Journal of Autism and Developmental Disorders (2021) 51:4054–4066
https://doi.org/10.1007/s10803-020-04864-y

ORIGINAL PAPER

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

Kristen L. Padilla[1] · Jessica S. Akers[1]

Accepted: 27 December 2020 / Published online: 14 January 2021

Abstract

The purpose of this study is to provide content validity evidence for the Verbal Behavior Milestones Assessment and Placement Program (VB-MAPP). A national panel of 13 experts 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 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.

Keywords Content validity · VB-MAPP · Applied behavior analysis · Autism · Verbal behavior

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 that have been shown to be very effective in treating individuals diagnosed with autism spectrum disorder (ASD; Axelrod et al. 2012; Foxx 2008; Lovaas 1987; National Autism Center 2015; Steinbrenner et al. 2020). 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 (1) rise in prevalence and (2) breadth of developmental 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 et al. 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).

Behavior analysts use a variety of assessments to assess an individual's strengths and weaknesses (Gould et al. 2011), identify the function of an individual's challenging behavior (Iwata et al. 1982), and develop goals (Sundberg 2014). Each type of assessments provides information about the present levels of the behavior of interest, which in turn offers insights for developing treatment plans for individuals with ASD. The assessment results can serve as the baseline against which further evaluations of the characteristic or behavior of interest will be compared to determine the extent of the behavior change.

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 et al. 2007) and skill acquisition assessments aide in the identification of an individual's current skill level and potential areas of growth. Regardless of the purpose, assessments based within the ABA framework heavily rely on direct observation of an individual's behavior.

There are two major types of assessments that are generally used to determine an individual's level of performance

✉ Kristen L. Padilla
Kristen_Padilla-Mainor@baylor.edu

[1] Department of Educational Psychology, Baylor University, One Bear Place #97301, Waco, TX 76798-7301, USA

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 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). 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 decisions or classifications (e.g., skill level, mastery, proficiency, certification) (Crocker and Algina 1986). 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 2020b). Regardless of the type or purpose of assessment, the availability of evidence supporting the assessment's use for an intended purpose is critical and the focus of the next section.

Validity Evidence in Behavior Assessment

Evidence supporting the use of an instrument should be collected and disseminated in order to have confidence in the decisions made based on assessment results. 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 1989). Such inferences should be based on theoretical and empirical data from multiple sources that align with the conclusions (Shultz et al. 2014). Several dimensions of ABA align with these principles of validity. Both frameworks are based on scientific theory that guides 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)—construct-, criterion-, and content-related validity evidence (Benson 1998; Cronbach and Meehl 1955) evidence.

Linehan (1980) stated that the types of inferences that are made in behavioral assessments, such as those in ABA, "necessitate attention to content validity" (p. 152). In fact, Linehan argued the need for content validity in most instances of behavior assessment. Sireci (1998) noted three common components of content validity that originated with the writings of Rulon (1946), Mosier (1947), and Gulliksen (1950), which were domain definition, domain representation, and domain relevance. 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. *Domain definition* refers to how the construct being measured is operationally defined (Sireci and Faulkner-Bond 2014). Providing evidence for domain definition involves subject matter experts (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 and Faulkner-Bond 2014). *Domain relevance* refers to the degree of importance or relevance each item has to its target domain. For domain relevance, SMEs commonly review and rate the relevance of each item as it relates to the target domain. The *appropriateness of the test development process* refers to how faithfully and fully processes were in creating instruments for measuring intended constructs (Sireci and Faulkner-Bond 2014).

Use of Standardized Instruments in ABA

Over the past 30 years, numerous instruments have been developed within the ABA framework that specifically target skill acquisition for individuals diagnosed with ASD. Until recently, there has been very limited research about the types of instruments used for individuals with ASD despite the critical role assessment plays in the diagnosis, treatment planning, and progress monitoring for this population (Ackley et al. 2019; Luiselli et al. 2001; Padilla 2020b). 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 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%).

