Journal of Autism and Developmental Disorders (2019) 49:2015–2023
https://doi.org/10.1007/s10803-019-03879-4

ORIGINALPAPER

Inter-Rater Agreement for the Milestones and Barriers Assessments of the Verbal Behavior Milestones Assessment and Placement Program (VB-MAPP)

Khrystle L. Montallana[1] · Brendan M. Gard[2] · Amin D. Lotfizadeh[3] · Alan Poling[4]

Published online: 19 January 2019

Abstract
We determined inter-rater agreement for the VB-MAPP, an instrument sometimes used in planning educational goals and evaluating intervention effects for young people with autism. A pair of raters independently rated each of 32 children diagnosed with autism. Intraclass correlation coefficients for the total Milestones and Barrier scores were 0.876 and 0.629, respectively, indicating good and moderate reliability. There was variability in reliability in the different domains of the Milestones Assessment, with most indicating moderate reliability, and most of the individual Barriers Assessment domains indicating poor reliability. These are the first data relevant to the reliability of the VB-MAPP, they suggest that further evaluation of its reliability is merited and that a high reliability for individual domains should not be assumed.

Keywords Verbal Behavior Milestones Assessment and Placement Program · Language assessment · Verbal behavior · Autism spectrum disorder · Inter-rater agreement · Reliability

Introduction

There are a variety of assessment tools to set goals for people with Autism Spectrum Disorder (ASD) and to evaluate the effects of interventions that target those goals. The majority of such assessments emphasize adaptive behavior measures, which do not necessarily correspond with the core skill deficits in ASD (Stolte et al. 2016). ASD is characterized by deficits in social and communication skills, as well as repetitive or restrictive behaviors (American Psychiatric Association 2013). An assessment tool and curriculum guide that evaluates skill deficit areas associated with ASD that has received considerable attention in clinical practice is the Verbal Behavior Milestones Assessment and Placement Program (VB-MAPP; Sundberg 2008, 2014).

The VB-MAPP is an assessment tool based on Skinner's (1957) analysis of verbal behavior. It takes a functional and topographical approach to assessing a range of early development skills up the age of 4 years. The VB-MAPP is divided into three general assessment areas: Milestones Assessment, Barriers Assessment, and Transition Assessment. In addition to these assessments, the VB-MAPP contains a Task Analysis and Supporting Skills section that is used to guide the development of teaching curricula and provides suggestions for Individualized Education Program (IEP) goals. The Task Analysis and Supporting Skills along with the curriculum guide are not assessment tools, rather, they provide general recommendations and a list of approximately 900 developmentally sequenced skills that can be taught. We will not consider them further in this manuscript.

The VB-MAPP Milestones Assessment is used to assess language, social, academic, and other related skill areas. It

✉ Amin D. Lotfizadeh
amin.lotfizadeh@essc.org

Khrystle L. Montallana
khrystle.montallana@essc.org

Brendan M. Gard
brendan.gard@essc.org

Alan Poling
alan.poling@wmich.edu

[1] Autism Services Training Division, Easterseals Southern California, Valencia, CA, USA

[2] Autism Services Assessment Division, Easterseals Southern California, Oxnard, CA, USA

[3] Autism Services Research Division, Easterseals Southern California, Valencia, CA, USA

[4] Psychology Department, Western Michigan University, Kalamazoo, MI, USA

contains 16 domains and each domain consists of between 5 and 15 milestones. The domains and milestones are developmentally sequenced and are divided into three levels based on typical development during the first 4 years of life. Level 1 represents skills from 0 to 18 months, Level 2 represents skills from 18 to 30 months, and Level 3 represents skills from 30 to 48 months. The Milestones Assessment yields a maximum cumulative score of 170. The score earned by an individual indicates the total number of milestones for which a participant has met the listed criteria.

The VB-MAPP Barriers Assessment is used to evaluate various impediments to learning. It details 24 learning and language acquisition barriers (e.g., defective mand, a term used to refer to defective requesting skills; defective tact, a term used to refer to defective labeling skills; defective echoic). Each barrier is ranked on a 5-point Likert scale, with a score of 0 indicating that a particular barrier is not a problem and a score of 4 indicating that the barrier is a severe problem. The Barriers Assessment yields a cumulative score of 96, with a higher score indicating more overall barriers to learning.

