indicating the (a) domain relevance (e.g., importance of "negative behaviors" to barriers assessment), (b) method of measurement appropriateness (i.e., response scale

appropriate for each barrier), and (c) domain representation. For domain relevance, SMEs used a three-point Likert-type response scale with the following categories: “Not Necessary,” “Useful, but Not Essential,” and “Essential.” For method of measurement appropriateness, SMEs used a three-point Likert-type response scale with the following categories: “Not Appropriate,” “Somewhat Appropriate,” and “Very Appropriate.” For VB-MAPP Barriers Assessment, SMEs rated the domain representation based on the number and types of barriers using a three-point Likert-type response scale with the following categories: “Inadequate,” “Somewhat Adequate,” and “Adequate.” Age appropriateness was not rated because the Barriers Assessment is not specific to any developmental age level.

Procedures

The design for this research is a traditional content validity study that uses the rating method as described in Chapter One. Prior to conducting the study, I obtained approval from the Institutional Review Board to ensure the research protocol, procedures, and informed consent were appropriate for the study.

For participant recruitment, I consulted with colleagues practicing in the field of ABA to determine a potential pool of participants to serve on the expert panel. Participants were recruited from various regions across the United States and contacted via email to invite them to participate in the study. The email contained the cover letter, participant inclusion criteria, and compensation for their time. If they agreed to participate, they were provided with a screening questionnaire to confirm their eligibility to participate in the study (Usry et al., 2018). Once confirmed, participants were provided conceptual information of the VB-MAPP, rating instructions, and a unique link to

complete the study. Participants were given six weeks to complete the study, and I provided a reminder to complete the study at weeks three and five.

At the completion of the study, ratings produced from each SME were recorded in a spreadsheet along with a unique random identifier for each participant; no personally identifiable information was recorded for these participants but included demographic information collected. The data were stored on an encrypted drive.

Data Analysis

After all data were collected, the following analyses were conducted using IBM SPSS version 25 (IBM Corp, 2017) and Microsoft Excel (Microsoft Corporation, 2018). First, the frequency distribution for each item was generated. These analyses demonstrated the number of SMEs who rated each item's (a) domain relevance, (b) developmental age appropriateness, (c) method of measurement/prompt appropriateness, and (d) domain representation. Popham (1992) recommended that 70% of SMEs endorsing an item's relevance to support content validity as sufficient, which is most closely approximated to nine out 13 SMEs (69.2%). All percentages were compared against the threshold of 69.2%, which is informed by Popham (1992). Second, the content validity ratio (CVR, Lawshe, 1975) was computed for each item as follows:

$$\text{CVR} = \frac{n_e - \frac{N}{2}}{\frac{N}{2}}$$

where n_e is the number of SMEs rating the item as "Essential," and N is the total number of SMEs who provided a rating. The CVR can range from +1 to -1, with higher scores indicating greater content validity evidence for the item. A CVR of 0 indicates that 50% of the SMEs rated the item as "Essential." Wilson, Pan, and Schumsky (2012)

recalculated the critical values for Lawshe's (1975) CVR based on differing levels of Type I error control and number of SMEs. The critical value for the CVR with 13 SMEs using a one-tailed test with Type I error rate of 5% is .456, and it was used as a comparison for all CVRs in this study. The critical value of .456 corresponds to 10 or more SMEs must rate an item as "Essential" in order to be considered statistically significant. The content validity index (CVI) was calculated as the average CVR across all items and can be interpreted as content validity evidence of the domain as a whole (Lynn, 1986; Shultz, Whitney, & Zickar, 2014). Additionally, the CVI was calculated for each level and the test as a whole.

It should be noted that the critical value of CVR requires more SMEs to rate an item as "Essential" than Popham's recommended criterion (i.e., 10 versus 9). Based on the size of the sample, each SME has considerable weight in the distribution of ratings. Ten out of 13 SMEs (76.9%) has a CVR of .54, and nine out of 13 SMEs (69.2%) has a CVR of .38. The difference between these two criteria is an example of the relationship between effect sizes, sample size, Type I error rate, and statistical power. An SME endorsement rating of 69.2% is considered meaningful even though the hypothesis test of the CVR is more conservative.

In order to examine whether there were differences in statistically significant CVRs and method of measurement appropriateness ratings across different methods of measurement, Cramer's *V* was computed. *V* is a measure of association between categorical variables, and can be expressed as:

$$V = \sqrt{\frac{\chi^2}{MN}},$$

where χ^2 is the model estimated value, *M* is the minimum of rows-1 or columns-1, and *N* is the sample size.

CHAPTER FOUR

Results

Completion Rates

Prior to analyzing the data, the number of missing responses was examined. For each item response, an SME provided 4 different ratings. Excluding open-ended responses, the survey required 609 ratings for each SME. A total of 31 out of 7,917 possible ratings (0.4%) were missing from the entire data set. For 29 of the rating opportunities, only one rating was missing from one SME for a total of 29 missing values. For one rating opportunity, two SMEs did not respond. According to Schafer (1999), a missing data rate of 5% or less is inconsequential. Percentages below are based on the number of responses for each rating opportunity.

VB-MAPP Milestones Overall Results

The entire Milestones Assessment, which includes all items for each domain across all three developmental levels, had a CVI = .32. As noted above, the CVI was computed for each Level as the average of the domain CVIs with that Level; the overall CVI for the Milestones Assessment was the average of three Levels' CVIs. The overall CVI estimate of .32 indicates that, in general, a majority of SMEs rated items as "Essential" although Levels 2 and 3 each contained one or more domains that were not rated as "Essential" by a majority of SMEs (i.e., CVR $\leq$ 0), on average. The domains with negative CVIs were Classroom Routines and Group Skills and Echoic within Level 2 as well as Visual Perceptual Skills and Matching-to-Sample in Level 3.

For each of 170 items the VB-MAPP Milestones Assessment, the CVRs were compared against the upper-tailed critical value of .456, which was associated with a Type I error rate (α) of 5%. Fifty-three of the 170 items (31%) had CVRs that exceeded .456; thus, the null hypothesis that these CVRs were equal to zero could be rejected in favor of the alternative hypothesis that the CVRs were statistically greater than zero. Each level had at least one domain that had no statistically significant items. In Level 1, 17 of the 45 items (38%) had CVRs that were statistically greater than zero. Within Level 1, the Independent Play and Spontaneous Vocal Behavior domains had no significant items that were statistically significant. In Level 2, 17 of the 60 items (28%) had CVRs that were statistically greater than zero. Four domains in Level 2 did not have any statistically significant items, which were Visual Perceptual Skills and Matching-to-Sample, Social Behavior and Social Play, Echoic, and Classroom Routines and Group Skills. In Level 3, 19 of the 65 items (29%) had CVRs statistically greater than zero. Two domains within this level – Visual Perceptual Skills and Matching-to-Sample and Writing – had no statistically significant items. Following are the results by level and domain, including a summary of commentary provided by the SMEs.

VB-MAPP Level 1 Results

The calculated CVI for Level 1 across all domains was .35. This indicates that on average, the majority of raters indicated that the items within domains were "Essential." Following is a summary of each domain within Level 1, which includes domain relevance, age appropriateness, measurement appropriateness, and domain representation. Item-level response distributions and CVRs for each level (with associated statistical significance classifications) are reported in Appendix A. It should be noted that item-

level results are presented collectively within each domain as a range of rating percentages across SMEs. For some items, all SMEs (i.e., 100%) used the same rating (e.g., Essential or Very Appropriate); as a result, some rating percentage ranges for an item will include 0%.

Mand

Domain relevance. The estimated CVI, which is the average CVR across items, was .54. This indicates that a large majority of SMEs rated items as "Essential," on average. The percentage of SMEs who rated each item as "Essential" exceeded 69% for all five items and ranged from 69% (*n* = 9, CVR = .38) to 92% (*n* = 12, CVR=.85). Three of the five items had estimated CVRs that were significantly greater than zero; and 86% or more of the SMEs rated the items as "Essential" or "Useful, but Not Essential" for all five items.

Age appropriateness. Related to developmental age appropriateness, the percentage of SMEs who rated each item as "Very Appropriate" ranged from 62% (*n* = 8) to 85% (*n* = 11). SMEs who rated each item as "Somewhat Appropriate" for this developmental age ranged from 8% (*n* = 1) to 31% (*n* = 4). For four of the five items within this domain, one SME indicated that the item was "Not Appropriate."

Measurement appropriateness. Regarding the appropriateness for the method of measurement, the percentage of SMEs who rated the method "Very Appropriate" ranged from 46% (*n* = 6) to 62% (*n* = 8). SMEs who rated each item as "Somewhat Appropriate"

ranged from 23% (*n* = 3) to 46% (*n* = 6). For two of the five items, one SME (8%) indicated that the item was "Not Appropriate."

Domain representation. Considering the number and content of items within this domain, 54% (*n* = 7) of SMEs considered the items and content to be an "Adequate" representation of the domain for this developmental age range. Thirty-eight percent (*n* = 5) of SMEs indicated that the number and content of items was "Somewhat Adequate." One SME (11%) reported that the number and content of the items was an "Inadequate" representation for the Mand domain for Level 1 (i.e., 0 – 18 months).

Tact

Domain relevance. The CVI for this domain was .53, which indicates that a large majority of SMEs rated items as "Essential," on average. The percentage of SMEs who rated Tact items as "Essential" ranged from 58% (*n* = 7, CVR = .17) to 85% (*n* = 11, CVR = .69). Four of the five items had estimated CVRs that were significantly greater than zero and have "Essential" rating percentages that exceeded 69%. For the remaining item, 92% of the SMEs rated the item as "Essential" or "Useful, but Not Essential." For two of the five items within this domain, one SME (8%) indicated that the item was "Not Necessary."

Age appropriateness. The percentage of SMEs who rated each item as "Very Appropriate" for children 0 to 18 months ranged from 67% (*n* = 8) to 92% (*n* = 12). SMEs who rated each item as "Somewhat Appropriate" for this developmental age ranged from 0% (*n* = 0) to 25% (*n* = 3). In this instance, 0% of SMEs rated an item as

"Somewhat Appropriate" because 100% of SMEs rated that item as "Very Appropriate." For three of the five Tact items, one SME (8%) indicated that the item was "Not Appropriate."

Measurement appropriateness. The percentage of SMEs who rated the method of measurement for these items as "Very Appropriate" ranged from 33% (*n* = 4) to 83% (*n* = 10) whereas SMEs who rated each method as "Somewhat Appropriate" ranged from 8% (*n* = 1) to 58% (*n* = 7). For three of the five items, one SME (8%) indicated that the method of measurement was "Not Appropriate" and for one of the five items, two SMEs indicated that the method of measurement was "Not Appropriate" (18%).

Domain representation. Considering the number and content of items within this domain, 68% (*n* = 8) of SMEs considered the items and content to be an "Adequate" representation, and 38% (n = 5) of SMEs indicated it was "Somewhat Adequate."

Listener Responding

Domain relevance. The CVI for Listener Responding was .51, which indicates that a large majority of SMEs rated items as "Essential," on average. The percentage of SMEs who rated each item as "Essential" exceeded 69% for four items and ranged from 46% (*n* = 6, CVR = -0.08) to 92% (*n* = 12, CVR=.85) overall. Four of the five items had estimated CVRs that were significantly greater than zero. For the remaining item that did not meet significance, 46% (*n* = 6) rated the item as "Essential" and 46% (*n* = 6) rated the item as "Useful, but Not Essential."

Age appropriateness. The percentage of SMEs who rated each item as "Very Appropriate" ranged from 46% (*n* = 6) to 92% (*n* = 12). SMEs who rated each item as "Somewhat Appropriate" for this developmental age ranged from 8% (*n* = 1) to 31% (*n* = 4). For one of the five items within this domain, three SMEs (23%) indicated that the item was "Not Appropriate" for this developmental age range.

Measurement appropriateness. The percentage of SMEs who rated the items "Very Appropriate" ranged from 23% (*n* = 3) to 85% (*n* = 11) and as "Somewhat Appropriate" ranged from 15% (*n* = 2) to 62% (*n* = 8). For two of the five items, two SMEs (15%) indicated that the item was "Not Appropriate" and for one of the five items, one SME indicated that the method of measurement was "Not Appropriate" (10%).

Domain representation. Considering the number and content of items within this domain, 62% (*n* = 8) of SMEs considered the domain representation to be "Adequate," and 38% (n = 5) of SMEs indicated the representation was "Somewhat Adequate."

Visual Perceptual Skills and Matching-to-Sample

Domain relevance. The CVI for this domain was .25, which indicates that a slight majority of SMEs rated items as "Essential," on average. The percentage of SMEs who rated each item as "Essential" ranged from 46% (*n* = 6, CVR = -0.08) to 85% (*n* = 11, CVR=.69). For two of the five items, more than 69% of the SMEs endorsed all items as "Essential" for this domain; however only one had an estimated CVR that was significantly greater than zero. For the four items that were not statistically significant, 92% or more of the SMEs rated the items as "Essential" or "Useful, but Not Essential."

Age appropriateness. The percentage of SMEs who rated each item as "Very Appropriate" ranged from 62% (*n* = 8) to 100% (*n* = 13). The percentage of SMEs who rated each item as "Somewhat Appropriate" for children aged 0 to 18 months ranged from 0% (*n* = 0) to 38% (*n* = 5). For two of the five items within this domain, one SME (8%) indicated that the item was "Not Appropriate" for this developmental age range.

Measurement appropriateness. The percentages of SMEs who rated the methods as "Very Appropriate" ranged from 46% (*n* = 6) to 77% (*n* = 10) and as "Somewhat Appropriate" ranged from 23% (*n* = 3) to 54% (*n* = 7). For one of the five items, an SME (8%) indicated that the method of measurement was "Not Appropriate."

Domain representation. Thirty-eight percent (*n* = 5) of SMEs indicated that the number and content of the items was an "Adequate" representation of the domain. The remaining SMEs (*n* = 8, 62%) indicated the items were a "Somewhat Adequate" representation for this developmental age range.

Independent Play

Domain relevance. The CVI for this domain was .20, which indicates that a slight majority of SMEs rated items as "Essential," on average. The percentage of SMEs who rated each item as "Essential" ranged from 46% (*n* = 6, CVR = -0.08) to 69% (*n* = 9, CVR=.38). None of the five items had estimated CVRs that were significantly greater than zero, but three items were rated as "Essential" by more than 69% of SMEs. With the exception of one SME rating one item as "Not Necessary," all other items were rated as "Essential" or "Useful, but Not Essential."

Age appropriateness. The percentages of SMEs who rated each item as "Very Appropriate" for this developmental age ranged from 54% (*n* = 7) to 85% (*n* = 11) and as "Somewhat Appropriate" ranged from 15% (*n* = 2) to 46% (*n* = 6). For one of the five items within this domain, an SME (8%) indicated that the item was "Not Appropriate" for this developmental age range.

Measurement appropriateness. The percentage of SMEs who rated the items "Very Appropriate" ranged from 31% (*n* = 4) to 67% (*n* = 8). SMEs who rated each item as "Somewhat Appropriate" ranged from 33% (*n* = 4) to 69% (*n* = 9). For one of the five items, one SME (8%) indicated that the method of measurement was "Not Appropriate."

Domain representation. About a third (*n* = 4, 31%) of SMEs considered the number and content of the items to be an "Adequate" representation of the domain for this developmental age range, and 62%f (*n* = 8) of SMEs indicated that the items were "Somewhat Adequate" for domain representation. One SME (8%) reported that the items were an "Inadequate" representation of Independent Play for this developmental age range.

Social Behavior and Social Play

Domain relevance. The CVI for this domain was .20, which indicates that a slight majority of SMEs rated items as "Essential," on average. The percentage of SMEs who rated each item as "Essential" ranged from 15% (*n* = 2, CVR = -0.69) to 77% (*n* = 10, CVR=.54). Three of the five items had estimated CVRs that were significantly greater

than zero and “Essential” rating by 69% or more of SMEs. All SMEs rated the remaining two items as “Essential” or “Useful, but Not Essential.”

Age appropriateness. The percentage of SMEs who rated each item as “Very Appropriate” ranged from 38% (*n* = 5) to 92% (*n* = 12). SMEs who rated each item as “Somewhat Appropriate” for this developmental age ranged from 8% (*n* = 1) to 62% (*n* = 8). None of the SMEs indicated that a Social Behavior and Social Play item was “Not Appropriate” for this developmental age.

Measurement appropriateness. The percentage of SMEs who rated the items “Very Appropriate” ranged from 31% (*n* = 4) to 46% (*n* = 6). SMEs who rated each item as “Somewhat Appropriate” ranged from 46% (*n* = 6) to 62% (*n* = 8). For four of the five items, one SME (8%) indicated that the method of measurement was “Not Appropriate.”