Austin and Thomas (2017) conducted a small-scale survey with 99 participants in the state of Washington regarding their clinical assessment practices for diagnostic and educational programming. The Austin and Thomas (2017) study was the first to focus on practices specific to professionals working in behavior analysis. A slight majority of respondents (56%) reported using the Verbal Behavior Milestones Assessment and Placement and Program (VB-MAPP; Sundberg 2014) and about 40% reported using the Assessment of Basic Language Learning Skills-Revised (ABLLS-R; Partington 2006) for programming.

Padilla (2020b) expanded this area of research by surveying 1,428 individuals who primarily practice in ABA throughout the world. The most widely reported assessment used for educational and curriculum programming was the VB-MAPP with 76% (n = 1,086) of the respondents reporting its use by itself or in addition to another assessment. Approximately 45% (n = 638) of ABA professionals reported using the ABLLS-R, 34% (n = 485) reported using the Vineland Adaptive Behavior Scales, and roughly 14% (n = 197) reported using the Promoting the Emergence of Advanced Knowledge (PEAK; Dixon 2014). The prevalent use of the VB-MAPP was reported across professional positions and certification levels as a part of goal identification, educational programming, and curriculum development. For instance, 80% (n = 623) of Board Certified Behavior Analysts (BCBAs), 78% (n = 94) of Board Certified Behavior Analyst, Doctoral designation (BCBA-Ds), 78% (n = 42) of Board Certified Assistant Behavior Analysts (BCaBAs), 68% (n = 290) of Registered Behavior Technicians (RBTs), 80% (n = 8) of those without any reported credential, and 69% (n = 27) of individuals with other types of credentials reported using the VB-MAPP. Furthermore, with respect to title, role, or position, 53% (n = 397) of clinical supervisors, 72% (n = 103) of faculty members, 74% (n = 554) of practitioners, and 77% (n = 140) of graduate students also reported using the VB-MAPP.

Description of VB-MAPP

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 to 48 months. 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, such as behavior problems and instructional control; (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 within the Milestones assessment; and (e) the Curriculum Placement and IEP Goals, which helps the individual designing the program develop an all-inclusive intervention plan (Sundberg 2014). The VB-MAPP also includes the Early Echoic Skills Assessment (EESA), which is a separate subtest used to evaluate an individual's ability to repeat speech sounds (e.g. phonemes, syllables, intonation) and is the basis for assessing the Echoic domain of VB-MAPP. The EESA includes five groups of items: Group 1—vowels and dipthongs; Groups 2 and 3—early consonants in 2- and 3-syllable combinations, respectively; and Groups 4 and 5—prosodic features of speech, including pitch, loudness, and vowel duration (Esch 2014). For each item, the examinee is provided a prompt/discriminative stimulus (S^D) to which they must echo to receive a correct response.

Purpose of Current Study

Understanding and conducting validity research is crucial for the scientific advancement of ABA assessment within research and practice given the prevalent use of skill acquisition assessments used within the field (Padilla 2020b). Within the field of ABA, decisions regarding assessment must be based on evidence and research. There is a strong need for both an evaluative tool for examining 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 its results (Padilla 2020a) and to support its continued used.

To our knowledge, two studies have collected the reliability and validity evidence available for ABA-based instruments and/or curricula (Ackley et al. 2019; Padilla et al. 2020). As reported by Padilla (2020b), the VB-MAPP is the most widely used instrument reported by ABA professionals, yet only one study was found across the aforementioned reviews that collected reliability evidence for VB-MAPP, as the primary focal instrument. Montallana et al. (2019) reported inter-reliability estimates of 0.88 and 0.63 for the Milestones and Barriers, respectively. No studies have collected validity evidence directly for the VB-MAPP, which is the focus of this study. As noted previously, content validity has four components: (a) domain definition, (b) domain relevance, (c) domain representation, and (d) appropriateness of test development process. The VB-MAPP is now in its second edition so domain definition and appropriateness of test development process 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?