The third area of assessment is the Transition Assessment, which provides a quantitative measure of overall skills to better assist with transitioning and school placement decisions. The transition assessment assesses three general categories of skills that are important for evaluating readiness to transition to less restrictive teaching environments. The three categories are language/social skills, learning patterns, and adaptive skills. Each category consists of six assessment areas that are either based on scores from specific areas on the Milestones Assessment, scores from the Barriers Assessment, or ratings on a series of 5-point Likert-scale questions.

Although the VB-MAPP components collectively provide an assessment tool and a curriculum guide for ASD interventions, a major limitation is that its psychometric properties are not well established (Carlson et al. 2017). With the exception of one study that evaluated the convergent validity of the VB-MAPP and the Promoting the Emergence of Advanced Knowledge Relational Training System (PEAK) assessment (Dixon et al. 2015), there are no published evaluations of the psychometric properties of the VB-MAPP. Dixon et al. administered two versions of the PEAK assessment, the PEAK Direct Training module (PEAK-DT; Dixon 2014) and the PEAK Generalization module (PEAK-G; Dixon et al.), along with the VB-MAPP to 40 individuals diagnosed with ASD. They summed the PEAK-DT and PEAK-G assessment scores for each participant to yield a cumulative PEAK-Combined score ranging from 0 to 386. The researchers calculated the Pearson correlation coefficient between the VB-MAPP and PEAK-Combined and concluded that the total VB-MAPP Milestones Assessment score was a strong predictor of the PEAK-Combined assessment total score. The results also showed that higher scores on the PEAK assessments corresponded with plateaued VB-MAPP scores. In other words, as PEAK scores increased, VB-MAPP scores approached and remained at a ceiling of 170. These findings suggest that the convergent validity of the two assessments is good.

The reliability of the VB-MAPP has not been reported. Reliability refers to the consistency of scores across repeated administrations of a testing tool, and high reliability is a feature of good assessment instruments (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education 2014). One form of reliability is inter-rater reliability, which refers to the degree of agreement between or among two or more raters who independently score the same individual or same instance of behavior. The current study evaluated the inter-rater reliability of the VB-MAPP Milestones Assessment and of the Barriers Assessment when used in clinical practice by trained clinicians familiar with verbal operants. Those components of the VB-MAPP were selected because they were routinely used to assess clients with ASD in the setting where the clinicians were employed.

Methods

Participants

We defined and recruited two types of participants: client participants and assessor participants. The number of participants was based on a power analysis, as described by Walter, Eliasziw, and Donner (1998), which yielded a sample size of 36 participants.

Client Participants

Client participants were individuals who had a primary diagnosis of ASD, as given by a physician or psychologist, were referred to a clinic for a behavioral assessment by the physician, and for whom a VB-MAPP assessment was recommended by the referring agency or the clinician supervising that client's case (who was the primary assessor participant). We only included clients who were referred for behavioral services or an assessment prior to receiving any applied behavior analysis (ABA) services from the organization where the primary assessor was employed. The reasoning for including only such clients was that, if we included participants who had already received behavioral services, the primary assessor would have been more familiar with the client than the second assessor, which may have resulted in the two assessors scoring the client differently. A total of 32 clients participated in the study. Their ages ranged from 1 to 9 years ($M = 3.5$, SD = 1.9). There were 27 male and five female client participants, all with a diagnosis of ASD;

Table 1 Participant demographic information

Demographic	*n* (%)
Gender	
Male	27 (84.4)
Female	5 (15.6)
Age group	
0–2 years	7 (21.9)
2–4 years	16 (50.0)
4–6 years	6 (18.8)
6–8 years	2 (6.2)
8–10 years	1 (3.1)
Diagnosis	
ASD only	30 (93.8)
Multiple	2 (6.2)

two participants were also diagnosed with comorbid conditions (one client was diagnosed with speech and language delay and another was diagnosed with exotropia, a form of eye misalignment, and amblyopia). The clients' average VB-MAPP Milestones and Barriers scores as measured by the primary assessors were 38.1 (SD = 32.1) and 40.8 (SD = 18.0), respectively. Table 1 provides more detailed demographic data regarding client participants.

Assessor Participants

Assessor participants were clinicians employed by a large Southern California organization that provided behavioral services and conducted assessments for clients with ASD and other developmental disabilities. The eligibility criteria for the assessor participants was based on the recommended guidelines for who should conduct the assessments in the VB-MAPP, as indicated in its manual. The eligibility criteria were as follows: (1) The assessors had independently conducted at least three VB-MAPP assessments while employed by the organization; (2) The assessors conducted assessments as a part of their regular job duties and, therefore, had completed the organization's online VB-MAPP training; (3) The assessors were Board Certified Behavior Analysts or held a master's degree from a behavior analysis program; (4) Assessors were in good professional standing in their current roles (i.e., no documented administrative or clinical concerns were reported to the organization that employed them).