Domain representation. Considering the number and content of items within this domain, 50% (*n* = 6) of SMEs considered the items to be an “Adequate” representation of the domain for this developmental age range. Forty-two percent (*n* = 5) of SMEs indicated that the items were “Somewhat Adequate.” One SME (8%) reported that the items were an “Inadequate” representation of this domain. One SME did not respond to this survey item.

Motor Imitation

Domain relevance. The CVI for this domain was .35, which indicates that a majority of SMEs rated items as “Essential,” on average. The percentage of SMEs who rated each item as “Essential” ranged from 62% (*n* = 8, CVR = .23) to 77% (*n* = 10,

CVR=.54). For three of the items, about 69% of the SMEs endorsed those items as "Essential;" however, only one item had an estimated CVR that was significantly greater than zero. All SMEs rated the four items that were not statistically significant as "Essential" or "Useful, but Not Essential."

Age appropriateness. All SMEs indicated the age appropriateness of the items were "Very Appropriate" or "Somewhat Appropriate." The percentages of SMEs who rated each item as "Very Appropriate" ranged from 69% (*n* = 9) to 85% (*n* = 11) and as "Somewhat Appropriate" ranged from 15% (*n* = 2) to 31% (*n* = 4).

Measurement appropriateness. The percentage of SMEs who rated the methods of measurement "Very Appropriate" ranged from 69% (*n* = 9) to 92% (*n* = 12). SMEs who rated each method of measurement as "Somewhat Appropriate" ranged from 8% (*n* = 1) to 31% (*n* = 4). None of the SMEs indicated that the methods of measurement were "Not Appropriate" for any item.

Domain representation. Considering the number and content of items within this domain, 77% (*n* = 10) considered the items and content to be an "Adequate" representation of the domain for this developmental age range. Twenty-three percent (n = 3) of SMEs indicated that the domain representation of the items was "Somewhat Adequate." No SME reported that the items were "Inadequate."

Echoic

Domain relevance. The CVI for the Echoic domain was .35, which indicates that a majority of SMEs rated items as "Essential," on average. The percentage of SMEs who

rated each item as “Essential” ranged from 54% (*n* = 7, CVR = 0.08) to 77% (*n* = 10, CVR=.54). For four of the items, approximately 69% of the SMEs endorsed all items as “Essential” for this domain. One item had an estimated CVR that was significantly greater than zero, but all SMEs rated the remaining four items as either “Essential” or “Useful, but Not Essential.”

Age appropriateness. The percentage of SMEs who rated each item as “Very Appropriate” ranged from 69% (*n* = 9) to 100% (*n* = 13). SMEs who rated each item as “Somewhat Appropriate” for this developmental age ranged from 0% (*n* = 0) to 31% (*n* = 4). None of the items were rated as “Not Appropriate” for this developmental age range by any SMEs.

Measurement appropriateness. All SMEs rated the method of measurement as “Very Appropriate” or “Somewhat Appropriate” for the Echoic domain. The percentage of SMEs who rated the methods of measurement “Very Appropriate” ranged from 69% (*n* = 9) to 85% (*n* = 11) and “Somewhat Appropriate” ranged from 15% (*n* = 2) to 31% (*n* = 4).

Domain representation. Sixty-two percent of SMEs (*n* = 8) considered the items and content to be an “Adequate” representation of the domain for this developmental age range. Thirty-eight percent (n = 5) indicated that the number and content of items was “Somewhat Adequate.” No SME reported that the number and content of the items was an “Inadequate” representation of Echoic domain for this developmental age range.

Spontaneous Vocal Behavior

Domain relevance. The CVI for this domain was .23, which indicates that a slight majority of SMEs rated items as "Essential," on average. The percentage of SMEs who rated each item as "Essential" ranged from 46% (n = 6, CVR = -0.08) to 69% (n = 9, CVR = .38). For three of the five items, about 69% of the SMEs endorsed all items as "Essential," but none had estimated CVRs that were significantly greater than zero. The majority of SMEs rated all items as "Essential" or "Useful, but Not Essential."

Age appropriateness. SMEs rated the age appropriateness of the Spontaneous Vocal Behavior items as either "Very Appropriate" or "Somewhat Appropriate." The percentage of SMEs who rated each item as "Very Appropriate" ranged from 62% (n = 8) to 92% (n = 12). SMEs who rated each item as "Somewhat Appropriate" for this developmental age ranged from 8% (n = 1) to 38% (n = 5).

Measurement appropriateness. The percentage of SMEs who rated the methods of measurement as "Very Appropriate" ranged from 62% (n = 8) to 77% (n = 10). SMEs who rated each method of measurement as "Somewhat Appropriate" ranged from 8% (n = 1) to 31% (n = 4). For four items, one SME (8%) indicated that the method of measurement was "Not Appropriate" and for one item, two SMEs (15%) indicated that the method of measurement was "Not Appropriate" (20%).

Domain representation. Considering the number and content of items within this domain, 54% (n = 7) considered the items to be an "Adequate" representation of the

domain for this developmental age range. Forty-six percent (n = 6) of SMEs indicated that the items were "Somewhat Adequate" for this developmental age range.

Table 4.1.

VB-MAPP Level 1 Content Validity Values

Domain	CVI	Minimum CVR	Maximum CVR	No. of Significant Items*
Mand	.54	.38	.85	3
Tact	.53	.17	.69	4
Listener Responding	.51	-.08	.85	4
Visual Perceptual Skills & Matching-To-Sample	.25	-.08	.69	1
Independent Play	.20	-.08	.38	0
Social Behavior & Social Play	.20	-.69	.54	3
Motor Imitation	.35	.23	.54	1
Echoic	.35	.08	.54	1
Spontaneous Vocal Behavior	.23	-.08	.38	0

* upper-tailed $p < .05$ for 13 SMEs

Note. CVI = Content validity index; Minimum CVR = Minimum content validity ratio among the five items for each domain; Maximum CVR = Maximum content validity ratio among the five items for each domain.

VB-MAPP Level 2 Results

The calculated CVI across all Level 2 domains was .30, which indicates that, on average, the majority of raters indicated that the items within domains were "Essential." Following is a summary of each domain within Level 2, which includes domain relevance, age appropriateness, measurement appropriateness, and domain representation.

Mand

Domain relevance. The CVI for this domain was .51, which indicates that a large majority of SMEs rated items as "Essential," on average. The percentage of SMEs who rated each item as "Essential" ranged from 54% (n = 7, CVR = .08) to 100% (n = 13, CVR=1). At least nine of the 13 SMEs (69% or more) endorsed four items as "Essential,"

but only two had estimated CVRs that were significantly greater than zero. With the exception of one SME rating one item as "Not Necessary", all SMEs rated the three items that were not statistically significant as "Essential" or "Useful, but Not Essential."

Age appropriateness. SMEs rated the age appropriateness for the items within this domain as either "Very Appropriate" or "Somewhat Appropriate." The percentage of SMEs who rated each item as "Very Appropriate" ranged from 54% (*n* = 7) to 100% (*n* = 13). SMEs who rated each item as "Somewhat Appropriate" for this developmental age ranged from 0% to 46% (*n* = 6).

Measurement appropriateness. The percentage of SMEs who rated the methods of measurement "Very Appropriate" ranged from 38% (*n* = 5) to 85% (*n* = 11). SMEs who rated each method of measurement as "Somewhat Appropriate" ranged from 15% (*n* = 2) to 54% (*n* = 7). For four of the five items, an SME indicated that the method of measurement was "Not Appropriate" (8%).

Domain representation. Considering the number and content of items within this domain, 54% (*n* = 7) considered the items to be an "Adequate" representation of the domain for this developmental age range. Forty-six percent (*n* = 6) of SMEs indicated that the items were "Somewhat Adequate" representation of this domain for this developmental age range.

Tact

Domain relevance. The CVI for this domain was .54, which indicates that a large majority of SMEs rated items as "Essential," on average. The percentage of SMEs who

rated each item as “Essential” ranged from 62% (*n* = 8, CVR = .23) to 92% (*n* = 12, CVR=.85). For four of the items, at least nine of the SMEs endorsed those items as “Essential” for this domain, but only two items had estimated CVRs that were significantly greater than zero. Although only two items reached statistical significance values, all SMEs rated all five items as “Essential” or “Useful, but Not Essential.”

Age appropriateness. The percentage of SMEs who rated each item as “Very Appropriate” ranged from 62% (*n* = 8) to 92% (*n* = 11). SMEs who rated each item as “Somewhat Appropriate” for this developmental age ranged from 8% (*n* = 1) to 38% (*n* = 5). None of the SMEs indicated that any items were “Not Appropriate.”

Measurement appropriateness. The percentage of SMEs who rated the methods of measurement “Very Appropriate” ranged from 62% (*n* = 8) to 92% (*n* = 12). SMEs who rated each method of measurement as “Somewhat Appropriate” ranged from 8% (*n* = 1) to 23% (*n* = 3). For one of the five items, one SME (8%) indicated that the item’s method of measurement was “Not Appropriate” and two SMEs (15%) indicated that another item’s method of measurement was “Not Appropriate.”

Domain representation. Considering the number and content of items within this domain, 62% (*n* = 8) considered the items to be an “Adequate” representation of the domain for this developmental age range. Thirty-eight percent (*n* = 5) of SMEs indicated that the number and content of items was “Somewhat Adequate” for this developmental age range.

Listener Responding

Domain relevance. The CVI for this domain was .20, which indicates that a slight majority of SMEs rated items as "Essential," on average. The percentage of SMEs who rated each item as "Essential" ranged from 38% (*n* = 5, CVR = -.23) to 92% (*n* = 11, CVR=.83) and exceeded 69% for one item. Only one item had an estimated CVR that was significantly greater than zero. With the exception of two items being rated as "Not Necessary" by one or two SMEs, all other items were rated as "Essential" or "Useful, but Not Essential."

Age appropriateness. The percentage of SMEs who rated each item as "Very Appropriate" ranged from 42% (*n* = 5) to 83% (*n* = 10). The percentage of SMEs who rated each item as "Somewhat Appropriate" for this developmental age ranged from 17% (*n* = 2) to 50% (*n* = 6). Two of the items were rated as "Not Appropriate" by one SME (8%).

Measurement appropriateness. The percentage of SMEs who rated the methods of measurement as "Very Appropriate" ranged from 54% (*n* = 7) to 100% (*n* = 13). The percentage of SMEs who rated each method of measurement as "Somewhat Appropriate" ranged from 0% (*n* = 0) to 31% (*n* = 4). For three of the five items, two SMEs (15%) indicated that the method of measurement for these items was "Not Appropriate." For one of the items, one SME (8%) indicated it was "Not Appropriate."

Domain representation. Considering the number and content of items within this domain, 38% (*n* = 5) considered the items to be an "Adequate" representation of the

domain for this developmental age range. Fifty-four percent (*n* = 7) of SMEs indicated that the items was "Somewhat Adequate." One SME (8%) reported that the number and content of the items was an "Inadequate" representation of this domain for this developmental age range.

Visual Perceptual Skills and Matching-to-Sample

Domain relevance. The CVI for this domain was .17, which indicates that a slight majority of SMEs rated items as "Essential," on average. The percentage of SMEs who rated each item as "Essential" ranged from 38% (*n* = 5, CVR = -.23) to 69% (*n* = 9, CVR=.38) but only exceeded 69% for one item. None of the five items had estimated CVRs that were significantly greater than zero; however, all SMEs rated all of the five items as "Essential" or "Useful, but Not Essential."

Age appropriateness. The percentage of SMEs who rated each item as "Very Appropriate" ranged from 38% (*n* = 5) to 77% (*n* = 10), and the percentage of SMEs who rated each item as "Somewhat Appropriate" for this developmental age ranged from 23% (*n* = 3) to 62% (*n* = 8). One SME (8%) indicated that one of the five items was "Not Appropriate" for this developmental age range.

Measurement appropriateness. The percentage of SMEs who rated the methods of measurement "Very Appropriate" ranged from 62% (*n* = 8) to 77% (*n* = 10). SMEs who rated each methods of measurement as "Somewhat Appropriate" ranged from 23% (*n* = 3) to 38% (*n* = 5). No SMEs indicated the method of measurement for any of the five items was "Not Appropriate."

Domain representation. Considering the number and content of items within this domain, 69% (*n* = 9) considered the items to be an "Adequate" representation of the domain for this developmental age range. Twenty-three percent (*n* = 3) of SMEs indicated that the number and content of items was "Somewhat Adequate." One SME (8%) reported that the number and content of the items was an "Inadequate" representation of this domain for this developmental age range.

Independent Play

Domain relevance. The CVI for this domain was .38, which indicates that a majority of SMEs rated items as "Essential," on average. The percentage of SMEs who rated each item as "Essential" ranged from 54% (*n* = 7, CVR = .08) to 92% (*n* = 12, CVR=.85). Two items had estimated CVRs that were significantly greater than zero and were rated as "Essential" more 69% of SMEs. For the three items that were not statistically significant, all SMEs indicated that they were "Essential" or "Useful, but Not Essential." None of the items were rated as "Not Necessary."

Age appropriateness. The percentage of SMEs who rated each item as "Very Appropriate" ranged from 69% (*n* = 9) to 100% (*n* = 13). SMEs who rated each item as "Somewhat Appropriate" for this developmental age ranged from 0% (*n* = 0) to 31% (*n* =4). None of the SMEs indicated that any item was "Not Appropriate" for this developmental age range.

Measurement appropriateness. The percentage of SMEs who rated the methods of measurement "Very Appropriate" ranged from 38% (*n* = 5) to 69% (*n* = 9). SMEs who

rated each method of measurement as “Somewhat Appropriate” ranged from 23% (*n* = 3) to 54% (*n* = 7). For four of the five items, one SME (8%) indicated that the method of measurement was “Not Appropriate.”

Domain representation. Considering the number and content of items within this domain, 62% (*n* = 8) considered the items to be an “Adequate” representation of the domain for this developmental age range. Thirty-eight percent (*n* = 5) of SMEs indicated that the number and content of items was a “Somewhat Adequate” representation of this domain for this developmental age range.

Social Behavior and Social Play

Domain relevance. The CVI, which is the average CVR across items, was .11, which indicates that a slight majority of SMEs rated items as “Essential,” on average. The percentage of SMEs who rated each item as “Essential” ranged from 46% (*n* = 6, CVR = -.08) to 62% (*n* = 8, CVR=.23). None of the five items had estimated CVRs that were significantly greater than zero, but all items were rated as “Essential” or “Useful, but Not Essential” by all SMEs.

Age appropriateness. The percentage of SMEs who rated each item as “Very Appropriate” ranged from 46% (*n* = 6) to 69% (*n* = 9). SMEs who rated each item as “Somewhat Appropriate” for this developmental age ranged from 31% (*n* = 4) to 46% (*n* =6). For three of the items one SME (8%) indicated that the item was “Not Appropriate” for this developmental age range.

Measurement appropriateness. The percentage of SMEs who rated the methods of measurement as "Very Appropriate" ranged from 23% (*n* = 3) to 38% (*n* = 5) while those who rated each method of measurement as "Somewhat Appropriate" ranged from 42% (*n* = 5) to 69% (*n* = 9). For three of the five items, one SME (8%) indicated that the method of measurement for those items was "Not Appropriate" (10%) and for one of the methods of measurement, two SMEs (17%) indicated that it was "Not Appropriate."

Domain representation. Thirty-one percent of SMEs (*n* = 4) considered the items and content to be an "Adequate" representation of the domain for this developmental age range. Fifty-four percent (*n* = 7) of SMEs indicated that the number and content of items was "Somewhat Adequate." Two SMEs (15%) reported that the number and content of the items was an "Inadequate" representation of this domain for this developmental age range.

Motor Imitation

Domain relevance. The CVI for this domain was .48, which indicates that a majority of SMEs rated items as "Essential," on average. The percentage of SMEs who rated each item as "Essential" ranged from 54% (*n* = 7, CVR = .08) to 100% (*n* = 13, CVR=1). Two items had estimated CVRs that were significantly greater than zero and were rated as "Essential" by at least 69% of SMEs. The remaining items were rated as "Essential" or "Useful, but Not Essential" by all SMEs.

Age appropriateness. The percentage of SMEs who rated each item as "Very Appropriate" ranged from 62% (*n* = 8) to 92% (*n* = 12). SMEs who rated each item as

"Somewhat Appropriate" for this developmental age ranged from 8% (*n* = 1) to 31% (*n* = 4). For two items, two SMEs (15%) indicated that the items were "Not Appropriate" for this developmental age range.