Methods

The design of the current study was modeled after Usry et al. (2018), which examined the content validity of less commonly used ABA-based assessment (i.e., ABLLS-R; Padilla 2020b). The current study incorporated the recommended improvements of Usry et al. (2018) so as to avoid the same limitations, which were 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.

Participants

Raymond and Reid (2001) suggested 10–15 experts for panels who are tasked with setting performance standards by determining one or more cut-off scores that demonstrates proficiency in a specific area. In this study, 15 SMEs agreed to participate, but two did not complete the survey due to effects of the COVID-19 pandemic. In order to qualify as an SME, participants were required to (a) be a certified practicing behavior analyst (i.e., BCBA, BCBA-D), (b) have at least 7 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 5 years, (e) have independently administered the VB-MAPP at least once prior to the study (adapted from Usry et al. 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

The 13 SMEs in this study 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 ($n=12$) self-identified as White and one SME (8%) self-identified as 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 1.

Instrumentation

In addition to the consent form and description of the study, each participating 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 and Davis 1997). For each of the five items in each Milestone domain, SMEs 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." A three-point Likert-type scale was used for the remaining categories for consistency with Lawshe (1975). For domain representation, SMEs rated 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." The same general format was followed for EESA and

Table 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

BCBA Board certified behavior analyst, *BCBA-D* Board certified behavior analyst, doctoral designation, *SME* subject matter expert, *Cert.* certification, *Exp.* years of experience

Barriers Assessment. Following Grant and Davis (1997), SMEs were allowed to provide recommendations for item revisions, additions, or deletions in an open-ended response format after rating each domain.

Procedures

Each SME was initially 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 using Qualtrics to complete the study. Participants were given six weeks to complete the study and reminders were sent at weeks three and five.

Data Analysis

The following analyses were conducted using IBM SPSS version 25 (IBM Corp 2017) and Microsoft Excel (Microsoft Corporation 2018) after all data were collected. First, the frequency distribution for each item was generated to show how SMEs 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 by nine out of 13 SMEs (69.2%). All percentages were compared against the threshold of 69.2%, which is informed by Popham (1992). The authors classified the strength of evidence based on percentages; that is, percentages 69% or greater were classified as strong, percentages between 50 and 60% as moderate, and percentages less than 50% as limited. 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. A CVR of 0 indicates that 50% of the SMEs rated the item as "Essential." Wilson et al. (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 0.456, which was used as a comparison for all CVRs in this study. The critical value of 0.456 corresponds to 10 or more SMEs out of 13 rating an item as "Essential" in order to be considered statistically significant (i.e., percentage of SME support exceeds 50%). 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 et al. 2014). Additionally, the CVI was calculated for each level across domains and the test as a whole. Following Lawshe (1975), CVRs and CVIs apply only to domain relevance.

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 0.54, and nine out of 13 SMEs

(69.2%) has a CVR of 0.38. Thus, an SME endorsement rating of 69.2% is considered meaningful even though the hypothesis test of the CVR is more conservative.

The relationships between the method of measurement appropriateness ratings across different methods of measurement were estimated using Cramér's *V*, which is a χ^2-based measure of association between categorical or nominal variables (Cramér 1946; Liebetrau 1983; Rea and Parker 1992).

As noted previously, SMEs could provide recommendations for item revisions, additions, or deletions on the items within each domain. The responses provided were analyzed qualitatively using thematic analysis of the text. That is, the responses were reviewed and categorized based on thematic elements that emerged with respect to recommended changes to the instrument.

Results

VB-MAPP Milestones

Domain Relevance

The entire Milestones Assessment, which includes all items for each domain across all three developmental levels, had a CVI = 0.32. This estimate of 0.32 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 ≤ 0), on average. The domains with negative CVIs were Classroom Routines and Group Skills (CR-GS) and Echoic within Level 2 as well as Visual Perceptual Skills and Matching-to-Sample (VPS-MTS) in Level 3. The calculated CVI for Levels 1, 2, and 3 across all domains, respectively, were 0.35, 0.30, and 0.30. This indicates that on average, the majority of raters indicated that the items within domains were "Essential."