There were a total of 24 assessors who participated in this study. Of the 24 assessors, 12 were scheduled to conduct a VB-MAPP assessment with a client who consented to participate and 12 were assessors who were paired with the primary assessors (see below for more details). Twenty-three assessors were BCBA® clinicians. The other assessor had received a master's degree in behavior analysis. All of the assessor participants were given a voluntary survey to complete. It requested information about their training in behavior analysis and certain demographic questions. Most respondents (21) indicated they had completed more than 12 VB-MAPP assessments, two respondents indicated that they had implemented 6–9 assessments, and one participant indicated that (s)he had implemented 3–6 VB-MAPP assessments. Nine assessors indicated that they had received formal VB-MAPP training and nine indicated they had not; six assessors did not provide a response to this question. Please see Table 2 for additional demographic information.

Setting and Materials

All assessments were conducted in the natural setting where the client participants engaged in their daily routines (e.g., home, daycare, park, community center). Assessment locations were based on the availability of client and assessor participants. The VB-MAPP Milestones and Barriers Assessment were scored using a laptop computer and the data were entered on the VB-MAPP Excel scoring sheet. As an incentive, we provided each assessor who completed the second VB-MAPP Milestones and Barriers Assessment for a client with a $50 gift card.

Procedures

Preliminary Training

Prior to the study, the organization arranged online VB-MAPP training for all clinicians, including the assessors in the present study. The training was self-paced and presented in a PowerPoint format with guided notes. The PowerPoint included information from the VB-MAPP manual and protocol booklets regarding the administration, methods of measurement, and scoring of the assessments. Specifically, the assessors were instructed to directly test and/or observe the

Table 2 Assessor demographic information

Demographic	*n* (%)
Assessors	
Primary	12 (50)
Secondary	12 (50)
Credential	
BCBA®	23 (95.8)
Master's degree only	1 (4.2)
Formal VB-MAPP training	
Yes	9 (50.0)
No	9(50.0)
Number of VB-MAPPs completed	
3–6	1(4.2)
6–9	2 (8.3)
More than 12	21 (87.5)

listed milestones based on the listed scoring criteria. At the conclusion of the training, there was a 17-question multiple-choice quiz assessing understanding of the procedures. The participants were required to obtain a score of 80% or higher, otherwise they were prompted to retake the training and quiz until they reached that criterion. The average score on the training across all assessor participants was 91%.

Demographic Survey

Prior to each assessment, we provided each of the assessors with a survey containing 16 demographic questions and nine multiple-choice questions related to the definitions of four commonly taught verbal operants (i.e., mand, tact, intraverbal, and echoic). The purpose of the questions related to the verbal operants was to verify that the participants had a general understanding of some of the verbal operants listed in the VB-MAPP. The average score on this quiz was 87.6%.

Assessment Administration

After a client received authorization for an assessment, the primary assessor presented the client's legal guardian with a flyer describing the study. If the guardian consented to participate and if the primary assessor deemed the VB-MAPP to be the appropriate assessment tool for that particular client, we randomly selected a second assessor from the pool of second assessors in the same geographic region. We assigned the second assessor randomly by assigning a number to each assessor and generating a number from a random number service website (http://www.random.org). If the first selected assessor could not complete the assessment, this process was repeated until an assessor was available to complete the assessment. Once the assessor pair was identified, both assessors were instructed to begin the Milestones and Barriers Assessments. The assessors were instructed to conduct the first day of testing within 2 weeks of one another and finish their individual assessment within 21 days of the first day of testing. For example, if the first assessor began administering the VB-MAPP on January 1st and the second assessor began administration on January 5th, the respective completion dates for the VB-MAPP assessments would be January 21st and January 26th. All assessor participants conducted the assessment during separate appointments so that the two assessors were not in direct contact at any point during an assessment.