Measurement appropriateness. The percentage of SMEs who rated the methods of measurement "Very Appropriate" ranged from 69% (*n* = 9) to 85% (*n* = 11). SMEs who rated each item's method of measurement as "Somewhat Appropriate" ranged from 15% (*n* = 2) to 23% (*n* = 3). For two of the five items, one SME (8%) indicated that the method of measurement for that item was "Not Appropriate" while for another item, two SMEs (15%) reported that it was "Not Appropriate."

Domain representation. Forty-six percent of SMEs (*n* = 6) considered the items and content to be an "Adequate" representation of the domain for this developmental age range. Fifty-four percent (*n* = 7) of SMEs indicated that the number and content of items was "Somewhat Adequate." None of the SMEs reported that the items were an "Inadequate" representation of this domain for this developmental age range.

Echoic

Domain relevance. The CVI for this domain was -.02. For four of the five items, about half of the SMEs rated each item as "Essential." For the remaining item, fewer than a third (*n* = 4) of the SMEs indicated that it was "Essential." The percentage of SMEs who rated each item as "Essential" ranged from 31% (*n* = 4, CVR = -.38) to 62% (*n* = 8, CVR=.23). None of the five items had estimated CVRs that were significantly greater

than zero, but all SMEs indicated that the items were “Essential” or “Useful, but Not Essential.”

Age appropriateness. The percentage of SMEs who rated each item as “Very Appropriate” ranged from 54% (*n* = 7) to 92% (*n* = 12). SMEs who rated each item as “Somewhat Appropriate” for this developmental age ranged from 8% (*n* = 1) to 46% (*n* =6). None of the SMEs indicated that any item was “Not Appropriate” for this developmental age range.

Measurement appropriateness. The percentage of SMEs who rated the methods of measurement “Very Appropriate” ranged from 69% (*n* = 9) to 77% (*n* = 10). SMEs who rated each method of measurement as “Somewhat Appropriate” ranged from 23% (*n* = 3) to 31% (*n* = 4). None of the SMEs indicated that the method of measurement for any item was “Not Appropriate.”

Domain representation. Sixty-two percent of SMEs (*n* = 8) considered the items to be an “Adequate” representation of the domain for this developmental age range. The remaining 38% (*n* = 5) of SMEs indicated that the items were “Somewhat Adequate” for this developmental age range.

Listener Responding by Function, Feature, and Class (LRFFC)

Domain relevance. The CVI for this domain was .29, which indicates that a slight majority of SMEs rated items as “Essential,” on average. The percentage of SMEs who rated each item as “Essential” ranged from 54% (*n* = 7, CVR = .08) to 85% (*n* = 11, CVR=.69). For two items, at least 69% of the SMEs rated the items as “Essential” but

only one of the items had an estimated CVR that was significantly greater than zero. The majority of all SMEs rated all items as "Essential" or "Useful, but Not Essential."

Age appropriateness. The percentage of SMEs who rated each item as "Very Appropriate" ranged from 62% (*n* = 8) to 69% (*n* = 9). SMEs who rated each item as "Somewhat Appropriate" for this developmental age ranged from 31% (*n* = 4) to 38% (*n* =5). None of the SMEs indicated that any item was "Not Appropriate" for this developmental age range.

Measurement appropriateness. The percentage of SMEs who rated the methods of measurement "Very Appropriate" ranged from 62% (*n* = 8) to 85% (*n* = 11). SMEs who rated each method of measurement as "Somewhat Appropriate" ranged from 15% (*n* = 2) to 31% (*n* = 4). For three of the five items, one SME (8%) indicated that the method of measurement for that item was "Not Appropriate."

Domain representation. Forty-six percent of SMEs (*n* = 6) considered the items and content to be an "Adequate" representation of the domain for this developmental age range. Fifty-four percent (*n* = 7) of SMEs indicated that the number and content of items was "Somewhat Adequate." None of the SMEs reported that the number and content of the items was an "Inadequate" representation of this domain for this developmental age range.

Intraverbal

Domain relevance. The CVI for this domain was .51, which indicates that a large majority of SMEs rated items as "Essential," on average. The percentage of SMEs who

rated each item as “Essential” ranged from 54% (*n* = 7, CVR = .08) to 92% (*n* = 12, CVR=.85). Four items had estimated CVRs that were significantly greater than zero and were rated as “Essential” more 69% or more of SMEs. The remaining item was rated as “Essential” or “Useful, but Not Essential” by 92% of SMEs (*n* = 12).

Age appropriateness. The percentage of SMEs who rated each item as “Very Appropriate” ranged from 54% (*n* = 7) to 100% (*n* = 13). SMEs who rated each item as “Somewhat Appropriate” for this developmental age ranged from 0% (*n* = 0) to 46% (*n* =6). For two of the items, an SME (8%) indicated that the item was “Not Appropriate” for this developmental age range.

Measurement appropriateness. The percentage of SMEs who rated the methods of measurement “Very Appropriate” ranged from 69% (*n* = 9) to 92% (*n* = 12). SMEs who rated each method of measurement as “Somewhat Appropriate” ranged from 8% (*n* = 1) to 31% (*n* = 4). For one of the five items, one SME (8%) indicated that the method of measurement for that particular item was “Not Appropriate.”

Domain representation. Considering the number and content of items within this domain, 54% (*n* = 7) considered the items and content to be an “Adequate” representation of the domain for this developmental age range. The remaining forty-six percent (*n* = 6) of SMEs indicated that the number and content of items was “Somewhat Adequate” for this developmental age range.

Classroom Routines and Groups Skills

Domain relevance. The CVI for this domain was -.02. For all five items, approximately 50% of SMEs rated each item as "Essential." The percentage of SMEs who rated each item as "Essential" ranged from 46% (*n* = 6, CVR = -.08) to 54% (*n* = 7, CVR=.08). None of the five items had estimated CVRs that were significantly greater than zero. This was one of the few domains that had varying results. For all five items, less than 70% of the SMEs endorsed those items as "Essential" for this domain. However, the majority of SMEs (77% - 92%) provided a rating of either "Essential" or "Useful, but Not Essential." The percentage of SME ratings indicating an item was "Not Necessary" ranged from 8% - 23%.

Age appropriateness. The percentage of SMEs who rated each item as "Very Appropriate" ranged from 38% (*n* = 5) to 62% (*n* = 8). SMEs who rated each item as "Somewhat Appropriate" for this developmental age ranged from 38% (*n* = 5) to 46% (*n* =6). The percentage of SMEs that indicated an item was "Not Appropriate" for this developmental age range ranged from 0% (*n* = 0) to 23% (*n* = 3).

Measurement appropriateness. The percentage of SMEs who rated the methods of measurement "Very Appropriate" ranged from 62% (*n* = 8) to 69% (*n* = 9). SMEs who rated each method of measurement as "Somewhat Appropriate" ranged from 23% (*n* = 3) to 31% (*n* = 4). For three of the five items, one SME (8%) indicated that the method of measurement for an item was "Not Appropriate." For one of the items, two SMEs (15%) indicated it was "Not Appropriate."

Domain representation. Fifty percent of SMEs (*n* = 6) considered the items and content to be an "Adequate" representation of the domain for this developmental age range. Thirty-three percent (*n* = 4) of SMEs indicated that the number and content of items was "Somewhat Adequate." Two SMEs (17%) reported that the number and content of the items was an "Inadequate" representation of this domain for this developmental age range.

Linguistic Structure

Domain relevance. The CVI for this domain was .42, which indicates that the majority of SMEs rated items as "Essential," on average. The percentage of SMEs who rated each item as "Essential" ranged from 38% (*n* = 5, CVR = -.23) to 85% (*n* = 11, CVR=.69). Three of the five items had estimated CVRs that were significantly greater than zero, and four items were rated as "Essential" by 69% or more of SMEs. However, all SMEs indicated that all items were "Essential" or "Useful, but Not Essential."

Age appropriateness. The percentage of SMEs who rated each item as "Very Appropriate" ranged from 69% (*n* = 9) to 92% (*n* = 12). SMEs who rated each item as "Somewhat Appropriate" for this developmental age ranged from 8% (*n* = 1) to 31% (*n* =4). None of the SMEs indicated an item was "Not Appropriate" for this developmental age range.

Measurement appropriateness. The percentage of SMEs who rated the methods of measurement "Very Appropriate" ranged from 54% (*n* = 7) to 77% (*n* = 10). SMEs who rated each method of measurement as "Somewhat Appropriate" ranged from 23% (*n*

= 3) to 31% (*n* = 4). For three of the five items, an SME (8%) indicated that the method of measurement for an item was "Not Appropriate." For one of the items, three SMEs (23%) indicated it was "Not Appropriate."

Domain representation. Thirty-eight percent of SMEs (*n* = 5) considered the items and content to be an "Adequate" representation of the domain for this developmental age range. The majority of SMEs (*n* = 8, 62%) indicated that the number and content of items was a "Somewhat Adequate" representation of this domain for this developmental age range.

Table 4.2.

VB-MAPP Level 2 Content Validity Values

Domain	CVI	Minimum CVR	Maximum CVR	No. of Significant Items*
Mand	.51	.08	1.00	2
Tact	.54	.23	.85	2
Listener Responding	.20	-.23	.83	1
Visual Perceptual Skills & Matching-To-Sample	.17	-.23	.38	0
Independent Play	.38	.08	.85	2
Social Behavior & Social Play	.11	-.08	.23	0
Motor Imitation	.48	.08	1.00	2
Echoic	-.02	-.38	.23	0
Listener Responding by Function, Feature, & Class	.29	.08	.69	1
Intraverbal	.51	.08	.85	4
Classroom Routines & Group Skills	-.02	-.08	.08	0
Linguistic Structure	.42	-.23	.69	3

* upper-tailed $p < .05$ for 13 SMEs

Note. CVI = Content validity index; Minimum CVR = Minimum content validity ratio among the five items for each domain; Maximum CVR = Maximum content validity ratio among the five items for each domain.

VB-MAPP Level 3 Results

The calculated CVI for Level 3 across all domains was .30. This indicates that on average, the majority of raters indicated that the items within domains were "Essential."

Following is a summary of each domain within Level 3, which includes domain relevance, age appropriateness, measurement appropriateness, and domain representation.

Mand

Domain relevance. The CVI for this domain was .51, which indicates that a large majority of SMEs rated items as "Essential," on average. The percentage of SMEs who rated each item as "Essential" ranged from 62% (*n* = 8, CVR = .23) to 85% (*n* = 11, CVR=.69). Three items had estimated CVRs that were significantly greater than zero, and four items were rated as "Essential" by at least 69% of SMEs. All SMEs indicated that all items were "Essential" or "Useful, but Not Essential."

Age appropriateness. The percentage of SMEs who rated each item as "Very Appropriate" ranged from 77% (*n* = 10) to 92% (*n* = 12). The percentage of SMEs who rated the items "Somewhat Appropriate" ranged from 8% (*n* = 1) to 23% (*n* =3) for this developmental age range.

Measurement appropriateness. The percentage of SMEs who rated the methods of measurement "Very Appropriate" ranged from 54% (*n* = 7) to 77% (*n* = 10). SMEs who rated each method of measurement as "Somewhat Appropriate" ranged from 23% (*n* = 3) to 46% (*n* = 6). None of the SMEs indicated that the method of measurement for an item was "Not Appropriate."

Domain representation. Considering the number and content of items within this domain, 69% (*n* = 9) considered the items and content to be an "Adequate" representation of the domain for this developmental age range. Twenty-three percent (*n* = 3) of SMEs

indicated that the number and content of items was “Somewhat Adequate.” One SME (8%) reported that the number and content of the items was an “Inadequate” representation of this domain for this developmental age range.

Tact

Domain relevance. The CVI for this domain was .38, which indicates that a majority of SMEs rated items as “Essential,” on average. The percentage of SMEs who rated each item as “Essential” ranged from 54% (*n* = 7, CVR = .08) to 77% (*n* = 10, CVR=.54). For four of the five items, at least nine out of 13 SMEs (69%) rated the items as “Essential,” but only two had estimated CVRs that were significantly greater than zero. All SMEs indicated that the items were “Essential” or “Useful, but Not Essential.”

Age appropriateness. The percentage of SMEs who rated each item as “Very Appropriate” ranged from 69% (*n* = 9) to 100% (*n* = 13). SMEs who rated each item as “Somewhat Appropriate” for this developmental age ranged from 0% (*n* = 0) to 31% (*n* =4). None of the SMEs indicated an item was “Not Appropriate.”

Measurement appropriateness. The percentage of SMEs who rated the methods of measurement “Very Appropriate” ranged from 46% (*n* = 6) to 85% (*n* = 11). SMEs who rated each method of measurement as “Somewhat Appropriate” ranged from 15% (*n* = 2) to 31% (*n* = 4). For one of the five items, three SMEs (23%) indicated that the method of measurement for that item was “Not Appropriate.”

Domain representation. Sixty-percent of SMEs (*n* = 8) considered the items and content to be an “Adequate” representation of the domain for this developmental age

range. Thirty-one percent (*n* = 4) of SMEs indicated that the number and content of items was "Somewhat Adequate." One SME (8%) reported that the number and content of the items was an "Inadequate" representation of this domain for this developmental age range.

Listener Responding

Domain relevance. The CVI for this domain was .32, which indicates that a majority of SMEs rated items as "Essential," on average. The percentage of SMEs who rated each item as "Essential" ranged from 54% (*n* = 7, CVR = .08) to 77% (*n* = 10, CVR=.54). For three of the five items, at least nine of the 13 (69%) SMEs endorsed those items as "Essential," but only one had an estimated CVR that was significantly greater than zero. All SMEs indicated that all items were either "Essential" or "Useful, but Not Essential."

Age appropriateness. The percentage of SMEs who rated each item as "Very Appropriate" ranged from 69% (*n* = 9) to 92% (*n* = 12). SMEs who rated each item as "Somewhat Appropriate" for this developmental age ranged from 8% (*n* = 1) to 31% (*n* =4). None of the SMEs indicated an item was "Not Appropriate" for this developmental age range.

Measurement appropriateness. The percentage of SMEs who rated the methods of measurement "Very Appropriate" ranged from 50% (*n* = 6) to 92% (*n* = 12). SMEs who rated each method of measurement as "Somewhat Appropriate" ranged from 8% (*n*

= 1) to 25% (*n* = 3). For one of the five items, three SMEs (25%) indicated that the method of measurement for that item was "Not Appropriate."

Domain representation. Sixty-nine percent of SMEs (*n* = 9) considered the items and content to be an "Adequate" representation of the domain for this developmental age range. Twenty-three percent (*n* = 3) of SMEs indicated that the number and content of items was "Somewhat Adequate." One SME (8%) reported that the number and content of the items was an "Inadequate" representation of this domain for this developmental age range.

Visual Perceptual Skills and Matching-to-Sample

Domain relevance. The CVI for this domain was -.09. For three of the items, less than half of the SMEs, rated the items as "Essential." For the other two items about two-thirds of the SMEs rated the items as "Essential." The percentage of SMEs who rated each item as "Essential" ranged from 23% (*n* = 3, CVR = -.54) to 67% (*n* = 8, CVR=.33). For all five items, less than 69% of the SMEs endorsed the items as "Essential" and no item had an estimated CVR that was significantly greater than zero. A majority of SMEs (92% or more) indicated that the items were "Essential" or "Useful, but Not Essential." Only two items were rated as "Not Necessary" by one SME (8%).

Age appropriateness. The percentage of SMEs who rated each item as "Very Appropriate" ranged from 23% (*n* = 3) to 62% (*n* = 8). SMEs who rated each item as "Somewhat Appropriate" for this developmental age ranged from 31% (*n* = 4) to 77% (*n*

=10). None of the SMEs indicated an item was “Not Appropriate” for this developmental age range.

Measurement appropriateness. The percentage of SMEs who rated the methods of measurement “Very Appropriate” ranged from 62% (*n* = 8) to 75% (*n* = 9). SMEs who rated each method of measurement as “Somewhat Appropriate” ranged from 23% (*n* = 3) to 38% (*n* = 5). For two of the five items, one SME (8%) indicated that the method of measurement for an item was “Not Appropriate.”

Domain representation. Sixty-two percent of SMEs (*n* = 8) considered the items and content to be an “Adequate” representation of the domain for this developmental age range. Twenty-three percent (*n* = 3) of SMEs indicated that the number and content of items was “Somewhat Adequate.” Two SMEs (15%) reported that the number and content of the items was an “Inadequate” representation of this domain for this developmental age range.