For each of 170 items the VB-MAPP Milestones Assessment, the CVRs were compared against the upper-tailed critical value of 0.456, which was associated with a Type I error rate (α) of 5%. Fifty-three of the 170 items (31%) had CVRs that exceeded 0.456; thus, the null hypothesis that these CVRs were equal to zero could be rejected. 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 (SVB) 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 VPS-MTS, Social Behavior & Social Play (SB-SP), Echoic, and CR-GS. In Level 3, 19 of the 65 items (29%) had CVRs statistically greater than zero. Two domains within Level 3—VPS-MTS and Writing—had no statistically significant items. Overall, there was moderate to strong evidence supporting domain relevance for the majority of domains across levels. A summary of the domain-level relevance results is presented in Table 2.

Developmental Age Appropriateness

There is moderate to strong evidence across the vast majority of domains across all levels for age appropriateness. In Level 1, there was strong support for developmental age appropriateness on seven of the nine domains. That is, the majority of SMEs rated the developmental age range on items as "Very Appropriate," on average, for Mand (77%), Tact (84%), Listener Responding (71%), VPS-MTS (80%), Motor Imitation (79%), Echoic (85%), and SVB (77%). There was moderate support for two of the nine domains in Level 1, Independent Play (65%) and SB-SP (66%). In Level 2, there was strong support (i.e., more than 69% of SME's rating the items as "Very Appropriate," on average) for age appropriateness for seven of the 12 domains. These domains included Mand (78%), Tact (77%), Independent Play (83%), Motor Imitation (78%), Intraverbal (71%), and Linguistic Skills (80%). There was moderate support for Listener responding (58%), VPS-MTS (57%), and Listener Responding by Function, Feature, and Class (LRFFC; 80%). Within Level 2, there was limited support for only one of the domains, CR-GS, with fewer than half of the SMEs (46%) rating the items as "Very Appropriate," on average. Based on the SME comments, the lower age appropriateness rating for CR-GS was 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).

In Level 3, there was strong support for nine of the 13 domains, which included Mand (82%), Tact (85%), Listener Responding (80%), Independent Play (85%), SB-SP (80%), LRFFC (76), Intraverbal (73%), CR-GS (69.4%), and Linguistic Skills (77%). There was moderate support for Reading (63%) and Math (67%). There was limited support with fewer than half of the SMEs rating the age appropriateness for the items, on average, for VPS-MTS (49%) and Writing (46%). For the VPS-MTS domain, the percentage of SMEs who rated age appropriateness as "Very Appropriate" decreased as level increased. Where lower ratings for developmental age appropriateness were observed, they 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. A summary of the strength of evidence

Table 2 VB-MAPP content validity values

Domain	CVI	Minimum CVR	Maximum CVR	No. of Significant Items*
Level 1				
Mand	0.54	0.38	0.85	3
Tact	0.53	0.17	0.69	4
Listener responding	0.51	−0.08	0.85	4
Visual perceptual skills & matching-to-sample	0.25	−0.08	0.69	1
Independent play	0.20	−0.08	0.38	0
Social behavior & social play	0.20	−0.69	0.54	3
Motor imitation	0.35	0.23	0.54	1
Echoic	0.35	0.08	0.54	1
Spontaneous vocal behavior	0.23	−0.08	0.38	0
Level 2				
Mand	0.51	0.08	1.00	2
Tact	0.54	0.23	0.85	2
Listener responding	0.20	−0.23	0.83	1
Visual perceptual skills & matching-to-sample	0.17	−0.23	0.38	0
Independent play	0.38	0.08	0.85	2
Social behavior & social play	0.11	−0.08	0.23	0
Motor imitation	0.48	0.08	1.00	2
Echoic	−0.02	−0.38	0.23	0
Listener responding by function, feature, & class	0.29	0.08	0.69	1
Intraverbal	0.51	0.08	0.85	4
Classroom routines & group skills	−0.02	−0.08	0.08	0
Linguistic structure	0.42	−0.23	0.69	3
Level 3				
Mand	0.51	0.23	0.69	3
Tact	0.38	0.08	0.54	2
Listener responding	0.32	0.08	0.54	1
Visual perceptual skills & matching-to-sample	−0.09	−0.54	0.33	0
Independent play	0.38	−0.08	0.69	2
Social behavior & social play	0.45	0.38	0.54	2
Reading	0.14	−0.23	0.69	1
Writing	0.11	−0.23	0.38	0
Listener responding by function, feature, & class	0.48	−0.08	0.69	3
Intraverbal	0.32	0.23	0.54	1
Classroom routines & group skills	0.26	−0.23	0.69	1
Linguistic structure	0.32	0.08	0.69	1
Math	0.26	0.08	0.54	2