Milestones Scoring

Both assessors conducted the Milestones Assessment as recommended by Sundberg (2014). Because the VB-MAPP assessment was administered for clinical treatment planning purposes, assessors were instructed to stop testing a single domain after they obtained three consecutive scores of zero. The VB-MAPP Milestones Assessment was administered through formal testing, observation, and/or timed observations. After testing each milestone, a score was generated for each milestone and all milestones' scores were summed to yield a total milestones score. Of the 170 milestones, a majority of them ($n = 166$) have possible scores of "0," "½," or "1," while the remaining four milestones only have possible scores of "0" or "1" (i.e., there is no "½" score). The score that is assigned to each milestone is based on whether or not the client participant has met the criteria listed for each milestone. For example, the assessors assigned a score of "0" for a milestone within a domain if the listed milestone criteria were not met, a score of "½" if the client participant met the criteria for the skills listed as half a score, and a score of "1" if the client participant met the full criteria listed for that milestone.

Barriers Scoring

Both assessors also completed the Barriers Assessment with each client participant. The assessor participants ranked each barrier with a score of 0 (no problem), 1 (occasional problem), 2 (moderate problem), 3 (persistent problem), or 4 (severe problem), based on the assessor's observations of the barriers. The VB-MAPP Guide and Protocol (Sundberg 2014) booklets provide additional scoring criteria for the Barriers Assessment that include some operational definitions for rating each of the barriers. The assessor participants used those operational definitions and rated each barrier accordingly. Finally, the participants added the individual barrier scores to obtain an overall barrier score.

Measurement and Analysis

We measured inter-rater reliability using intraclass correlation coefficients (ICC). ICCs are well established measures of inter-rater agreement (Shrout and Fleiss 1979). We calculated ICCs for the total Milestones Assessment score, for each domain within the Milestones Assessment, for the total Barriers Assessment score, and for each specific barrier listed in the Barriers Assessment. For the ICC calculations we utilized a single-rating, absolute-agreement, two-way random effect model as recommended by Koo and Li (2016). Accordingly, we considered ICC values of less than 0.5 as poor reliability, 0.5–0.74 as moderate, 0.75–0.9 as good, and values greater than 0.9 as indicative of excellent reliability. Alternatively, a less stringent strength of agreement criteria, proposed by Landis and Koch (1977), considers ICC values of less than 0.00 as poor, 0.00–0.20 as slight, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–80 as substantial, and 81–1.00 as almost perfect. Hereafter, we will utilize the criteria proposed by Koo and Li (2016).

The Transition Assessment was not evaluated in this study because not all of the participants were eligible for such assessments at the start of their treatments. For example, some participants were not old enough to require an IEP assessment or transitioning was not a clinical priority at the start of services, therefore, those components of the VB-MAPP were not conducted for all clients by the organization at intake.

Interobserver Agreement (IOA) for Data Entry

In order to assure that data entry was conducted in a reliable manner, we measured the agreement in data entry for 33% of the assessor pairs. Two researchers independently inputted the assessment data for a particular client into an Excel data sheet. We compared each of the corresponding data entry boxes on the Excel sheet and, if the two researchers inputted data the same, it was considered an agreement. If the corresponding data box on the Excel sheet did not match across the two researchers inputting data, it was considered a disagreement. Next, we divided the total agreements by the total disagreements and multiplied the result by 100. The IOA coefficient for data entry was 98.7%.

Results

The intraclass correlations (ICC), with 95% confidence intervals and results of statistical comparison to chance values, are reported in Table 3. The ICC for the total milestones score was 0.876 ($p < 0.001$), which indicates that it has good reliability as defined by Koo and Li (2016). ICCs for all of the individual milestones, except Group score, exceeded chance values at the $p < 0.01$ level. Figure 1 shows the percentage of ICCs that indicated poor (6.3%), moderate (56.3%), good (31.3%), and excellent (6.3%) reliability. Most indicated moderate or good reliability.

The ICC for the total barriers score was 0.629 ($p < 0.001$), which indicates moderate reliability. ICCs for 10 of the 24 individual barriers (41.7%) exceeded chance levels at $p < 0.01$ level, and for 5 others (20.8%) at the $p < 0.05$ levels. With respect to barriers, 87.5% were of poor reliability and the rest were of moderate reliability, as shown in Fig. 1.

Discussion

Although no relevant data have been published, recent articles report that the VB-MAPP is frequently used to make educational decisions regarding children with ASD and other developmental disabilities. For example, Barnes et al. (2014) reported that "Many educational settings use the instrument to establish language goals and objectives for individuals with autism spectrum disorder and other developmental disabilities" (p. 36). Dixon et al. (2018) reached a similar conclusion. They wrote, "One assessment tool that is commonly used to identify the language deficits experienced by individuals with autism is the Verbal Behavior Milestones and Placement Program…" (p. 224).