Independent Play

Domain relevance. The CVI for this domain was .38, which indicates that a majority of SMEs rated items as “Essential,” on average. The percentage of SMEs who rated each item as “Essential” ranged from 46% (*n* = 6, CVR = -.08) to 85% (*n* = 11, CVR=.69). For four of the five items, at least nine of the 13 SMEs (69%) endorsed those items as “Essential” for this domain, but only two had estimated CVRs that were significantly greater than zero. None of the items were rated as “Not Necessary” by an SME.

Age appropriateness. The percentage of SMEs who rated each item as "Very Appropriate" ranged from 54% (*n* = 7) to 100% (*n* = 13). SMEs who rated each item as "Somewhat Appropriate" for this developmental age ranged from 0% (*n* = 0) to 46% (*n* =6). None of the SMEs indicated an item was "Not Appropriate" for this developmental age range.

Measurement appropriateness. The percentage of SMEs who rated the methods of measurement "Very Appropriate" ranged from 69% (*n* = 9) to 77% (*n* = 10). SMEs who rated each method of measurement as "Somewhat Appropriate" ranged from 15% (*n* = 2) to 23% (*n* = 3). For four of the five items, one SME (8%) indicated that the method of measurement for an item was "Not Appropriate." For one of the items, two SMEs (15%) indicated it was "Not Appropriate."

Domain representation. Seventy-seven percent (*n* =10) considered the items and content to be an "Adequate" representation of the domain for this developmental age range. One SME (8%) indicated that the number and content of items was "Somewhat Adequate." Two SMEs (15%) reported that the number and content of the items was an "Inadequate" representation of this domain for this developmental age range.

Social Behavior and Social Play

Domain relevance. The CVI for this domain was .45, which indicates that a majority of SMEs rated items as "Essential," on average. The percentage of SMEs who rated each item as "Essential" ranged from 69% (*n* = 9, CVR = .38) to 77% (*n* = 10, CVR=.54). For all five items, approximately 70% or more of the SMEs endorsed those

items as “Essential” but only two of had estimated CVRs that were significantly greater than zero. None of the SMEs indicated that an item was “Not Necessary.”

Age appropriateness. The percentage of SMEs who rated each item as “Very Appropriate” ranged from 77% (*n* = 10) to 85% (*n* = 11). SMEs who rated each item as “Somewhat Appropriate” for this developmental age ranged from 15% (*n* = 2) to 23% (*n* =3). None of the SMEs indicated an item was “Not Appropriate” for this developmental age range.

Measurement appropriateness. The percentage of SMEs who rated the methods of measurement “Very Appropriate” ranged from 38% (*n* = 5) to 69% (*n* = 9). SMEs who rated each method of measurement as “Somewhat Appropriate” ranged from 23% (*n* = 3) to 54% (*n* = 7). For each of the five items, one SME (8%) indicated that the method of measurement for the item was “Not Appropriate.”

Domain representation. Thirty-eight percent of SMEs (*n* = 5) considered the items and content to be an “Adequate” representation of the domain for this developmental age range. Forty-six percent (*n* = 6) of SMEs indicated that the number and content of items was “Somewhat Adequate.” Two SMEs (15%) reported that the number and content of the items was an “Inadequate” representation of this domain for this developmental age range.

Reading

Domain relevance. The CVI for this domain was .14, which indicates that a slight majority of SMEs rated items as “Essential,” on average. The percentage of SMEs who

rated each item as “Essential” ranged from 38% (*n* = 5, CVR = -.23) to 85% (*n* = 11, CVR=.69) but only one exceeded 69%. Although only one item had an estimated CVR that was significantly greater than zero, at least 92% of SMEs indicated that all items were “Essential” or “Useful, but Not Essential.” Only one (8%) SME indicated that an item was “Not Necessary.”

Age appropriateness. The percentage of SMEs who rated each item as “Very Appropriate” ranged from 31% (*n* = 4) to 92% (*n* = 12). SMEs who rated each item as “Somewhat Appropriate” for this developmental age ranged from 8% (*n* = 1) to 54% (*n* =7). Two SMEs (15%) indicated an item was “Not Appropriate” for this developmental age range.

Measurement appropriateness. The percentage of SMEs who rated the methods of measurement “Very Appropriate” ranged from 69% (*n* = 9) to 92% (*n* = 12). SMEs who rated each method of measurement as “Somewhat Appropriate” ranged from 0% (*n* = 0) to 15% (*n* = 2). For three of the five items, one SME (8%) indicated that the method of measurement for the item was “Not Appropriate.” For one of the items, two SMEs (15%) indicated it was “Not Appropriate.”

Domain representation. Sixty-two percent of SMEs (*n* = 8) considered the items and content to be an “Adequate” representation of the domain for this developmental age range. Fifteen percent (*n* = 2) of SMEs indicated that the number and content of items was “Somewhat Adequate.” Three SMEs (23%) reported that the number and content of the items was an “Inadequate” representation of this domain for this developmental age range.

Writing

Domain relevance. The CVI for this domain was .11, which indicates that a slight majority of SMEs rated items as "Essential," on average. The percentage of SMEs who rated each item as "Essential" ranged from 38% (n = 5, CVR = -.23) to 69% (n = 9, CVR=.38). None of the five items had estimated CVRs that were significantly greater than zero, but one item was rated as "Essential" by 69% of the SMEs. For four of the five items, at least 92% of SMEs indicated that the items were "Essential" or "Useful, but Not Essential." For the remaining item, three SMEs (23%) indicated that the item was "Not Necessary."

Age appropriateness. The percentage of SMEs who rated each item as "Very Appropriate" ranged from 23% (n = 3) to 69% (n = 9). SMEs who rated each item as "Somewhat Appropriate" for this developmental age ranged from 31% (n = 4) to 62% (n =8). For two of the items, two SMEs (15%) indicated the item was "Not Appropriate" for this developmental age range.

Measurement appropriateness. The percentage of SMEs who rated the methods of measurement "Very Appropriate" was 77% (n = 10) across all items. SMEs who rated each method of measurement as "Somewhat Appropriate" ranged from 15% (n = 2) to 23% (n = 3). For one of the five items, an SME (8%) indicated that the method of measurement for an item was "Not Appropriate."

Domain representation. Thirty-eight percent of SMEs (n = 5) considered the items and content to be an "Adequate" representation of the domain for this

developmental age range. Thirty-eight percent (*n* = 5) of SMEs indicated that the number and content of items was "Somewhat Adequate." Three SMEs (23%) reported that the number and content of the items was an "Inadequate" representation of this domain for this developmental age range.

Listener Responding by Function, Feature, and Class (LRFFC)

Domain relevance. The CVI for LRFFC was .48, which indicates that a large majority of SMEs rated items as "Essential," on average. The percentage of SMEs who rated each item as "Essential" ranged from 46% (*n* = 6, CVR = -.08) to 85% (*n* = 11, CVR=.69). For four of the five items, at least 69% of the SMEs endorsed those items as "Essential" for this domain, but only three of the items had estimated CVRs that were significantly greater than zero. Although only three of the items were statistically significant, none of the SMEs indicated that an item was "Not Necessary."

Age appropriateness. The percentage of SMEs who rated each item as "Very Appropriate" ranged from 54% (*n* = 7) to 85% (*n* = 11). SMEs who rated each item as "Somewhat Appropriate" for this developmental age ranged from 15% (*n* = 2) to 46% (*n* =6). None of the SMEs indicated an item was "Not Appropriate" for this developmental age range.

Measurement appropriateness. The percentage of SMEs who rated the methods of measurement "Very Appropriate" ranged from 46% (*n* = 6) to 85% (*n* = 11). SMEs who rated each method of measurement as "Somewhat Appropriate" ranged from 15% (*n*

= 2) to 38% (*n* = 5). For one of the five items, two SMEs (15%) indicated that the method of measurement for that item was "Not Appropriate."

Domain representation. Sixty-two percent of SMEs (*n* = 8) considered the items and content to be an "Adequate" representation of the domain for this developmental age range. Thirty-eight percent (*n* = 5) of SMEs indicated that the number and content of items was "Somewhat Adequate." None of the SMEs reported that the number and content of the items was an "Inadequate" representation of this domain for this developmental age range.

Intraverbal

Domain relevance. The CVI for this domain was .32, which indicates that a majority of SMEs rated items as "Essential," on average. The percentage of SMEs who rated each item as "Essential" ranged from 62% (*n* = 8, CVR = .23) to 77% (*n* = 10, CVR=.54). For two items, at least 69% of the SMEs provided an "Essential" rating, however only one item had an estimated CVRs that was significantly greater than zero. No item was rated as "Not Necessary."

Age appropriateness. The percentage of SMEs who rated each item as "Very Appropriate" ranged from 62% (*n* = 8) to 85% (*n* = 11). SMEs who rated each item as "Somewhat Appropriate" ranged from 15% (*n* = 2) to 38% (*n* =5). None of the SMEs indicated an item was "Not Appropriate" for this developmental age range.

Measurement appropriateness. The percentage of SMEs who rated the methods of measurement "Very Appropriate" ranged from 46% (*n* = 6) to 85% (*n* = 11). SMEs

who rated each method of measurement as "Somewhat Appropriate" ranged from 15% (*n* = 2) to 46% (*n* = 6). For one of the five items, one SME (8%) indicated that the method of measurement for an item was "Not Appropriate."

Domain representation. Thirty-one percent of SMEs (*n* = 4) considered the items and content within this domain to be an "Adequate" representation of the domain for this developmental age range. The majority, (*n* = 9, 60%) of SMEs indicated that the number and content of items was "Somewhat Adequate."

Classroom Routines and Group Skills

Domain relevance. The CVI for this domain was .26, which indicates that a slight majority of SMEs rated items as "Essential," on average. The percentage of SMEs who rated each item as "Essential" ranged from 38% (*n* = 5, CVR = -.23) to 85% (*n* = 11, CVR=.69). For three of the five items, at least nine SMEs endorsed those items as "Essential". Only one item had an estimated CVR that was significantly greater than zero. The majority of SMEs rated the majority of items as "Essential" or "Useful, but for one item three SMEs (23%) indicated that it was "Not Necessary."

Age appropriateness. The percentage of SMEs who rated each item as "Very Appropriate" ranged from 38% (*n* = 5) to 85% (*n* = 11). SMEs who rated each item as "Somewhat Appropriate" for this developmental age ranged from 15% (*n* = 2) to 31% (*n* =4). For one of the items, one SME indicated the item was "Not Appropriate" for this developmental age range. For one item, four SMEs (31%) indicated that it was "Not Appropriate."

Measurement appropriateness. The percentage of SMEs who rated the methods of measurement "Very Appropriate" ranged from 62% (*n* = 8) to 77% (*n* = 10). SMEs who rated each method of measurement as "Somewhat Appropriate" ranged from 15% (*n* = 2) to 31% (*n* = 4). For three of the five items, one SME (8%) indicated that the method of measurement for an item was "Not Appropriate." For two of the items, 15% of the SMEs (n = 2) indicated that the method of measurement for these two items was "Not Appropriate."

Domain representation. Considering the number and content of items within this domain, 46% (*n* = 6) considered the items and content to be an "Adequate" representation of the domain for this developmental age range. Thirty-eight percent (*n* = 5) of SMEs indicated that the number and content of items was "Somewhat Adequate." Two SMEs (15%) reported that the number and content of the items was an "Inadequate" representation of this domain for this developmental age range.

Linguistic Structure

Domain relevance. The CVI for this domain was .32, which indicates that a majority of SMEs rated items as "Essential," on average. The percentage of SMEs who rated each item as "Essential" ranged from 62% (*n* = 8, CVR = .23) to 85% (*n* = 11, CVR=.69). For two of the five items, at least nine SMEs endorsed those items as "Essential". Only one item had an estimated CVR that was significantly greater than zero. No item was rated as "Not Necessary."

Age appropriateness. The percentage of SMEs who rated each item as "Very Appropriate" ranged from 62% (*n* = 8) to 92% (*n* = 12). The percentage of SMEs who rated the items as "Somewhat Appropriate" ranged from 8% (*n* = 1) to 38% (*n* =5) for this developmental age ranged.

Measurement appropriateness. The percentage of SMEs who rated the methods of measurement "Very Appropriate" ranged from 54% (*n* = 7) to 69% (*n* = 9). SMEs who rated each method of measurement as "Somewhat Appropriate" ranged from 31% (*n* = 4) to 38% (*n* = 5). For one of the five items, one SME (8%) indicated that the method of measurement for an item was "Not Appropriate."

Domain representation. Sixty-nine percent of SMEs (*n* = 9) considered the items and content to be an "Adequate" representation of the domain for this developmental age range. Twenty-three percent (*n* = 3) of SMEs indicated that the number and content of items was "Somewhat Adequate." One SME (8%) reported this was an "Inadequate" representation of this domain for this developmental age range.

Math

Domain relevance. The CVI for this domain was .26, which indicates that a slight majority of SMEs rated items as "Essential," on average. The percentage of SMEs who rated each item as "Essential" ranged from 54% (*n* = 7, CVR = .08) to 77% (*n* = 10, CVR=.54). Two of the five items had estimated CVRs that were significantly greater than zero and were rated as "Essential" by at least 69% of SMEs. All SMEs indicated that four

of the five items were “Essential” or “Useful, but Not Essential” (*n* = 13, 100%). For one of the items, two SMEs (15%) indicated that the item was “Not Necessary.”

Table 4.3.

VB-MAPP Level 3 Content Validity Values

Domain	CVI	Minimum CVR	Maximum CVR	No. of Significant Items*
Mand	.51	.23	.69	3
Tact	.38	.08	.54	2
Listener Responding	.32	.08	.54	1
Visual Perceptual Skills & Matching-To-Sample	-.09	-.54	.33	0
Independent Play	.38	-.08	.69	2
Social Behavior & Social Play	.45	.38	.54	2
Reading	.14	-.23	.69	1
Writing	.11	-.23	.38	0
Listener Responding by Function, Feature, & Class	.48	-.08	.69	3
Intraverbal	.32	.23	.54	1
Classroom Routines & Group Skills	.26	-.23	.69	1
Linguistic Structure	.32	.08	.69	1
Math	.26	.08	.54	2

* upper-tailed $p < .05$ for 13 SMEs

Note. CVI = Content validity index; Minimum CVR = Minimum content validity ratio among the five items for each domain; Maximum CVR = Maximum content validity ratio among the five items for each domain.

Age appropriateness. The percentage of SMEs who rated each item as “Very Appropriate” ranged from 54% (*n* = 7) to 85% (*n* = 11). SMEs who rated each item as “Somewhat Appropriate” for this developmental age ranged from 15% (*n* = 2) to 46% (*n* =6). None of the SMEs indicated an item was “Not Appropriate” for this developmental age range.

Measurement appropriateness. The percentage of SMEs who rated the methods of measurement “Very Appropriate” ranged from 69% (*n* = 9) to 77% (*n* = 10). SMEs who rated each method of measurement as “Somewhat Appropriate” ranged from 23% (*n*

= 3) to 31% (*n* = 4). None of the SMEs indicated that the method of measurement for any item was "Not Appropriate."

Domain representation. Sixty-two percent of SMEs (*n* = 8) considered the items and content to be an "Adequate" representation of the domain for this developmental age range. The remaining SMEs (*n* = 5, 38%) indicated that the number and content of items was "Somewhat Adequate" for this developmental age range.

General Commentary

In many cases, SMEs provided additional comments related to item ratings and/or suggestions for revisions that were consistent across domains and milestone levels. SMEs expressed the most concerns regarding the age appropriateness of the items. The commentary indicated that some criteria did not match age level of zero to 18 months; that is, SMEs noted that for some items the behavior criterion was too advanced or too rudimentary for the developmental age range. Many SMEs also reported that the age range was too wide to accurately assess skill level. Recommendations were also made to add items to better assess skills of children at the lower end of the age range, and to include activities/tasks that may be more common for this age range (i.e., lack of exposure). The availability of normative data to use as a criterion for making comparisons to a neurotypical child was widely and consistently recommended by SMEs.

With respect to the methods of measurement used and domain representation, SMEs offered several recommendations. Some methods were not considered appropriate based on the developmental age range, and the domains could include additional behaviors to meet the criteria and, ultimately, make the assessment more comprehensive.

Clarification on assessment procedures, scoring protocols, and environmental arrangements were also included in the SMEs' commentary. That is, SMEs indicated that more standardized procedures and clarification of assessment procedures are needed to ensure consistency across examiners and more accurate evaluations.