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

*Upper-tailed $p < 0.05$ for 13 SMEs

by domain and level for age appropriateness presented in Table 4.

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. Within Level 1, there was strong support (i.e., more than 69% of SME's rating the items as "Very Appropriate," on average) for only three of the nine domains; the domains included Motor Imitation (85%), Echoic (77%), and SVB (69%). There was moderate support for Mand (63.2%), Tact (68%), Listener Responding (62%), and VPS-MTS (60%). There was limited support for

Independent Play (49%) and SB-SP (43%) which indicates that on average, fewer than half of the SMEs indicated that the method of measurement was "Very Appropriate." Within Level 2, there was strong support for the Tact (78%), VPS-MTS (71%), Motor Imitation (77%), Echoic (74%), LRFFC (77%), and Intraverbal (77%) domains. There was moderate support for Mand (54%), Listener Responding (66%), Independent Play (57%), CR-GS (67%), and Linguistic Skills (65%) domains. There was limited support for SB-SP (33%). In Level 3, all 13 domains were classified as having strong or moderate evidence for method of measurement appropriateness. Interestingly, the strength of support for method of measurement appropriateness generally increased as developmental age increased. In addition, the domains with high average domain relevance ratings tended not to have high average method of measurement ratings, on average. In general, in the instances when method of measurement appropriateness ratings were slightly lower, SMEs explained that lower ratings 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).

The method of measurement appropriateness ratings were also compared across methods of measurement. Ratings were recoded based on the percentage of SMEs indicating that the method of measurement was "Very Appropriate." The estimated Cramer's V was 0.50 ($df = 3$, $p < 0.001$), which suggests there was a relatively strong association between the method of measurement and the appropriateness rating (Rea and 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 strength of evidence by domain and level for method of measurement appropriateness presented in Table 4.

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, only one domain, Motor Imitation, was rated as "Adequate" by the majority of SMEs (77%). There was moderate support for six of the domains (Mand, Tact, Listener Responding, SB-SP, Echoic, and SVB) with percentages ranging from 54 to 68%. The remaining two domains (VPS-MTS, Independent Play) had limited support where fewer than half of the SMEs rated the number and content of items as "Adequate." In Level 2, only one domain, VPS-MTS, had strong support where a majority of SMEs (69%) rated the domain representation as "Adequate." Six domains had moderate support (*range* = 50% to 62%), which included Mand, Tact, Independent Play, Echoic, Intraverbal, and CR-GS. The remaining five domains had limited support (*range* = 31% to 46%). In Level 3, there was strong support for four of the 13 domains (Mand, Listener Responding, Independent Play, Linguistic Skills) where on average, the majority of SMEs rated the domain representation as "Adequate" (*range* = 69% to 77%). There was moderate support (62%) for five of the 13 domains and limited support (range = 31% to 46%) for the remaining four domains. These results suggest that the number and/or content of the items may not adequately represent their domains across levels. A summary of the strength of evidence by domain and level for domain representation is presented in Table 4.

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 open-responses indicated that some criteria did not match the age level for 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.