Recent studies have also used the VB-MAPP as an outcome measure in published research (e.g., Lotfizadeh et al. 2018). As noted previously, Dixon et al. (2015) compared scores on two versions of the PEAK assessment to scores on the VB-MAPP. Dixon et al. (2018) used the VB-MAPP as one measure of the effectiveness of PEAK training, and Dunne et al. (2014) used it as one measure of the effects of providing relational frame training for young children with autism. Grannan and Rehfeldt (2012) used VB-MAPP scores to set instructional goals for their participants and Gunby et al. (2010) used the VB-MAPP to index their participants' general verbal repertoires. None of these studies reported inter-rater agreement for the VB-MAPP, although the studies were in the vein of applied behavior analysis and it has long been common practice for researchers in this area to determine inter-rater agreement when they report target behavior (Page and Iwata 1986).

Reliability is a key aspect of the quality of an assessment instrument (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education 2014), and inter-rater agreement is a meaningful aspect of reliability (Watkins and Pacheco 2000). Given that no prior reports of inter-rater agreement for the VB-MAPP, or other indices of its reliability, have appeared, the present results should interest people who are using, or plan to use, the instrument.

The VB-MAPP has a number of characteristics that make it a useful instrument for planning and evaluating behavioral interventions for young people with ASD (Lotfizadeh et al. 2018; Gould et al. 2011). As Gould et al. indicated, "The greatest limitation of the VB-MAPP is the lack of psychometric evaluation. Sufficient reliability and validity of assessments is not a default assumption, but rather, a consideration that requires empirical evaluation" (p. 998).

The present results suggest reliability of the Total Milestones Assessment of the VB-MAPP, in terms of inter-rater agreement, was good, as indicated by the relatively high ICC coefficient. That was not the case for the individual barriers. There was substantial variability in reliability in the different domains of the Milestones Assessment. Reliability for the majority of them was moderate according to the criteria espoused by Koo and Li (2016). Reliability was good or excellent for the mand, tact, intraverbal, echoic, imitation, and writing domains and poor for the groups domain.

The ICC for the total Barriers Assessment was moderate, but the reliability of the individual domains of the Barriers Assessment was unimpressive. None of them had good

Table 3 Results of intra-class correlation coefficient (ICC) calculations in SPSS using single-rating, absolute-agreement, 2-way random effect model