CVR by Method of Measurement

The CVRs were compared across methods of measurement (i.e., direct testing, observation, either direct testing or observation, or timed observation). On average, those items requiring either direct testing or observation had the highest CVRs (M = .42, min = .00, max = 1.00). The average CVR was .28 (min = -.23, max = .28) for items requiring observation, .29 (min = -.54, max = 1.00) for items requiring direct testing, and .30 (min = -.69, max = .69) for items requiring timed observation. The relationship between the method of measurement and statistical significance was .11 (V), which suggest there is no meaningful relationship between these two variables. The CVR values were the same or very similar regardless of the method of measurement used for the items.

The method of measurement appropriateness ratings was also compared across methods of measurement. Ratings were recoded based on the percentage of SMEs indicating that the method of measurement was "Very Appropriate." Popham's (1992) recommendation of 70% SME endorsement was used as the threshold for classification. Due to the number of SMEs in the study, the threshold was 69% (i.e., nine out of 13 SMEs) to approximate Popham's (1992) recommendation. The estimated Cramer's V was .50 (df = 3, $p < .001$), which suggests there was a relatively strong association between the method of measurement and the appropriateness rating (Rea & Parker, 1992). The method of measurement that SMEs tended to rate as "Very Appropriate" most often was

direct testing. Out of 87 items requiring direct testing, 71 (82%) were rated as "Very Appropriate" by at least 9 SMEs. The method of measurement for which SMEs rated "Very Appropriate" least frequently was timed observation. Out of 30 items, only five (17%) were rated as "Very Appropriate" by at least 9 SMEs. A summary of the ratings by method of measurement is presented in Table 4.4.

Early Echoic Skills Assessment (EESA) Results

The calculated CVI for the EESA across groups was .35. This indicates that on average, the majority of raters indicated that the groups of items were "Essential." Following is a summary of each group within the EESA, which includes domain relevance, age appropriateness, measurement appropriateness, and domain representation.

Table 4.4.

Summary of Appropriateness Ratings by Methods of Measurement

	High Appropriateness		Medium or Low Appropriateness	
Method of Measurement	*No. of Items*	*Percentage**	*No. of Items*	*Percentage**
Timed Observation (TO)	5	17	25	83
Direct Testing (T)	71	82	16	18
Observation (O)	19	70	8	30
Either T or O	13	50	13	50

* Row percentages

Note. High Appropriateness = More than 69% of SMEs (9 or more out of 13) rated the method of measurement as "Very Appropriate", Medium or Low Appropriateness = 69% or fewer SMEs (8 or fewer out of 13) rated the method of measurement as "Very Appropriate."

Group-level response distributions and CVRs (with associated statistical significance classifications) are reported in Table 4.5

Group 1: Simple and Reduplicated Syllables

Domain relevance. The CVR for Group 1 was .69, which indicates that a large majority of SMEs rated this group of items as "Essential." Eighty-five percent (n = 11) rated this group of items as "Essential." This group of items had an estimated CVR that was significantly greater than zero. Fifteen percent of SMEs (n = 2) rated this group of items as "Useful, but Not Essential." None of the SMEs rated this group of items as "Not Necessary."

Age appropriateness. Regarding the developmental age appropriateness, 85% of SMEs (n = 11) rated this group of items as "Very Appropriate." Fifteen percent of SMEs (n = 2) rated this group of items as "Somewhat Appropriate" for this developmental age range. None of the SMEs indicated that this group of items was "Not Appropriate" for this developmental age range.

Prompt appropriateness. As noted previously, each group within the EESA has a particular set of prompts (i.e., discriminative stimulus, S^D) in which the examinee's best response is scored. Group 1 has a total of 25 prompts that include simple and reduplicated syllables. For this, 54% (n = 7) of SMEs rated prompts within Group 1 as "Essential." Thirty-eight percent of SMEs (n = 5) rated these prompts within Group 1 as "Useful, but Not Essential." One SME (8%) rated prompts within this group as "Not Necessary."

Group 2: 2-Syllable Combinations

Domain relevance. The CVR for Group 2 was .54, which indicates that a large majority of SMEs (n = 10, 77%) rated this group of items as "Essential." This group of

items had an estimated CVR that was significantly greater than zero. Twenty-three percent of SMEs (*n* = 3) rated this group of items as "Useful, but Not Essential." None of the SMEs rated this group of items as "Not Necessary."

Age appropriateness. Regarding the developmental age appropriateness, 77% of SMEs (*n* = 10) rated this group of items as "Very Appropriate." Twenty-three percent of SMEs (*n* = 3) rated this group of items as "Somewhat Appropriate" for this developmental age range. None of the SMEs indicated that this group of items was "Not Appropriate" for this developmental age range.

Prompt appropriateness. Group 2 has a total of 30 prompts (i.e., S^D) that include two-syllable combinations. For this group, 46% (*n* = 6) of SMEs rated prompts within Group 2 as "Essential." Forty-six percent of SMEs (*n* =6) rated these prompts within Group 2 as "Useful, but Not Essential." One SME (8%) rated prompts within this group as "Not Necessary."

Group 3: 3-Syllable Combinations

Domain relevance. The CVR for Group 3 was .38, which indicates that a majority of SMEs (*n* = 9, 69%) rated this group of items as "Essential," on average. However, using the published critical values for statistical significance, this group of items did not have an estimated CVR that was significantly greater than zero. Twenty-three percent of SMEs (*n* = 3) rated this group of items as "Useful, but Not Essential." One SME (8%) rated this group of items as "Not Necessary."

Age appropriateness. Regarding the developmental age appropriateness, 69% of SMEs (*n* = 9) rated this group of items as "Very Appropriate." Thirty-one percent of SMEs (*n* = 4) rated this group of items as "Somewhat Appropriate" for this developmental age range. None of the SMEs indicated that this group of items was "Not Appropriate" for this developmental age range.

Prompt appropriateness. Group 3 has a total of 30 prompts (i.e., S^D) that include three-syllable combinations. For this group, 46% (*n* = 6) of SMEs rated prompts within Group 3 as "Essential." Thirty-eight percent of SMEs (*n* =5) rated these prompts within Group 3 as "Useful, but Not Essential." Two SMEs (15%) rated prompts within this group as "Not Necessary."

Group 4: Prosody: Spoken Phrases

Domain relevance. The CVR for Group 4 was .08, which indicates that a slight majority of SMEs (*n* = 7, 54%) rated this group of items as "Essential," on average. However, this group of items did not have an estimated CVR that was significantly greater than zero. Forty-six percent of SMEs (*n* = 6) rated this group of items as "Useful, but Not Essential." None of the SMEs rated this group of items as "Not Necessary."

Age appropriateness. Regarding the developmental age appropriateness, 69% of SMEs (*n* = 9) rated this group of items as "Very Appropriate." Twenty-three percent of SMEs (*n* = 3) rated this group of items as "Somewhat Appropriate" for this developmental age range. One SME (8%) indicated that this group of items was "Not Appropriate" for this developmental age range.

Prompt appropriateness. Group 4 has a total of 10 prompts (i.e., items) that assess an examinee's ability to imitate pitch, loudness, and vowel duration. For Group 4, 69% (*n* = 9) of SMEs rated these prompts as "Essential." Twenty-three percent of SMEs (*n* =3) rated these prompts within Group 4 as "Useful, but Not Essential." One SME (8%) rated prompts within this group as "Not Necessary."

Group 5: Prosody: Other Contexts

Domain relevance. The CVR for Group 5 was .08, which indicates that a slight majority of SMEs (*n* = 7, 54%) rated this group of items as "Essential," on average. However, using the published critical values for statistical significance, this group of items did not have an estimated CVR that was significantly greater than zero. Forty-six percent of SMEs (*n* = 6) rated this group of items as "Useful, but Not Essential." None of the SMEs rated this group of items as "Not Necessary."

Age appropriateness. Regarding the developmental age appropriateness, 62% of SMEs (*n* = 8) rated this group of items as "Very Appropriate." Thirty-one percent of SMEs (*n* = 4) rated this group of items as "Somewhat Appropriate" for this developmental age range. One SME (8%) indicated that this group of items was "Not Appropriate" for this developmental age range.

Prompt appropriateness. Group 5 has a total of five prompts (i.e., S^D) that assess an examinee's ability to imitate pitch, loudness, and vowel duration in other contexts. For Group 5, 62% (*n* = 8) of SMEs rated these prompts as "Essential." Thirty-one percent of

Table 4.5.

EESA Content Validity Ratios and Percentage of Responses

		Domain Relevance			Age Appropriateness			Prompt Appropriateness		
Group	CVR	*N*	*U*	*E*	*NA*	*SA*	*VA*	*NA*	*SA*	*VA*
Group 1: Simple and reduplicated syllables	.69*	0	15	85	0	15	85	8	38	54
Group 2: 2-syllable combinations	.54*	0	23	77	0	23	77	8	46	46
Group 3: 3-syllable combinations	.38	8	23	69	0	31	69	15	38	46
Group 4: Prosody: spoken phrases	.08	0	46	54	8	23	69	8	23	69
Group 5: Prosody: other contexts	.08	0	46	54	8	31	62	8	31	62

* upper-tailed $p < .05$

Note. CVR = Content validity ratio; "E" = Essential; "U" = Useful, but Not Essential; "N" = Not necessary; "VA" = Very Appropriate; "SA" = Somewhat Appropriate; "NA" = Not Appropriate.

SMEs (n =4) rated these prompts within Group 4 as "Useful, but Not Essential." One SME (8%) rated prompts within this group as "Not Necessary."

Domain representation. Considering the number and content of items within the entire EESA, 77% (n = 10) considered the items and content to be an "Adequate" representation of echoic skills for this developmental age range. Fifteen percent (n = 2) of SMEs indicated that the number and content of items was "Somewhat Adequate." One SME (8%) reported that the number and content of the items was an "Inadequate" representation of this domain for this developmental age range.

VB-MAPP Barriers Assessment Result

The calculated CVI for the Barriers Assessment across categories was .60. This indicates that on average, a large majority of raters indicated that the categories of barriers included in this assessment were "Essential." Following is a summary of the Barriers Assessment, which includes domain relevance and measurement

appropriateness. Category-level response distributions and CVRs (with associated statistical significance classifications) are reported in Table 4.6.

Domain Relevance

The CVR across barrier categories ranged from was .08 (*n* = 7, 54%) to .85 (*n* = 12, 92%), which indicates that a majority of SMEs rated barrier categories as "Essential," on average. For 21 of the 24 categories, approximately 70% or more of the SMEs endorsed those categories as "Essential" for this assessment and 18 of these had estimated CVRs that were significantly greater than zero. SMEs who indicated that items were "Useful, but Not Essential" ranged from 8% (*n* = 1) to 46% (*n* = 6). None of the SMEs indicated that any of the barrier categories was "Not Necessary."

Measurement Appropriateness

As previously mentioned, each barrier is scored on a Likert-type scale ranging from 0 = No problem to 4 = Severe Problem. Each score has associated examples to assist the examiner in determining the best representation of the behavior. SMEs who rated the method of measurement as "Very Appropriate" ranged from 46% (n = 6) to 69% (n = 9) to across barriers. SMEs who rated the method of measurement across barriers as "Somewhat Appropriate" ranged from 15% (*n* = 2) to 31% (*n* = 4). Regarding SMEs who rated the method of measurement as "Not Appropriate," percentages ranged from 0% (*n* = 0) to 23% (*n*= 3).

Domain Representation

Considering the number of types of barriers within the entire Barriers Assessment, 46% (*n* = 6) considered the items and types of barriers to be an "Adequate" representation of barriers that might impede a child's progress for this developmental age

Table 4.6.

Barriers Assessment Content Validity Ratios and Percentage of Responses

		Domain Relevance			Measurement Appropriateness		
Barrier	CVR	*N*	*U*	*E*	*NA*	*SA*	*VA*
Negative Behaviors	.85*	0	8	92	8	31	62
Instructional Control (Escape and Avoidance of Instructional Demands)	.85*	0	8	92	8	31	62
Absent, Weak, or Impaired Mand Repertoire	.85*	0	8	92	23	15	62
Absent, Weak, or Impaired Tact Repertoire	.69*	0	15	85	23	15	62
Absent, Weak, or Impaired Motor Imitation	.85*	0	8	92	15	23	62
Absent, Weak, or Impaired Echoic Repertoire	.54*	0	23	77	23	31	46
Absent, Weak, or Impaired Visual Perceptual and Matching-to-Sample	.54*	0	23	77	23	15	62
Absent, Weak, or Impaired Listener Repertoire (e.g. LD, LRFFC)	.85*	0	8	92	23	15	62
Absent, Weak, or Impaired Intraverbal Repertoire	.38	0	31	69	23	15	62
Absent, Weak, or Impaired Social Skills	.69*	0	15	85	8	31	62
Prompt Dependent	.69*	0	15	85	8	23	69
Scrolling Responses	.38	0	31	69	8	31	62
Impaired Scanning Skills	.69*	0	15	85	8	23	69
Failure to Make Conditional Discriminations (C^D_S)	.69*	0	15	85	15	15	69
Failure to Generalize	.69*	0	15	85	15	15	69
Weak or Atypical Motivating Operations (MOs)	.69*	0	15	85	8	23	69
Response Requirement Weakens the MO	.54*	0	23	77	8	23	69
Reinforcement Dependent	.69*	0	15	85	8	23	69
Self-Stimulation	.69*	0	15	85	8	25	67
Articulation Problems	.08	0	46	54	0	31	69
Obsessive-Compulsive Behavior	.23	0	38	62	8	31	62
Hyperactive Behavior	.23	0	38	62	8	23	69
Failure to Make Eye Contact, or Attend to People	.38	0	31	69	8	31	62
Sensory Defensiveness	.54*	0	23	77	8	23	69

* upper-tailed $p < .05$

Note. CVR = Content validity ratio; "E" = Essential; "U" = Useful, but Not Essential; "N" = Not necessary; "VA" = Very Appropriate; "SA" = Somewhat Appropriate; "NA" = Not Appropriate.

range. Forty-six percent (*n* = 6) of SMEs indicated that the number and types of barriers was "Somewhat Adequate." One SME (8%) reported that the number and types of

barriers was an “Inadequate” representation of barriers for this domain within this developmental age range.

Conclusion

Overall, the evidence is generally supportive for the use of the VB-MAPP. For most domains across the VB-MAPP, the CVRs and CVIs were positive, which indicates that a majority of SMEs rated items as “Essential.” However, only 31% of the CVRs were statistically significant. Although a majority of CVRs were not statistically significant, a majority of SMEs rated items as “Essential” or “Useful, but Not Essential.” For example, 100% of SMEs rated 140 of the 170 items as “Essential” or “Useful, but Not Essential.” For 22 of the 170 items, 92% of SMEs rated the items as “Essential” or “Useful, but Not Essential” while 85% rated the remaining 5 items of the 170 similarly. For the remaining three items, 77% provided the “Essential” or “Useful, but Not Essential” rating. Regarding age appropriateness of the items the VB-MAPP, the distribution was nearly identical to the domain relevance ratings. For 140 of the 170 items, 100% of the SMEs rated the age appropriateness of the item as “Very Appropriate” or “Somewhat Appropriate.” For the remaining items, 77% or more rated the items as “Very Appropriate” or “Somewhat Appropriate.” In regard to the method of measurement appropriateness, 92% or more SMEs rated the method of measurement for 151 items as “Very Appropriate” or “Somewhat Appropriate.” For the remaining items, 75% or more SMEs rated the method of measurement as “Very Appropriate” or “Somewhat Appropriate.” Considering the number and content of items, at least 92% of SMEs rated the domain representativeness as “Adequate” or “Somewhat Adequate” for 162 of the 170 items. These findings demonstrate that SMEs generally supported the

domain relevance, age appropriateness, method of measurement appropriateness, and domain representation of the items within the VB-MAPP Milestones.

For groups of items within the EESA, 92% or more of SMEs rated the groups as "Essential" or "Useful, but Not Essential." Regarding the age appropriateness of the groups of items, 92% or more rated the groups as "Very Appropriate" or "Somewhat Appropriate." Eighty-five percent or more of SMEs rated the prompts for the items within the groups as "Very Appropriate" or "Somewhat Appropriate." For the Barriers Assessment, 100% of the SMEs rated the categories as "Essential" or "Useful, but Not Essential." Regarding the method of measurement for the categories, 77% or more indicated that it was "Very Appropriate" or "Somewhat Appropriate."