Early Echoic Skills Assessment (EESA)

The calculated CVI for the EESA across groups was 0.35. That is, on average, 68% of SMEs rated the groups of items as "Essential," which indicates moderate support for its domain relevance. There was strong support for age appropriateness for the EESA, but 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 method of measurement (i.e., prompt [S^D]) appropriateness, there was moderate support across groups of items for the EESA. 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." SMEs commented that specific item content on EESA may be outside the expertise of behavior analysts. Thus, consulting with a speech-language pathologist may result in more accurate inferences regarding echoic skills.

VB-MAPP Barriers Assessment

The calculated CVI for the Barriers Assessment across categories was 0.60. 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. 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. The Barriers Assessment was not specific to any Level so the developmental age appropriateness was not collected from SMEs. A summary of the Barriers evidence is presented in Table 3, and the strength of all evidence categories are presented in Table 4 for the VB-Milestones, EESA, and Barriers Assessment.

Discussion

Teaching techniques and interventions based in ABA are the most empirically supported treatments to address the skill deficits in children with ASD (Axelrod et al. 2012; Foxx 2008; Lovaas 1987; National Autism Center 2015; Steinbrenner et al. 2020). Treatment plans are developed based on

Table 3 Barriers assessment content validity ratios and percentage of responses

Barrier	CVR	Domain Relevance			Measurement Appropriateness		
		N	*U*	*E*	*NA*	*SA*	*VA*
Negative behaviors	0.85*	0	8	92	8	31	62
Instructional control (Escape and avoidance of instructional demands)	0.85*	0	8	92	8	31	62
Absent, weak, or impaired Mand repertoire	0.85*	0	8	92	23	15	62
Absent, weak, or impaired tact repertoire	0.69*	0	15	85	23	15	62
Absent, weak, or impaired motor imitation	0.85*	0	8	92	15	23	62
Absent, weak, or impaired echoic repertoire	0.54*	0	23	77	23	31	46
Absent, weak, or impaired visual perceptual and matching-to-sample	0.54*	0	23	77	23	15	62
Absent, weak, or impaired listener repertoire (e.g. LD, LRFFC)	0.85*	0	8	92	23	15	62
Absent, weak, or impaired intraverbal repertoire	0.38	0	31	69	23	15	62
Absent, weak, or impaired social skills	0.69*	0	15	85	8	31	62
Prompt dependent	0.69*	0	15	85	8	23	69
Scrolling responses	0.38	0	31	69	8	31	62
Impaired scanning skills	0.69*	0	15	85	8	23	69
Failure to make conditional discriminations ($C^{D}{}_{S}$)	0.69*	0	15	85	15	15	69
Failure to generalize	0.69*	0	15	85	15	15	69
Weak or atypical motivating operations (MOs)	0.69*	0	15	85	8	23	69
Response requirement weakens the MO	0.54*	0	23	77	8	23	69
Reinforcement dependent	0.69*	0	15	85	8	23	69
Self-stimulation	0.69*	0	15	85	8	25	67
Articulation problems	0.08	0	46	54	0	31	69
Obsessive–compulsive behavior	0.23	0	38	62	8	31	62
Hyperactive behavior	0.23	0	38	62	8	23	69
Failure to make eye contact, or attend to people	0.38	0	31	69	8	31	62
Sensory defensiveness	0.54*	0	23	77	8	23	69

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

*Upper-tailed $p < 0.05$

Table 4 Summary of content validity evidence strength by level, domain, and category

Level	Domain	Domain relevance	Age appropriateness	Measurement appropriateness	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
	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

Shading used to aid in readability. 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

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 (2020b), approximately 80% of ABA professionals reported administering the VB-MAPP. Despite being the most widely used instrument, there were no studies until the current one that focused explicitly on collecting or evaluating validity evidence for the VB-MAPP (Ackley et al. 2019; Padilla et al. 2020). Thus, the VB-MAPP was 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.