VB-MAPP assessment area		Assessor 1: *M* (SD)	Assessor 2: *M* (SD)	95% Confidence interval			*F* test with true value 0			
				Intraclass correlation	Lower bound	Upper bound	Value	*df*	*df2*	*p*
ICC rank	Milestones score									
	VB-MAPP milestones total	38.1 (32.1)	37.0 (34.0)	0.876	0.760	0.938	14.790	29	58	< 0.001
1	Tact	3.4 (4.1)	3.3 (4.0)	0.940	0.879	0.970	31.457	29	58	< 0.001
2	Mand	2.8 (3.3)	3.3 (4.1)	0.864	0.739	0.932	14.138	29	58	< 0.001
3	Intraverbal	0.8 (2.1)	0.9 (2.2)	0.859	0.727	0.929	12.786	29	58	< 0.001
4	Echoic	2.8 (3.5)	3.2 (3.6)	0.843	0.701	0.921	11.671	29	58	< 0.001
5	Imitation	2.5 (2.8)	2.4 (2.9)	0.819	0.658	0.909	9.828	29	58	< 0.001
6	Writing	0.5 (1.5)	0.4 (1.3)	0.799	0.623	0.898	8.722	29	58	< 0.001
7	Vocal	2.9 (1.9)	3.2 (1.7)	0.697	0.462	0.841	5.580	29	58	< 0.001
8	Listener	3.4 (3.3)	2.9 (3.2)	0.672	0.426	0.827	5.126	29	58	< 0.001
9	LRFFC	0.7 (2.4)	0.4 (1.8)	0.627	0.359	0.800	4.375	29	58	< 0.001
10	Linguistics	1.0 (1.8)	1.5 (2.4)	0.624	0.358	0.798	4.467	29	58	< 0.001
11	VP/MTS	5.5 (2.5)	5.0 (2.5)	0.614	0.341	0.792	4.209	29	58	< 0.001
12	Math	0.5 (1.2)	0.7 (1.3)	0.612	0.336	0.792	4.122	29	58	< 0.001
13	Social	3.4 (2.4)	3.1 (2.9)	0.577	0.283	0.771	3.664	29	58	< 0.001
14	Reading	0.8 (1.6)	0.8 (1.5)	0.570	0.270	0.767	3.561	29	58	< 0.001
15	Play	6.7 (3.3)	5.6 (3.5)	0.546	0.252	0.750	3.543	29	58	< 0.001
16	Group	0.5 (1.7)	0.4 (1.2)	−0.116	−0.464	0.252	0.798	29	58	0.73
	Barrier score									
	Total barriers	40.8 (18.0)	34.5 (16.0)	0.629	0.339	0.806	4.985	29	58	< 0.001
1	Defective echoic	2.4 (1.7)	2.0 (1.8)	0.734	0.519	0.862	6.775	29	58	< 0.001
2	Reinforcer dependent	1.6 (1.6)	1.1 (1.2)	0.594	0.303	0.782	4.337	29	58	< 0.001
3	Defective listener	2.1 (1.5)	1.7 (1.6)	0.559	0.268	0.758	3.629	29	58	< 0.001
4	Hyperactive behavior	1.5 (1.5)	1.3 (1.2)	0.495	0.179	0.719	2.957	29	58	0.002
5	Defective mand	2.7 (1.3)	2.7 (1.7)	0.479	0.165	0.707	2.897	29	58	0.002
6	Defective articulation	1.8 (1.6)	2.1 (1.7)	0.467	0.142	0.702	2.739	29	58	0.004
7	Defective tact	2.1 (1.8)	1.6 (1.8)	0.439	0.117	0.680	2.630	29	58	0.005
8	Failure to make eye contact	1.5 (1.4)	1.7 (1.2)	0.424	0.084	0.675	2.439	29	58	0.009
9	Defective social	2.7 (1.4)	1.9 (1.8)	0.423	0.098	0.670	2.740	29	58	0.004
10	Defective imitation	2.2 (1.4)	1.8 (1.6)	0.413	0.082	0.664	2.429	29	58	0.009
11	Scrolling	0.5 (0.9)	0.4 (1.0)	0.361	0.009	0.632	2.105	29	58	0.023
12	Behavior problems	2.0 (1.2)	1.8 (0.8)	0.356	0.009	0.627	2.099	29	58	0.023
12	Weak motivators	1.6 (1.5)	1.2 (1.4)	0.356	0.019	0.624	2.137	29	58	0.021
13	Self-stimulation	1.1 (1.1)	0.8 (1.1)	0.304	−0.042	0.588	1.889	29	58	0.043

Table 3 (continued)

VB-MAPP assessment area		Assessor 1: *M* (SD)	Assessor 2: *M* (SD)	95% Confidence interval			*F* test with true value 0			
				Intraclass correlation	Lower bound	Upper bound	Value	*df*	*df2*	*p*
14	Defective scanning	1.3 (1.6)	1.5 (1.4)	0.296	−0.066	0.587	1.821	29	58	0.053
15	Defective intraverbal	2.1 (1.9)	1.9 (1.7)	0.282	−0.080	0.576	1.768	29	58	0.062
16	Defective VP-MTS	1.5 (1.5)	0.8 (1.3)	0.280	−0.043	0.562	1.887	29	58	0.044
17	Obsessive-compulsive behavior	1.1 (1.4)	0.8 (1.1)	0.246	−0.106	0.546	1.660	29	58	0.085
18	Prompt dependent	2.0 (1.5)	1.9 (1.5)	0.234	−0.134	0.542	1.595	29	58	0.103
19	Instructional control	2.0 (1.1)	1.8 (0.9)	0.214	−0.143	0.523	1.545	29	58	0.119
20	Failure to generalize	0.8 (1.3)	0.6 (1.1)	0.194	−0.166	0.508	1.479	29	58	0.144
21	Response requirement weakens MO	2.2 (1.4)	2.2 (1.2)	0.165	−0.209	0.491	1.381	29	58	0.191
22	Sensory defensiveness	0.8 (1.2)	0.5 (1.0)	0.076	−0.281	0.414	1.163	29	58	0.341
23	Defective conditional discriminations	1.4 (1.6)	1.0 (1.3)	0.016	−0.336	0.364	1.032	29	58	0.465

Domains with the same ICC values were ranked the same

or excellent reliability. Reliability was moderate for the impaired echoic, impaired listener, and reinforcer dependent domains. Reliability was poor for the other 21 barriers. A possible reason why the domains have poor reliability is that some of the individual barriers do not have clear operational definitions. For example, the scoring criteria to assess prompt dependency states, "Give the child a score of 0 on the Barriers Assessment if he is consistently learning new skills and does not show any signs of prompt dependency" (Sundberg 2008, p. 114), but it does not define "consistently" or "any signs" with measurable criteria.