CHAPTER FIVE

Discussion

Teaching techniques and interventions based in ABA are the most empirically supported treatments to address the skill deficits in children with ASD (Axelrod, McElrath, & Wine, 2012; Lovaas, 1987). Treatment plans are developed based on identified skill deficits, which are derived from the results of an individual's performances on the collection of items that make up instruments. Treatment planning entails selecting goals and identifying interventions that ultimately impact the trajectory of a child's skill development. It is imperative that researchers and practitioners utilize instruments that have evidence to support their use as the scores and information obtained from these instruments are used to make decisions about specific skill deficits. Evaluating the quality and content of items in an instrument is an important first step in increasing confidence in one's choice of an instrument. The content of the items is the focus of evaluation in order to provide evidence for content validity. Under the overarching umbrella of construct validity, the items included on an instrument can be viewed as a sample of all possible items to measure that construct. Researchers and practitioners are limited by the number of items they can administer to measure behavior, thus the specific items within a given instrument should have strong evidence supporting their use. Therefore, evaluating the characteristics (e.g., content, representation) of the items included on an assessment is critical because the responses to the items are used as the basis for decisions about the underlying phenomenon of interest.

According to Padilla (2019), approximately 80% of ABA professionals reported administering the VB-MAPP, which is a criterion-referenced instrument developed within the ABA framework, for curriculum development and treatment planning for children with ASD. Despite being the most widely used instrument, there are currently no studies that focus explicitly on collecting or evaluating validity evidence for the VB-MAPP (Padilla et al., 2019). Thus, the VB-MAPP is in need of evidence supporting its use because thousands of researchers and practitioners are reportedly administering the instrument as the basis for their decisions about children. Collecting and evaluating content validity evidence was precisely the focus of the current study.

The design of the current study was modeled after Usry et al. (2017), which examined the content validity of the ABLLS-R, and also incorporated recommendations for improvement based on limitations they presented. The limitations Usry and colleagues (2017) identified in their study included the small number of SMEs, lack of geographic representation of the SMEs, wide range of inclusion criteria to qualify as an SME, and the need for the evaluation of CVRs using critical values established by the work of Wilson et al. (2012). Each of these limitations was directly addressed in the design of the current study. A panel of 13 SMEs from five geographic regions in the United States evaluated the VB-MAPP items with respect to each item's domain relevance, age appropriateness, method of measurement appropriateness, and domain representation. These evaluations (i.e., content validity evidence) were conducted for the VB-MAPP Milestones, EESA, and Barriers Assessment. The sample size was determined based on recommendations from Raymond and Reid (2001). The inclusion criteria were based on recommendations from the *Standards for Educational and Psychological*

Testing (AERA et al., 2014), inclusion criteria adapted from Usry et al. (2017) as well as the assessor qualifications stipulated in the *VB-MAPP Guide* (Sundberg, 2014). Ultimately, these criteria were applied to identify SMEs who had a specific area of expertise, professional credentials, and experience with the target population for the VB-MAPP.

The results of the current study provide moderate to strong content-related validity evidence of the VB-MAPP for nearly all of the domains across levels of the Milestones Assessment, EESA, and Barriers Assessment. The validity evidence for each of the assessments and the implications for their use are discussed in more detail below. This chapter concludes with general discussion and recommendations, limitations, and future research.

Milestones

Domain Relevance

There was moderate to strong evidence supporting domain relevance for the majority of domains across levels. Mand, Tact, Listener Responding, Visual Perceptual Skills and Matching-to-Sample, Independent Play, and Social Behavior and Social Play are measured across all three levels within the VB-MAPP. There was strong support for Mand and Tact across all three levels, which is not surprising given the abundance of research for these two verbal operants in the literature. DeSouza, Akers, and Fisher (2017) conducted a systematic review of studies that focused on Skinner's verbal operants in interventions for children with ASD, which was an extension of Sundberg and Michael (2011). DeSouza and colleagues (2017) found that the majority of studies

targeted mands (53%), followed by tact (33%), intraverbal (23%), and echoic (2%), which was consistent with Sundberg and Michael (2001). Mand studies included in the review targeted acquisition, generalization, and maintenance of mands using a variety of intervention methods. Tact studies included comparison of different teaching strategies, data collection methods, prompting strategies, as well as error correction, instructional feedback, and different arrangements of reinforcement. Of the studies involving tact training, the majority of studies focused on teaching tacts using direct teaching methods. Because mands and tacts are the most researched verbal operant (DeSouza et al., 2017), it is important that the current study showed strong evidence for these domains in the VB-MAPP.

There was moderate to strong support for Listener Responding, Independent Play, and Social Behavior and Social Play across all three levels. There was stronger support for Listener Responding in Level 1 and Social Behavior and Social Play in Level 3, which may be due to a relationship between these domains with developmental age. That is, the higher average relevance rating in Level 3 could be due to higher developmental age and related exposure to social settings. Within the Social Behavior and Social Play domain, SMEs commented on the need for instruction clarification, specific item definitions, and a reduction in the number of skill demonstrations required and/or revising those items to assess for generalization. Additionally, in DeSouza, et al. (2017), the authors noted that many of the mand and tact studies often integrated listener responding teaching components to promote skill acquisition.

Some domains are measured in two of three levels – Echoic (Level 1 & 2), Motor Imitation (Level 1 & 2), LRFFC (Level 2 & 3), Intraverbal (Level 2 & 3), Classroom

Routines & Groups Skills (Level 2 & 3), and Linguistic Structure (Level 2 & 3). There was moderate to strong evidence for Motor Imitation, LRFFC, Intraverbal, and Linguistic Structure, but limited to moderate evidence for Echoic and Classroom Routines & Group Skills. In the review conducted by DeSouza and colleagues (2017), intraverbals had the third highest percentage of studies. Forty out of 172 studies involved intraverbal responses as the primary outcome variable. Intraverbal training included transfer-of-stimulus control, effects of antecedents or consequences, and response variability. Sundberg and Sundberg (2011) stated that intraverbals are controlled by multiple verbal antecedent stimuli, which underscores the importance of the conditions under which complex intraverbals are emitted (DeSouza et al., 2017). The moderate to strong evidence provided for the Intraverbal domain is an important contribution of this work.

There is a very small body of literature targeting echoics. DeSouza et al. (2017) identified only four studies published since 2001 that targeted echoics as the primary outcome variable. Given that the echoic skill can facilitate the use of their controlling prompts to teach other verbal operants (DeSouza, et al. 2017), this is an area of improvement for the VB-MAPP. It is important to note that domain relevance ratings for Echoic may have been influenced by other concerns reported by SMEs such as age appropriateness, age range, length of the assessment for this age range, and scoring difficulties. Additionally, the Echoic domain is completely dependent on the EESA scores so further evaluation of the EESA items is recommended to ensure this domain is measured accurately.

In regards to Classroom Routines and Group Skills, several SMEs commented on the inability to assess this skill due to an individual's potential lack of exposure,

behavioral expectations, and developmental age. That is, Level 2 is designed for children 18 to 30 months of age so it may be unlikely that they have experience in educational settings. All of these factors make it difficult for this skill to be accurately and objectively assessed. The evidence for this domain was stronger in Level 3 than Level 2, the VB-MAPP authors may consider revisions to the Classroom Routines and Group Skills domain or possibly only include the domain at Level 3.

Age Appropriateness

There is moderate to strong evidence across the vast majority of domains across all levels for age appropriateness. There was limited support for Classroom Routines and Group Skills (Level 2), Visual Perceptual Skills and Matching-to-Sample (Level 3), and Writing (Level 3). Based on the SME comments, the lower age appropriateness rating for Classroom Routines and Group Skills is due to the age range being too wide for the embedded assessment tasks and criteria (e.g., tasks may not be appropriate for children at the lower end of the developmental age range). For the Visual Perceptual Skills Matching-to-Sample domain, the percentage of SMEs who rated age appropriateness as "Very Appropriate" decreased as level increased. The lower rating may have been due to what SMEs described as assessment procedures lacking operational definitions (e.g., "messy array") as well as lack of exposure to tasks or activities embedded within assessment items. SMEs also expressed concerns related to the wide range for developmental age and lack of normed comparisons. For the Writing domain, the SMEs reported that the items were too advanced for this developmental age range. Children may begin to imitate writing early as two years old, but emergent writing skills tend to

develop toward the upper bound of Level 3's target developmental age range (Dennis & Votteler, 2013; Rowe & Neitzel, 2010).

Method of Measurement Appropriateness

There was moderate to strong support for the method of measurement appropriateness for the vast majority of domains across all three levels. There was limited support for Independent Play (Level 1) and Social Behavior and Social Play (Level 1 & 2). Interestingly, the strength of support for method of measurement appropriateness generally increased as level increased. In addition, the domains with high average domain relevance ratings tended not to have high average method of measurement ratings, on average.

Overall, the use of direct testing was the most strongly supported method of measurement by SMEs, and it was also the most used method of measurement (87 of 170 items, 51%). Some domains relied on direct testing more heavily than other methods of measurement. That is, direct testing was used for all 10 Echoic items, eight of 10 Intraverbal items, 13 of 15 Listener Responding items, nine of 10 LRFFC items, all five Math items eight of 10 Motor Imitation items, four of five Reading items, 11 of 15 Tact items, and all five Writing items. Out of 87 items, 71 (82%) received a high percentage of SMEs reported the method of measurement as "Very Appropriate." These ratings suggest that direct testing is considered a strong and generally appropriate method for collecting data on skills across numerous domains.

For timed observations in the VB-MAPP, assessors determine whether a skill is exhibited within a certain timeframe. Generally, timed observation received the lowest level of support from SMEs and tended to be used more often within four domains. That

is, of the 30 items that required timed observation, 26 of the items appear in the Social Behavior and Social Play (10), Independent Play (6), Mand (5), and Spontaneous Vocal Behavior (5) domains. Timed observation items tended to be rated much lower in appropriateness across domains, which may have confounded the domain-specific results or evidence. Of the 30 timed observation items, only five had a percentage of "Very Appropriate" ratings from SMEs. The lower ratings may be attributed to the length of time required to observe (e.g., 30 or 60 minutes) for a particular skill or there were multiple skills that needed to be simultaneously assessed with varying timeframes. Additionally, other criterion-referenced tests based in ABA (e.g., ABLLS-R, PEAK) do not typically use this type of method of measurement. These findings suggest that timed observation may pose a challenge for researchers and practitioners administrating the VB-MAPP, particularly for the domains that rely heavily on timed observations (i.e., Social Behavior and Social Play and Independent Play). Such challenges may negatively affect the quality of decisions made about these domains.

The remaining items specified the use of observation or allowed the assessor to use either observation or direct testing. The appropriateness ratings were high for 50% of the 26 items requiring either observation or direct testing and 70% of the 27 items that required observation. A possible improvement in terms of method of measurement would be to specify which method should be employed for those items that currently allow assessors to choose because both observation and direct testing each have high support.

In general, reasons given by SMEs for lower method of measurement appropriateness rating tended to be due to misalignment between the method of measurement and skill being assessed, length of timed observations being too long, and

item wording conflicting with the specified method of measurement. That is, the protocol specifies direct testing although the wording of the item indicates that data may also be obtained from another source (e.g., caregiver-provided information). Future editions of VB-MAPP may benefit from reconsidering or restructuring the use of timed observation and/or providing clarity on specific methods to use on items for which multiple methods are currently allowed.

Domain Representation

Overall, the evidence for domain representation was not as strong as for other areas evaluated by SMEs. There was limited to moderate evidence for domain representation for the majority of domains across levels. In Level 1, seven of the nine domains were rated as "Adequate" by a majority of SMEs (one of nine exceeding 69%). In Level 2, seven of 12 domains were rated as "Adequate" by a majority of SMEs, with only one domain at 69%. In Level 3, 10 of the 13 domains were rated as "Adequate" by a majority of SMEs with four at or above 69%. These results suggest that the number and/or content of the items may not adequately represent their domains across levels. In general, domains that had a high domain relevance rating tended not to have high domain representation ratings. This suggests that the items within certain domains were generally considered relevant to the domain, but the number of items was insufficient. Scores for domains that are underrepresented may not accurately reflect a child's true skill level. As a result, practitioners may develop curriculum guides and/or treatment plans that may not specifically address the child's individualized needs. Adding or revising items may improve domain representation for domains where ratings were lower.

EESA

Although the EESA is a separate instrument embedded within the VB-MAPP to measure the Echoic domains, SMEs were asked to evaluate it on the same properties. On average, 68% of SMEs rated the groups of items as "Essential," which indicates moderate support for its domain relevance. Overall SME ratings for the EESA generally align with the Echoic domain representation ratings. There was strong support for age appropriateness for the EESA. Interestingly, the age appropriateness ratings for EESA tended to slightly decrease as the complexity of skills assessed increased, which aligns with the Echoic ratings from the Milestones Assessment for age appropriateness. Regarding prompt appropriateness (S^D), there was moderate support across groups of items for the EESA. There was strong support for domain representation for the EESA with 77% of SMEs rating it as "Adequate." That said, SMEs commented that specific item content may be outside the expertise of behavior analysts. Thus, consulting with a speech-language pathologist may result in more accurate inferences regarding echoic skills.

Barriers Assessment

There was strong evidence for domain relevance across the 24 barriers assessed. On average, 80% of SMEs rated each barrier as "Essential" for evaluating barriers to assessment. Age appropriateness was not evaluated by SMEs for the Barriers Assessment. Regarding method of measurement, there was moderate support for appropriateness across barriers within this assessment. This may be attributed to the multiple considerations for evaluating barriers. A Likert-type response scale is provided as a general guide to measure the severity of the problem behavior. In addition, each

barrier also has its own loosely operationally defined severity rating scale for problem behaviors that may impede assessment and/or skill acquisition. There was limited support for domain representation for the Barriers Assessment. Fewer than half of the SMEs (46%) rated it as "Adequate." Given that the other areas had strong support, this may suggest that the assessment could include additional barriers for a more comprehensive assessment. Comments from SMEs indicate that the Barriers Assessment could be revised by changing criteria and/or operational definitions that would result in a more representative assessment of barriers.

General Discussion and Recommendations

Overall, the content validity evidence for the VB-MAPP Milestones, EESA, and Barriers Assessment was moderate to strong across the evaluated areas although there were areas with limited or conflicting support. Evidence for domain relevance was moderate to strong for 91% of domains (31 out of 34) measured across Milestone levels. The domains with the strongest overall support across levels were also the most researched verbal operants – Mand, Tact, and Intraverbal (DeSouza et al., 2017). For *all* domains, the vast majority of SMEs (85% or more) rated all items as "Essential" or "Useful, but Not Essential," which indicates that items within the VB-MAPP are necessary to some degree to measure the behavior constructs. The same pattern generally held for age appropriateness and method of measurement appropriateness ratings across the evaluated areas within the Milestones, EESA, and Barriers Assessment. Regarding domain representation, there was moderate to strong support for 68% (23 out of 34) of the domains across Milestone levels. For all domains, the vast majority of SMEs (77% or more) rated domain representation as "Adequate" or "Somewhat Adequate," which

suggest that items adequately represent their targeted behavior construct to some degree. Domains with higher relevance ratings tended to have higher age appropriateness ratings. Domains with higher method of measurement ratings tended to have higher domain representation ratings. The evaluated areas did not follow a discernable pattern otherwise. Independent Play in Level 3 was the only domain that had strong evidence across all four categories.

Domain relevance and domain representation are two of the four content validity areas identified by Sireci (1998). In this study, there were more domains with high relevance ratings than there were domains with high representation ratings. This suggests that item content within the VB-MAPP is important but more items may be necessary to provide a more comprehensive assessment of the targeted behavior constructs. Overall, the domain relevance evidence is considered moderate to strong but domain representation is mixed. Thus, the evidence suggests that the scores of the VB-MAPP provides information relevant to the target behaviors of interest but may not fully represent the construct for a few domains. When the VB-MAPP is used by itself, researchers and practitioners can have reasonable confidence in the results for many domains but should exercise caution for some domains across levels. That said, it is recommended that the VB-MAPP be used in conjunction with other sources of assessment information, which is recommended for assessment in general. According to the National Council of Measurement in Education (1995), "Persons who interpret, use, and communicate assessment results have a professional responsibility to use multiple sources and types of relevant information about persons or programs whenever making educational decisions" (Section 6.7). Moreover, AERA et al. (2014) states that "a

decision...that will have major impact on a student should not be made on the basis of a single test score. Other relevant information should be taken into account if it will enhance the overall validity of the decision (Standard 13.7)."

The results of the current study could also inform revisions to future editions of the VB-MAPP. With some targeted revisions, the VB-MAPP could serve as a comprehensive assessment with strong validity evidence for this developmental age range. A summary of the strength of evidence across categories for the VB-MAPP Milestone domains, EESA, and Barriers Assessment is provided in the Appendix (Table A.4).