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. For domains where there was limited and/or moderate support, the decisions made about an individual's skill repertoire based on the VB-MAPP results may not be fully reflective of the individual's actual skill repertoire. Thus, treatment plans or progress monitoring efforts may be incomplete without supplemental assessments measuring the same or related skills.

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" (Sect. 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)." For instance, given the varying levels of support for the Echoics, Linguistic Skills, and Listener Responding domains, it may be beneficial to supplement the assessment process with instruments evaluating receptive and language skills similar to that measured in those domains. Additionally, there was mostly limited-to-moderate support across categories for the SB-SP, Independent play, and Interverbal domains. Thus, supplementing the VB-MAPP with assessments that measure related skills such as socialization, communication, interaction, and play would provide complimentary information that could be helpful for treatment planning. Lastly, it is also important for test users to include other forms of assessment, such as caregiver and/or teacher interviews, reviews of educational and medical records, and direct observation.

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 Table 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.

Acknowledgments The author would like to thank Nicholas Benson, Sara Dolan, John Ferron, Grant Morgan, and Terrill Saxon for their helpful input for this study.

Author Contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Kristen L. Padilla. The first draft of the manuscript was written by Kristen L. Padilla and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

References

Ackley, M., Subramanian, J. W., Moore, J. W., Litten, S., Lundy, M. P., & Bishop, S. K. (2019). A review of language development protocols for individuals with autism. *Journal of Behavioral Education, 28,* 362–388.

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

Anastasi, A. (1988). *Psychological testing*. New York: 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.

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

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

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

Cramér, H. (1946). A contribution to the theory of statistical estimation. *Scandinavian Actuarial Journal, 1946*(1), 85–94.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart, and Winston.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). New York: American Council on Education.

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

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*. Carbondale: Shawnee Scientific Press.

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

Esch, B. E. (2014). Early echoic skills assessment (EESA). In M. L. Sundberg (Ed.), *VB-MAPP: Verbal behavior milestones assessment and placement program: A language and social skills assessment program for children with autism or other developmental disabilities* (2nd ed., pp. 42–48). Concord: AVB Press.

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

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.

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.

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

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

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.

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

Liebetrau, A. M. (1983). *Measures of association* (Vol. 32). Newbury Park: Sage.

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.

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.

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

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

Montallana, K. L., Gard, B. M., Lotfizadeh, A. D., & Poling, A. (2019). Inter-rater agreement for the milestones and barriers assessments of the verbal behavior milestones assessment and placement program (VB-MAPP). *Journal of Autism and Developmental Disorders, 49*(5), 2015–2023.

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

National Autism Center. (2015). *Findings and conclusions: National Standards Project, Phase2*. National Autism Center.

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*. Bombay: McGraw-Hill.

Padilla, K. L. (2020a). *Content validity evidence for the verbal behavior milestones assessment and placement program* (Document No. 11211). [Doctoral Dissertation, Baylor University]. ProQuest Dissertations & Theses Global.

Padilla, K. L. (2020b). Global assessment use and practices in applied behavior analysis: Surveying the field. *Research in Autism Spectrum Disorders*. https://doi.org/10.1016/j.rasd.2020.101676.

Padilla K. L., Morgan, G. B., Weston, R., Lively, P., & O'Guinn, N. (2020). Validity and reliability of behavior analytic assessments: A review of the literature. Unpublished manuscript.

Partington, J. W. (2006). *Assessment of basic language and learning skills, revised*. California: Behavior Analysts Inc.

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

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

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

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

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

Steinbrenner, J. R., Hume, K., Odom, S. L., Morin, K. L., Nowell, S. W., Tomaszewski, B., Szendrey, S., McIntyre, N. S., Yücesoy-Özkan, S., & Savage, M. N. (2020). *Evidence-based practices for children, youth, and young adults with Autism*. The University of North Carolina at Chapel Hill, Frank Porter Graham Child Development Institute, National Clearinghouse on Autism Evidence and Practice Review Team.

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*. Concord, CA: AVB Press.

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.

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.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.