The present findings suggest that the reliability of the individual domains comprised by the Barriers Assessment of the VB-MAPP, in terms of inter-rater agreement, is not good and the reliability of the individual domains of the Milestones Assessment is moderate. These findings suggest that it is appropriate to report inter-rater agreement when the VB-MAPP is used in research settings, and to exercise caution in making important instructional decisions based on VB-MAPP scores yielded by a single rater. These findings also suggest that further examination of the reliability of the VB-MAPP is justified, especially in view of the limitations of the present study.

An obvious, and important, limitation of our study is that the trainers may have received inadequate training. The VB-MAPP manual and protocol booklets provide information regarding the administration, methods of measurement, and scoring of the assessments, and the on-line training our assessor participants received was based on this information. All of them passed a post-test that we devised to index mastery of the material. It is probable, but not assured, that the characteristics of our assessors are similar to those of practitioners who regularly use the VB-MAPP in clinical settings. It is also probable that more rigorous training on how to implement the VB-MAPP will influence the reliability of the results it yields.

Evidence that rigor of training influences how the VB-MAPP is administered was recently provided by Barnes et al. (2014), who reported that two school psychologists who were "asked to read the VB-MAPP protocol and guide book in order to prepare to implement Levels 1 and 2 of the Milestones Assessment" did not consistently implement the assessment with accuracy. Their performance substantially improved following behavioral skills training that comprised five components (instruction, modeling, rehearsal and feedback, and remedial teaching as needed). Although all of our assessor clients received on-line training in accordance with information provided with the VB-MAPP, only half of them indicated that they had received formal VB-MAPP training similar to that outlined by Barnes et al. It is impossible to know how our training compared to theirs, and our data were not collected in a manner that allow for the effects of training on reliability

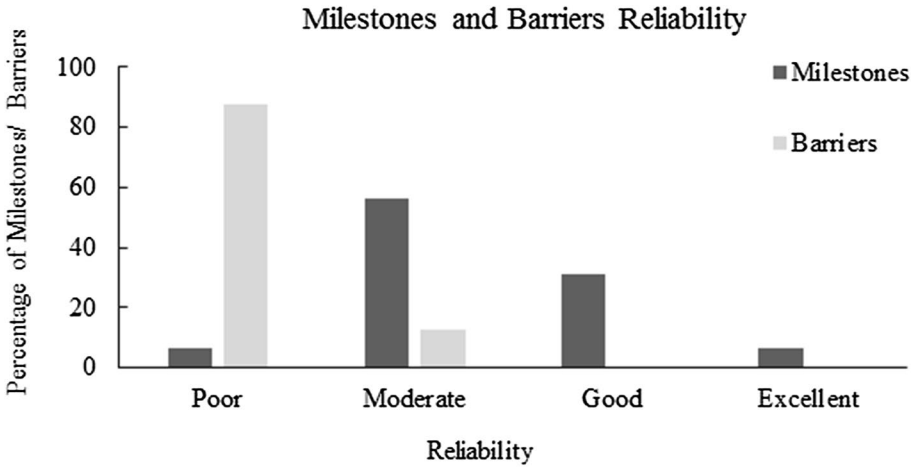

Fig. 1 Percentage of VB-MAPP Milestones and Barriers Assessment domains with poor ICC (less than 0.5), moderate ICC (between 0.5 and 0.75), good ICC (between 0.75 and 0.9), or excellent ICC (above 0.9)

to be determined. Determining IOA for assessors trained in the manner described by Barnes et al., and in other, less rigorous, ways that would be easier to arrange in most clinical settings, is a worthy task for future research. So, too, is examining IOA across a variety of client participants.

In addition to the specific training that our assessors received, other factors may have influenced our findings. All of our assessors were BCBA®s who had substantial, but not extensive, knowledge of client participants. Although this was by design, perhaps greater familiarity with clients may have resulted in higher reliability because more exposure to the clients would allow for more observation of a skill. Moreover, all of our clients were relatively young people (1–9 years of age) with an ASD diagnosis who had relatively low skill levels, which may have affected the obtained results and limited the generality of our findings. Therefore, additional studies are warranted to evaluate the reliability of VB-MAPP when administered to a range of clients by diverse assessors.