Limitations

As with all studies, this content validity study is not without potential limitations. First, although the sample size used in this study is within the recommended range, the inclusion of more SMEs may have slightly affected the results because each SME's responses would be weighted less heavily. Second, the sample predominantly identified as female and/or White. The distribution of sex/gender and race/ethnicity in the population is unknown so the sample may or may not representative. Third, the SMEs were not provided with the full VB-MAPP Guide; rather, they were given study guidelines, general VB-MAPP information, and all items from the VB-MAPP Milestones, EESA, and Barriers Assessment. The VB-MAPP Guide provides more detailed information such as rationale, examples, and scoring considerations. Although the goal was to have SMEs evaluate item content in terms of the relevance and representation, having the full VB-MAPP Guide may have influenced ratings. Fourth, the Transition Assessment and Task Analysis and Skills Tracking Assessment of the VB-

MAPP were excluded from this study. Finally, the COVID-19 pandemic occurred during the data collection phase of this study. Due to shelter-in-place mandates resulting in school and business closures and disruption to daily professional and personal activities, the length of time needed to complete the evaluation was extended.

Conclusions and Future Research

In general, the VB-MAPP has moderate to strong evidence supporting its domain relevance, age appropriateness, method of measurement appropriateness, and domain representation. The VB-MAPP is used by thousands of individuals practicing behavior analysis worldwide to make decisions and develop treatment plans for children in these content areas. The current study lends support to research and clinical practice based on the VB-MAPP. Based on the ratings and comments provided by SMEs, additions and/or revisions to some items within domains would only strengthen the content validity evidence of the VB-MAPP. Future reliability and validity research on VB-MAPP is also recommended. The consistency with which behaviors can be scored should be systematically evaluated by examining different types of reliability, such as interrater reliability, test-retest, and generalizability. Additionally, criterion-related validity evidence should be collected by comparing the scores of VB-MAPP with another, validated instrument or outcome to determine whether or not the results correlate as expected. Such studies are critical to support the continued widespread use of the VB-MAPP.

APPENDIX

APPENDIX

Supplemental Tables

Following are supplemental tables with additional information regarding SME ratings as well content validity evidence.

Table A.1.

Content Validity Ratio (CVR) and Summary of SME Item Response Distribution by Domain in Level 1

		Domain Relevance			Age Appropriateness			Method of Measurement Appropriateness		
Domain	CVR	*N*	*U*	*E*	*NA*	*SA*	*VA*	*NA*	*SA*	*VA*
Mand										
Item 1	.38	15	15	69	0	15	85	8	31	62
Item 2	.85	0	8	92*	8	8	85	0	46	54
Item 3	.38	0	31	69	8	31	62	0	23	77
Item 4	.54	0	23	77*	8	23	69	8	46	46
Item 5	.54	0	23	77*	8	25	67	0	23	77
Tact										
Item 1	.17	8	33	58	8	25	67	18	9	73
Item 2	.69	0	15	85*	0	8	92	8	8	83
Item 3	.54	0	23	77*	8	0	92	8	25	67
Item 4	.69	0	15	85*	0	15	85	8	58	33
Item 5	.54	8	15	77*	8	8	85	0	17	83
Listener Responding										
Item 1	.69	0	15	85*	0	31	69	15	62	23
Item 2	.85	0	8	92*	0	8	92	15	31	54
Item 3	.54	0	23	77*	0	23	77	0	23	77
Item 4	.54	0	23	77*	8	23	69	0	15	85
Item 5	-.08	8	46	46	23	31	46	8	23	69
Visual Perceptual Skills & Matching-To-Sample										
Item 1	.38	0	31	69	0	0	100	8	46	46
Item 2	-.08	8	46	46	0	38	62	0	54	46
Item 3	.69	8	8	85*	8	8	85	0	31	69
Item 4	.23	0	38	62	0	8	92	0	23	77
Item 5	.00	0	50	50	8	31	62	0	38	62
Independent Play										
Item 1	.38	0	31	69	0	15	85	0	33	67
Item 2	.38	0	31	69	8	31	62	8	46	46
Item 3	-.08	8	46	46	0	46	54	0	69	31
Item 4	-.08	0	54	46	0	46	54	0	54	46
Item 5	.38	0	31	69	0	31	69	0	46	54

(Continued)

Domain	CVR	Domain Relevance			Age Appropriateness			Method of Measurement Appropriateness		
		N	*U*	*E*	*NA*	*SA*	*VA*	*NA*	*SA*	*VA*
Social Behavior & Social Play										
Item 1	.54	0	23	77*	0	31	69	8	46	46
Item 2	.54	0	23	77*	0	8	92	8	46	46
Item 3	-.69	0	85	15	0	62	38	8	62	31
Item 4	.54	0	23	77*	0	31	69	0	54	46
Item 5	.08	0	46	54	0	38	62	8	46	46
Motor Imitation										
Item 1	.38	0	31	69	0	15	85	0	15	85
Item 2	.54	0	23	77*	0	15	85	0	15	85
Item 3	.38	0	31	69	0	15	85	0	8	92
Item 4	.23	0	38	62	0	31	69	0	31	69
Item 5	.23	0	38	62	0	31	69	0	8	92
Echoic										
Item 1	.54	0	23	77*	0	0	100	0	15	85
Item 2	.38	0	31	69	0	8	92	0	15	85
Item 3	.38	0	31	69	0	15	85	0	23	77
Item 4	.08	0	46	54	0	23	77	0	31	69
Item 5	.38	0	31	69	0	31	69	0	31	69
Spontaneous Vocal Behavior										
Item 1	.38	8	23	69	0	8	92	15	8	77
Item 2	.38	0	31	69	0	15	85	8	31	62
Item 3	.08	0	46	54	0	31	69	8	23	69
Item 4	.38	0	31	69	0	23	77	8	23	69
Item 5	-.08	0	54	46	0	38	62	8	23	69

* upper-tailed $p < .05$

Note. CVR = Content validity ratio; “E” = Essential; “U” = Useful, but Not Essential; “N” = Not necessary; “VA” = Very Appropriate; “SA” = Somewhat Appropriate; “NA” = Not Appropriate. Item wording excluded due to copyright protection.

Table A.2.

Content Validity Ratio (CVR) and Summary of SME Item Response Distribution by Domain in Level 2

		Domain Relevance			Age Appropriateness			Method of Measurement Appropriateness		
Domain	CVR	*N*	*U*	*E*	*NA*	*SA*	*VA*	*NA*	*SA*	*VA*
Mand										
Item 1	.38	0	31	69	0	46	54	8	38	54
Item 2	1.00	0	0	100*	0	0	100	0	15	85
Item 3	.69	0	15	85*	0	8	92	8	46	46
Item 4	.38	0	31	69	0	23	77	8	46	46
Item 5	.08	8	38	54	0	31	69	8	54	38
Tact										
Item 1	.85	0	8	92*	0	8	92	0	8	92
Item 2	.38	0	31	69	0	23	77	0	23	77
Item 3	.85	0	8	92*	0	15	85	0	8	92
Item 4	.23	0	38	62	0	31	69	8	23	69
Item 5	.38	0	31	69	0	38	62	15	23	62
Listener Responding										
Item 1	.08	15	31	54	8	33	58	15	31	54
Item 2	-.23	8	54	38	8	50	42	15	31	54
Item 3	.83	0	8	92*	0	17	83	0	0	100
Item 4	.23	0	38	62	0	42	58	8	23	69
Item 5	.08	0	46	54	0	50	50	15	31	54
Visual Perceptual Skills & Matching-To-Sample										
Item 1	.23	0	38	62	0	38	62	0	31	69
Item 2	.38	0	31	69	0	23	77	0	23	77
Item 3	.23	0	38	62	0	46	54	0	31	69
Item 4	.23	0	38	62	8	38	54	0	23	77
Item 5	-.23	0	62	38	0	62	38	0	38	62
Independent Play										
Item 1	.54	0	23	77*	0	0	100	0	38	62
Item 2	.85	0	8	92*	0	0	100	8	23	69
Item 3	.23	0	38	62	0	23	77	8	31	62
Item 4	.08	0	46	54	0	31	69	8	54	38
Item 5	.23	0	38	62	0	31	69	8	38	54

(Continued)

Domain	CVR	Domain Relevance			Age Appropriateness			Method of Measurement Appropriateness		
		N	*U*	*E*	*NA*	*SA*	*VA*	*NA*	*SA*	*VA*
Social Behavior & Social Play										
Item 1	-.08	0	54	46	0	38	62	8	69	23
Item 2	.23	0	38	62	0	31	69	8	62	31
Item 3	.00	0	50	50	8	46	46	0	62	38
Item 4	.23	0	38	62	8	38	54	17	42	42
Item 5	.17	0	42	58	8	42	50	8	62	31
Motor Imitation										
Item 1	.23	0	38	62	0	23	77	0	15	85
Item 2	.23	0	38	62	15	23	62	8	23	69
Item 3	.08	0	46	54	0	31	69	8	15	77
Item 4	.85	0	8	92	0	8	92	0	15	85
Item 5	1.00	0	0	100*	0	8	92	15	15	69
Echoic										
Item 1	.23	0	38	62	0	8	92	0	23	77
Item 2	.08	0	46	54	0	8	92	0	23	77
Item 3	.08	0	46	54	0	15	85	0	23	77
Item 4	-.38	0	69	31	0	31	69	0	31	69
Item 5	-.08	0	54	46	0	46	54	0	31	69
Listener Responding by Function, Feature, & Class										
Item 1	.38	0	31	69	0	38	62	8	15	77
Item 2	.23	0	38	62	0	38	62	8	15	77
Item 3	.69	0	15	85*	0	31	69	0	15	85
Item 4	.08	0	46	54	0	38	62	0	15	85
Item 5	.08	8	38	54	0	38	62	8	31	62
Intraverbal										
Item 1	.54	0	23	77*	0	15	85	8	23	69
Item 2	.85	0	8	92*	0	0	100	0	8	92
Item 3	.08	8	38	54	8	38	54	0	31	69
Item 4	.54	0	23	77*	0	46	54	0	23	77
Item 5	.54	0	23	77*	8	31	62	0	23	77
Classroom Routines & Group Skills										
Item 1	-.08	8	46	46	0	38	62	0	31	69
Item 2	-.08	8	46	46	15	46	38	8	25	67
Item 3	.08	15	31	54	15	38	46	8	23	69
Item 4	.08	15	31	54	8	46	46	8	23	69

(Continued)

Domain	CVR	Domain Relevance			Age Appropriateness			Method of Measurement Appropriateness		
		N	*U*	*E*	*NA*	*SA*	*VA*	*NA*	*SA*	*VA*
Item 5	-.08	23	31	46	23	38	38	15	23	62
Linguistic Structure										
Item 1	.54	0	23	77*	0	23	77	8	31	62
Item 2	.69	0	15	85*	0	8	92	8	31	62
Item 3	.69	0	15	85*	0	8	92	8	23	69
Item 4	-.23	0	62	38	0	31	69	0	23	77
Item 5	.38	0	31	69	0	31	69	23	23	54

* upper-tailed $p < .05$

Note. CVR = Content validity ratio; “E” = Essential; “U” = Useful, but Not Essential; “N” = Not necessary; “VA” = Very Appropriate; “SA” = Somewhat Appropriate; “NA” = Not Appropriate. Item wording excluded due to copyright protection.

Table A.3.

Content Validity Ratio (CVR) and Summary of SME Item Response Distribution by Domain in Level 3

		Importance			Age Appropriateness			Method of Measurement Appropriateness		
Domain	CVR	*N*	*U*	*E*	*NA*	*SA*	*VA*	*NA*	*SA*	*VA*
Mand										
Item 1	.23	0	38	62	0	23	77	0	46	54
Item 2	.54	0	23	77*	0	23	77	0	31	69
Item 3	.69	0	15	85*	0	15	85	0	38	62
Item 4	.69	0	15	85*	0	8	92	0	23	77
Item 5	.38	0	31	69	0	23	77	0	31	69
Tact										
Item 1	.38	0	31	69	0	0	100	0	15	85
Item 2	.54	0	23	77*	0	31	69	0	23	77
Item 3	.54	0	23	77*	0	15	85	0	15	85
Item 4	.38	0	31	69	0	15	85	0	23	77
Item 5	.08	0	46	54	0	15	85	23	31	46
Listener Responding										
Item 1	.38	0	31	69	0	15	85	0	8	92
Item 2	.23	0	38	62	0	31	69	0	8	92
Item 3	.54	0	23	77*	0	8	92	0	15	85
Item 4	.38	0	31	69	0	23	77	0	8	92
Item 5	.08	0	46	54	0	23	77	25	25	50
Visual Perceptual Skills & Matching-To-Sample										
Item 1	-.23	0	62	38	0	46	54	8	23	69
Item 2	.33	0	33	67	0	31	69	0	25	75
Item 3	-.54	8	69	23	0	77	23	8	23	69
Item 4	.23	0	38	62	0	38	62	0	31	69
Item 5	-.23	8	54	38	0	62	38	0	38	62
Independent Play										
Item 1	.54	0	23	77*	0	0	100	8	15	77
Item 2	.38	0	31	69	0	8	92	8	23	69
Item 3	.38	0	31	69	0	15	85	8	15	77
Item 4	.69	0	15	85*	0	8	92	8	15	77
Item 5	-.08	0	54	46	0	46	54	15	15	69

(Continued)

Domain	CVR	Importance			Age Appropriateness			Method of Measurement Appropriateness		
		N	*U*	*E*	*NA*	*SA*	*VA*	*NA*	*SA*	*VA*
Social Behavior & Social Play										
Item 1	.38	0	31	69	0	23	77	8	31	62
Item 2	.38	0	31	69	0	23	77	8	54	38
Item 3	.54	0	23	77*	0	23	77	8	38	54
Item 4	.54	0	23	77*	0	15	85	8	23	69
Item 5	.38	0	31	69	0	15	85	8	31	62
Reading										
Item 1	.69	0	15	85*	0	8	92	15	15	69
Item 2	.08	0	46	54	0	31	69	8	0	92
Item 3	.08	0	46	54	0	38	62	0	8	92
Item 4	.08	0	46	54	0	38	62	8	8	85
Item 5	-.23	8	54	38	15	54	31	8	8	85
Writing										
Item 1	.38	0	31	69	0	31	69	0	23	77
Item 2	.23	0	38	62	0	38	62	8	15	77
Item 3	.23	8	31	62	0	54	46	0	23	77
Item 4	-.08	8	46	46	15	54	31	0	23	77
Item 5	-.23	23	38	38	15	62	23	0	23	77
Listener Responding by Function, Feature, & Class										
Item 1	.69	0	15	85*	0	15	85	0	23	77
Item 2	.69	0	15	85*	0	15	85	0	15	85
Item 3	.38	0	31	69	0	31	69	0	31	69
Item 4	.69	0	15	85*	0	15	85	0	31	69
Item 5	-.08	0	54	46	0	46	54	15	38	46
Intraverbal										
Item 1	.38	0	31	69	0	23	77	0	46	54
Item 2	.23	0	38	62	0	15	85	8	46	46
Item 3	.23	0	38	62	0	38	62	0	31	69
Item 4	.23	0	38	62	0	38	62	0	46	54
Item 5	.54	0	23	77*	0	23	77	0	15	85
Classroom Routines & Group Skills										
Item 1	.69	8	8	85*	0	23	77	15	23	62
Item 2	.38	8	23	69	0	15	85	8	15	77
Item 3	.38	8	23	69	0	15	85	8	15	77
Item 4	.08	8	38	54	8	31	62	15	15	69

(Continued)

Domain	CVR	Importance			Age Appropriateness			Method of Measurement Appropriateness		
		N	*U*	*E*	*NA*	*SA*	*VA*	*NA*	*SA*	*VA*
Item 5	-.23	23	38	38	31	31	38	8	31	62
Linguistic Structure										
Item 1	.23	0	38	62	0	31	69	8	38	54
Item 2	.08	0	46	54	0	38	62	0	38	62
Item 3	.69	0	15	85*	0	8	92	0	31	69
Item 4	.38	0	31	69	0	15	85	0	31	69
Item 5	.23	0	38	62	0	23	77	0	31	69
Math										
Item 1	.08	0	46	54	0	38	62	0	31	69
Item 2	.08	0	46	54	0	46	54	0	31	69
Item 3	.54	0	23	77*	0	23	77	0	31	69
Item 4	.54	0	23	77*	0	15	85	0	23	77
Item 5	.08	15	31	54	0	42	58	0	31	69

* upper-tailed $p < .05$

Note. CVR = Content validity ratio; “E” = Essential; “U” = Useful, but Not Essential; “N” = Not necessary; “VA” = Very Appropriate; “SA” = Somewhat Appropriate; “NA” = Not Appropriate. Item wording excluded due to copyright protection.