In conclusion, the present study is the first formal assessment of the inter-rater reliability of the VB-MAPP. Although the overall Milestones Assessment and Barriers Assessment provided good and moderate reliability, respectively, the individual domains within each assessment were less reliable. Additional research is needed to ascertain the generality of these findings, and to determine whether the psychometric properties of the VB-MAPP make it an appropriate instrument for selecting goals for students with ASD and for evaluating the effects of interventions intended to attain those goals.

Acknowledgments Special thanks to Alyssa Kavner for her assistance with data entry and IOA calculations.

Author Contributions All authors contributed equally to this study and to the preparation of this manuscript.

Compliance with Ethical Standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical Approval All procedures performed were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. This study was approved by a Human Subjects Institutional Review Board (HSIRB).

Informed Consent Informed consent was obtained from all individual participants included in the study prior to the start of services.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th edn.). Arlington: American Psychiatric Publishing.

Barnes, C. S., Mellor, J. R., & Rehfeldt, R. A. (2014). Implementing the verbal behavior milestones assessment and placement program (VB-MAPP). *Analysis of Verbal Behavior, 30*, 36–47. https://doi.org/10.1007/s40616-013-0004-5.

Carlson, J. F., Geisinger, K. F., & Jonson, J. L. (Eds.). (2017). *The twentieth mental measurements yearbook*. Lincoln: Buros Center for Testing.

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

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*, 223–233. https://doi.org/10.1007/s10882-014-9410-4.

Dixon, M. R., Wiggins, S. H., & Belisle, J. (2018). The effectiveness of the peak relational training system and corresponding changes on

the VB-MAPP for young adults with autism. *Journal of Applied Behavior Analysis, 51*, 321–334. https://doi.org/10.1002/jaba.448.

Dunne, S., Foody, M., Barnes-Holmes, Y., Barnes-Holmes, D., & Murphy, C. (2014). Facilitating repertoires of coordination, opposition distinction, and comparison in young children with autism. *Behavioral Development Bulletin, 19*, 37–47. https://doi.org/10.1037/h0100576

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 behavioral intervention programs for autism spectrum disorders. *Research in Autism Spectrum Disorders, 5*, 990–1002. https://doi.org/10.1016/j.rasd.2011.01.012.

Grannan, L., & Rehfeldt, R. (2012). Emergent intraverbal responses via tact and match-to-sample instruction. *Journal of Applied Behavior Analysis, 45*, 601–605. https://doi.org/10.1901/jaba.2012.45-601.

Gunby, K. V., Carr, J. E., & LeBlanc, L. A. (2010). Teaching abduction-prevention skills to children with autism. *Journal of Applied Behavior Analysis, 41*, 107–*112*. doi:0.1901/jaba.2010.43-107.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*, 155–163. https://doi.org/10.1016/j.jcm.2016.02.012.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.

Lotfizadeh, A. D., Kazemi, E., Pompa-Craven, P., & Eldevik, S. (2018). Moderate effects of low-intensity behavioral intervention. *Behavior Modification*. https://doi.org/10.1177/0145445518796204.

Page, T. J., & Iwata, B. A. (1986). Interobserver agreement. In A. Poling & R. W. Fuqua (Eds.), *Research methods in applied behavior analysis* (pp. 99–126). Boston: Springer.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420–428. https://doi.org/10.1037/0033-2909.86.2.420.

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

Stolte, M., Hodgetts, S., & Smith, V. (2016). A critical review of outcome measures used to evaluate the effectiveness of comprehensive, community based treatment for young children with ASD. *Research in Autism Spectrum Disorders, 23*, 221–234. https://doi.org/10.1016/j.rasd.2015.12.009.

Sundberg, M. L. (2008). *Verbal behavior milestones assessment and placement program: The VB-MAPP*. Concord: AVB Press.

Sundberg, M. L. (2014). *The verbal behavior milestones assessment and placement program: The VB-MAPP* (2nd edn.). Concord: AVB Press.

Walter, S. D., Eliasziw, M., & Donner, A. (1998). Sample size and optimal designs for reliability studies. *Statistics in Medicine, 17*(1), 101–110. https://doi.org/10.1002/(SICI)1097-0258(19980115)17:1%3C101::AID-SIM727%3E3.0.CO;2-E.

Watkins, M. W., & Pacheco, M. (2000). Interobserver agreement in behavioral research: Importance and calculation. *Journal of Behavioral Education, 10*, 205–212. https://doi.org/10.1023/A:1012295615144.