Table A.4.

Summary of Content Validity Evidence Strength by Level, Domain, and Category

Level	Domain	Domain Relevance	Age Approp.	Measurement Approp.	Domain Representation
1	Mand	Strong	Strong	Moderate	Moderate
	Tact	Strong	Strong	Moderate	Moderate
	Listener Responding	Strong	Strong	Moderate	Moderate
	Visual Perceptual Skills & Matching-To-Sample	Moderate	Strong	Moderate	Limited
	Independent Play	Moderate	Moderate	Limited	Limited
	Social Behavior & Social Play	Moderate	Moderate	Limited	Moderate
	Motor Imitation	Moderate	Strong	Strong	Strong
	Echoic	Moderate	Strong	Strong	Moderate
	Spontaneous Vocal Behavior	Moderate	Strong	Strong	Moderate
2	Mand	Strong	Strong	Moderate	Moderate
	Tact	Strong	Strong	Strong	Moderate
	Listener Responding	Moderate	Moderate	Moderate	Limited
	Visual Perceptual Skills & Matching-To-Sample	Moderate	Moderate	Strong	Strong
	Independent Play	Strong	Strong	Moderate	Moderate
	Social Behavior & Social Play	Moderate	Moderate	Limited	Limited
	Motor Imitation	Strong	Strong	Strong	Limited
	Echoic	Limited	Strong	Strong	Moderate
	Listener Responding by Function, Feature, & Class	Moderate	Moderate	Strong	Limited
	Intraverbal	Strong	Strong	Strong	Moderate
	Classroom Routines & Group Skills	Limited	Limited	Moderate	Moderate
	Linguistic Structure	Strong	Strong	Moderate	Limited
3	Mand	Strong	Strong	Moderate	Strong
	Tact	Strong	Strong	Strong	Moderate
	Listener Responding	Moderate	Strong	Strong	Strong
	Visual Perceptual Skills & Matching-To-Sample	Limited	Limited	Moderate	Moderate
	Independent Play	Strong	Strong	Strong	Strong
	Social Behavior & Social Play	Strong	Strong	Moderate	Limited
	Reading	Moderate	Moderate	Strong	Moderate
	Writing	Moderate	Limited	Strong	Limited

(Continued)

Level	Domain	Domain Relevance	Age Approp.	Measurement Approp.	Domain Representation
	Listener Responding by Function, Feature, & Class	Strong	Strong	Strong	Moderate
	Intraverbal	Moderate	Strong	Moderate	Limited
	Classroom Routines & Group Skills	Moderate	Strong	Strong	Limited
	Linguistic Structure	Moderate	Strong	Moderate	Strong
	Math	Moderate	Moderate	Strong	Moderate
N/A	EESA	Moderate	Strong	Moderate	Strong
N/A	Barriers	Strong	N/A	Moderate	Limited

Note. Shading used to aid in readability. Age Approp. = Age Appropriateness; Measurement Approp. = Method of Measurement Appropriateness; EESA = Early Echoic Skills Assessment. Limited Evidence = Fewer than half of SMEs rated the items at highest response category (e.g., Essential, Very Appropriate, Adequate), on average; Moderate Evidence = Between than 50% and 68.9% of SMEs rated the items at highest response category (e.g., Essential, Very Appropriate, Adequate), on average; Strong Evidence = 69% or more of SMEs rated the items at highest response category (e.g., Essential, Very Appropriate, Adequate), on average. Neither EESA nor Barriers Assessment were specific to any level from the VB-MAPP Milestones.

BIBLIOGRAPHY

Aiken, L. R. (1980). Content validity and reliability of single items or questionnaires. *Educational and Psychological Measurement, 40*(4), 955-959.

American Educational Research Association, American Psychological Association, and National Council on Measurement in education, & Joint Committee on Standards for Educational and Psychological Testing. (2014). *Standards or Educational and Psychological Testing*. AERA

American Psychological Association (2017a). *Cognitive and social skills to expect from 18 to 36 months.* Retrieved from https://www.apa.org/act/resources/fact-sheets/development-36-months

American Psychological Association (2017b). *Cognitive and social skills to expect from 3 to 5 years*. Retrieved from https://www.apa.org/act/resources/fact-sheets/development-5-years

Anastasi, A. (1988). *Psychological Testing*. MacMillan Publishing Company.

Austin, K., & Thomas, J. (2017, May). *Administering assessments for comprehensive behavior analytic programs: Analyzing practitioner skills and reliability*. Symposium at the 43rd Annual Convention for the Association for Behavior Analysis International, Denver, CO.

Axelrod, S., McElrath, K. K., & Wine, B. (2012). Applied behavior analysis: Autism and beyond. *Behavioral Interventions, 27*, 1-15.

Baer, D. M., Wolf, M.M., & Risley, T.R. (1987). Some still-current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis, 20*, 313-327.

Benson, J. (1998). Developing a strong program of construct validation: A test anxiety example. Educational measurement: *Issues and Practice, 17*(1), 10-17.

Bond, L. A. (1996) Norm- and Criterion-Referenced Testing, *Practical Assessment, Research, and Evaluation, 5*, 1-3.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56(2)*, 81-105.

CCSSO SEC Collaborative Project (2005) (2005). *Surveys of enacted curriculum: a guide for sec collaborative state and local coordinators.* Council of Chief State School Officers.

Centers for Disease Control and Prevention. (2016). *Autism and developmental disabilities monitoring (ADDM) network*. Retrieved from https://www.cdc.gov/ncbddd/autism/facts.html

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6,* 284-290.

Coleman, R. S., Whitman, T. C., Johnson, M. R. (1979). Suppression of self-stimulatory behavior of profoundly retarded boy across staff and settings: An assessment of situational generalization. *Behavior Therapy, 10,* 266-280.

Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: Theory and application. *The American Journal of Medicine, 119,* 166.e7-166.e16.

Cooper, J. O., Heron, T. E., & Heward, W. L. (2007). *Applied behavior analysis* (2nd ed.) Pearson Education, Inc.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Harcourt Brace Jovanovich.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 443-507). American Council on Education.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281.

Dennis, L. R., & Votteler, N. K. (2013). Preschool teachers and children's emergent writing: Supporting diverse learners. *Early Childhood Education Journal, 41*(6), 439-446.

DeSouza, A. A., Akers, J. S., & Fisher, W. W. (2017). Empirical application of Skinner's Verbal behavior to interventions for children with autism: A review. *The Analysis of Verbal Behavior, 33*(2), 229-259.

Dixon, M. R. (2014). *The PEAK relational training system: Direct training module*. Carbondale, IL: Shawnee Scientific Press.

Dixon, M. R., Belisle, J., Stanley, C., Rowsey, K., Daar, J. H., & Szekely, S. (2015). Toward a behavior analysis of complex language for children with autism: Evaluating the relationship between PEAK and the VB-MAPP. *Journal of Developmental and Physical Disabilities, 27*(2), 223-233.

Dixon, M. R., Belisle, J., Whiting, S. W., & Rowsey, K. E. (2014). Normative sample of the PEAK relational training system: Direct training module and subsequent comparisons to individuals with autism. *Research in Autism Spectrum Disorders, 8(11),* 1597-1606.

Dixon, M. R, Rowsey, K., Gunnarsson, K. F., Belisle, J., Stanley, C. R., & Daar, J. H. (2017). Normative sample of the PEAK relational training system: Generalization module with comparison to individuals with autism. *Journal of Behavioral Education, 26*, 101-122.

Durand, V. M. (1989). *The Motivation Assessment Scale*. Pergamon Press.

Ebel, R. L. (1956). Obtaining and reporting evidence on content validity. *Educational and Psychological Measurement, 16*(3), 269-282.

Fitzpatrick, A. R. (1983). The meaning of content validity. *Applied Psychological Measurement, 7*(1), 3-13.

Florida Center on Self-Injury (2000). *Functional Assessment Screening Tool*. Retrieved from http://www.rodspecialeducation.org/uploads/Functional%20Analysis%20Screening%20Tool%20FAST.pdf

Foxx, R. M. (2008). Applied behavior analysis treatment of autism: the state of the art. *Child and Adolescent Psychiatric Clinics of North America, 17*, 821-834.

Grant, J. S., & Davis, L. L. (1997). Selection and use of content experts for instrument development. *Research in Nursing & Health, 20*, 269-274.

Gould, E., Dixon, D. R., Najdowski, A. C., Smith, M. N., & Tarbox, J. (2011). A review of assessments for determining the content of early intensive behavior intervention programs for autism spectrum disorders. *Research in Autism Spectrum Disorders, 5*, 990-1002.

Gullikson, H. (1950). *Theory of mental tests.* Wiley.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items*. Routledge.

Hawkins, R. P. (1979). The functions of assessment: implications for selection and development of devices for assessing repertoires in clinical, educational, and other settings. *Journal of Applied Behavior Analysis, 12*, 501-616.

Haynes, S. N., Richard, D., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological assessment, 7*(3), 238.

Higgins, P. A., & Straub, A. J. (2006). Understanding the error of our ways: Mapping the concepts of validity and reliability. *Nurse Outlook, 54*, 23-29.

Hubley, A. M., & Zumbo, B. D. (2013). Psychometric characteristics of assessment procedures: an overview. In *APA handbook of testing and assessment in psychology, volume 1* (pp. 3-19). American Psychological Association.

IBM Corp. (2017). *IBM SPSS statistics for Windows, version 25*. IBM Corp.

Iwata, B. A., Dorsey, M. F., Slifer, K. J., Bauman, K. E., & Richman, G. S. (1982). Toward a functional analysis of self-injury. *Analysis and Intervention in Developmental Disabilities, 2*, 3-20.

Johnson, R. L., & Morgan, G. B. (2016). *Survey scales: a guide to development, analysis, and reporting*. Guilford Press.

Johnson, R. L., Penny, J. A., & Gordon, B. (2008). *Assessing performance: Designing, scoring, and validating performance tasks*. Guilford Press.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17-64). American Council on Education.

Lawshe, C.H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*, 563-575.

Linehan, A. A. (1977). Issues in behavioral interviewing. In J. D. Cone & R. P. Hawkins (Eds.), *Behavioral assessment: New directions in clinical psychology* (pp. 30-51). Brunner/Mazel.

Linehan, M. M. (1980). Content validity: Its relevance to behavioral assessment. *Behavioral Assessment, 2*, 147-159.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*(3), 635-694.

Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B. L., DiLavore, P. C., ... & Rutter, M. (2012). The Autism Diagnostic Observation Schedule—Generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of Autism and Developmental Disorders, 30*(3), 205-223.

Lovaas, O. I. (1987). Behavioral treatment and normal educational and intellectual functioning in young autistic children. *Journal of Counseling and Clinical Psychology, 55*, 3-9.

Luiselli, J. K., Campbell, S., Cannon, B., DiPietro, E., Ellis, J. T., Taras, M., & Lifter, K. (2001). Assessment instruments used in the education and treatment of persons with autism: Brief report of a survey of a national service centers. *Research in Developmental Disabilities, 22,* 389-398.

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing research, 35*(6) 382-385.

Meadows, T. J. (2017). [Test review of Verbal Behavior Milestones Assessment and Placement Program]. In J. F. Carlson, K. F. Geisinger, & J. L. Jonson (Eds.), The twentieth mental measurements yearbook. Retrieved from http://marketplace.unl.edu/buros/

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist, 30*(10), 955-966.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*(11), 1012-1027.

Messick, S. (1989a). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5-11.

Messick, S. (1989b). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). Macmillan.

Microsoft Corporation (2018). *Microsoft Excel*. Retrieved from https://office.microsoft.com/excel

Mosier, C. I. (1947). A critical examination of the concepts of face validity. *Educational and Psychological Measurement, 7*(2), 191-205.

National Council of Measurement in Education (1995). *Code of Professional Responsibilities.* Retrieved from https://www.ncme.org/resources/library/professional-responsibilities

Nunnally, J. (1967). *Psychometric methods*. McGraw-Hill.

Padilla, K. L. (2019). Global assessment use and practices in applied behavior analysis: Surveying the field. Manuscript submitted for publication.

Padilla K. L., Morgan, G. B., Weston, R., Lively, P., & O'Guinn, N. (2019). Validity and Reliability of Behavior Analytic Assessments: a Review of the Literature. In progress.

Partington, J. W. (2006). *Assessment of Basic Language and Learning Skills, Revised*. Behavior Analysts, Inc.

Partington, J. W., Bailey, A., & Partington, S. W. (2018). A pilot study examining the test-retest and internal consistency reliability of the ABLLS-R. *Journal of Psychoeducational Assessment, 36*(4), 405-410.

Popham, W. J. (1992). Appropriate expectations for content judgments regarding teacher licensure tests. *Applied Measurement in Education, 5*, 285-301.

Porter, A.C, & Smithson, J.L. (2001). *Defining, developing, and using curriculum indicators*. Consortium for Policy Research in Education. Retrieved from http://www.cpre.org/sites/default/files/researchreport/788_rr48.pdf

Raymond, M. R., & Reid, J. B. (2001). Who made thee a judge? Selecting and training participants for standard setting. In G. J. Cizek (Ed.), *Setting performance standards: Theory and application* (pp. 133-172). Lawrence Erlbaum Associates.

Rea, L. M., & Parker, R. A. (1992). *Designing and conducting survey research*. Jossey-Bass.

Rothman, R., Slattery, J.B., Vranek, J.L., & Resnick, L.B. (2002). Benchmarking and Alignment of Standards and Testing (Technical Report No. 566). National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Retrieved from http://www.cse.ucla.edu/products/reports/TR566.pdf

Rovinelli, R. J., Hambleton, R. K. (1977). On the use of content specialists in the assessment of criterion-referenced test item validity. *Dutch Journal for Educational Research*, 2, 49–60.

Rowe, D. W., & Neitzel, C. (2010). Interest and agency in 2-and 3-year-olds' participation in emergent writing. *Reading Research Quarterly, 45*(2), 169-195.

Rubio, D. M., Berg-Weger, M., Tebb, S. S., Lee, E. S., & Rauch, S. (2003). Objectifying content validity: Conducting a content validity study in social work research. *Social Work Research, 27*, 94-104.

Rulon, P. J. (1946). On the validity of educational tests. *Harvard Educational Review, 16*, 290-296.

Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research, 8*, 3-15.

Schopler, E., Van Bourgondien, M. E., Wellman, G. J., & Love, S. R. (2010) Childhood Autism Rating Scale, Second Edition (CARS-2). Pearson Assessments.

Shultz, K. S., Whitney, D. J., & Zickar, M. J. (2014). *Measurement theory in action: case studies and exercises*. Routledge.

Sireci, S. (1998). The construct of content validity. Social Indicators Research, 45, 83-117.

Sireci, S., & Faulkner-Bond, M. (2014). Validity evidence based on test content. Psicothema, 26, 100-107.

Skinner, B. F. (1957). *Verbal behavior*. Prentice Hall Englewood Cliffs.

Sparrow, S. S., Balla, D.A., & Cicchetti, D. V. (1984). Vineland Adaptive Behavior Scales. American Guidance Service.

Sparrow, S. S., Cicchetti, D. V., & Saulnier, C. A. (2016). *Vineland Adaptive Behavior Scales – Third Edition*. Pearson Assessments.

Sparrow, S. S., Cicchetti, D. V., & Balla, D. A. (2005). *Vineland Adaptive Behavior Scales – Second Edition*. Pearson Assessments.

Sudman, S., & Bradburn, N. M. (1982). *Asking questions: a practice guide to questionnaire design.* Josey-Bass.

Sundberg, M. L. (2014). *VB-MAPP: Verbal behavior milestones assessment and placement program: A language and social skills assessment program for children with autism or other developmental disabilities*. AVB Press.

Sundberg, M. L., & Michael, J. (2001). The benefits of Skinner's analysis of verbal for children with autism. *Behavior Modification, 25*, 698-724.

Usry, J., Partington, S. W., & Partington, J. W. (2018). Using expert panels to examine the content validity and inter-rater reliability of the ABLLS-R. *Journal of Developmental & Physical Disabilities, 30*(1), 27–38.

Waugh, C. K. & Gronlund, N. E. (2012). *Assessment of Student Achievement* (10th ed.). Pearson.

Webb, N. (1997). Criteria for alignment of expectations and assessments in mathematics and science education. National Institute for Science Education Madison, WI. Retrieved from http://facstaff.wceruw.org/normw/WEBBMonograph6criteria.pdf.

Wilson, F. R., Pan, W., & Schumsky, D. A. (2012). Recalculation of the critical values for Lawshe's content validity ratio. *Measurement and Evaluation in Counseling and Development, 45*(3), 197-210.