J Dev Phys Disabil (2018) 30:27–38
https://doi.org/10.1007/s10882-017-9574-9

ORIGINAL ARTICLE

Using Expert Panels to Examine the Content Validity and Inter-Rater Reliability of the *ABLLS-R*

Jennifer Usry[1] · Scott W. Partington[2] · James W. Partington[2]

Published online: 17 November 2017

Abstract The assessment literature cites several instruments used to assess the skills of children with an autism spectrum disorder (ASD) diagnosis, but many lack adequate empirical support for their psychometric properties. The *Assessment of Basic Language and Learning Skills-Revised* (*ABLLS-R*) is a popular assessment used by clinicians to measure the skills of children with ASD. Despite its widespread use, the *ABLLS-R* contains limited research on its psychometric properties. The current study sought to extend the recent research on the psychometric properties of the *ABLLS-R* by using data obtained from two separate panels of expert raters to evaluate its content validity and the inter-rater reliability of its scores. Our results demonstrate evidence of content validity as at least five out of six expert panel members rated 441 out of the 544 *ABLLS-R* items (or 81% of the items in the assessment) as "essential." We also found evidence of excellent inter-rater reliability (intraclass correlation coefficient = .95, $p < .001$) across the *ABLLS-R* scores obtained from a second panel of expert raters. These findings extend the existing literature and further document the *ABLLS-R* as a valid instrument that yields reliable scores.

Keywords Autism · Assessments · *ABLLS-R* · Reliability · Validity · Psychometric properties

The prevalence of autism spectrum disorder (ASD) has more than doubled since 2002 and now impacts one in every 68 children in the United States (Center for Disease Control and Prevention 2015). The alarming increase in the prevalence of ASD further highlights the increasing need for effective intervention and teaching strategies. Given that individuals with ASD display a wide range of skill deficiencies and behaviors, researchers encourage the use

✉ Jennifer Usry
jthreatt@liberty.edu

[1] Liberty University, Lynchburg, VA, USA

[2] Behavior Analysts, Inc., 309 Lennon Lane Suite 104, Walnut Creek, CA, USA

of and characterize best practice as accurately assessing the strengths and skill deficits of each client to determine appropriate teaching strategies (Guldberg 2010). In addition, both professional organizations (American Educational Research Association et al. 2014) and state and federal law (Every Student Succeeds Act 2015; Individuals with Disabilities Education Act 2004) call for the use of valid and reliable assessments.

The literature cites several assessment tools available to parents and professionals (e.g., behavior analysts, speech and language pathologists, educators, etc.) who serve children with ASD. Unfortunately, many criterion-referenced assessments review a wide range of skills, but contain limited empirical support for their psychometric properties (e.g., *Verbal Behavior Milestones Assessment and Placement Program*; Sundberg 2008). In contrast, other assessments contain evidence of validity and reliability, but fail to review an extensive range of essential skills (e.g., *Vineland II*; Sparrow et al. 2005). To engage in best practice necessitates the use of a comprehensive assessment that reviews a wide range of skills and contains an appropriate level of empirical support for its psychometric properties.

Professional organizations and leading researchers in the field of behavior analysis recognize the *Assessment of Basic Language and Learning Skills-Revised* (*ABLLS-R*; Partington 2010a) as a powerful tool that guides parents and professionals seeking to teach language and critical learner skills to individuals with ASD (Aman et al. 2004; American Medical Association 2014; Schwartz et al. 2001; Thompson 2007; Thompson 2011). The *ABLLS-R* is a criterion-referenced assessment comprised of 544 items and provides a comprehensive review of 25 skill areas (i.e., repertoires) including language, social interaction, academic, self-help, and motor skills. Its international recognition and widespread use by parents and professionals demonstrate its strong clinical significance while recent research found that the *ABLLS-R* yields reliable scores (Partington et al. 2016) and contains evidence of validity (Malkin et al. 2016). Despite these positive attributes, the *ABLLS-R* contains limited empirical support for its psychometric properties and would benefit from further research in this area.

At present, only two published studies examined the psychometrics of the *ABLLS-R*. One research team obtained evidence of convergent validity by showing that *ABLLS-R* scores strongly correlate with scores obtained from the *Vineland II* and *Promoting the Emergence of Advanced Knowledge Relational Training System-Direct Training Module* assessments (Malkin et al. 2016). A second study obtained strong evidence of test-retest and internal consistency reliability—a finding that demonstrates that the *ABLLS-R* yields reliable scores (Partington et al. 2016). While these research reports establish empirical support for the validity of the *ABLLS-R* and the reliability of its scores, the assessment literature would benefit from examining other forms of validity and reliability. The current study sought to extend the research by Malkin et al. (2016) and Partington et al. (2016) by obtaining measures of other psychometric properties of the *ABLLS-R* including its content validity and the inter-rater reliability of its scores.

Method

Participants

Prior to recruiting our participant sample, we consulted the literature to determine the appropriate number of individuals that should comprise each expert panel in order to establish evidence of content validity and inter-rater reliability. Based on the guidelines set forth by Salkind (2013), establishing evidence of content validity would require ratings from at least two subject matter experts. However, the number of participants needed to establish evidence of inter-rater reliability appears unclear as researchers do not adhere to a common set of guidelines (Wolak et al. 2012). These resources indicate that we need to recruit *a minimum* of two participants per expert panel, but ultimately, we decided to include data from as many qualified participants as possible in order to better meet our research objectives. With these parameters in mind, we then began taking steps toward establishing our participant sample.

We first identified the characteristics that members of each panel should possess and the specific criteria that they needed to meet for inclusion onto one of our two expert panels. Since we used data obtained from the first expert panel to measure the content validity of the *ABLLS-R*, we sought to include individuals onto our panel that possessed a wealth of experience with working with individuals with ASD and were familiar with the *ABLLS-R*. To recruit individuals with these characteristics, the researchers established and required participants to meet two main sets of inclusion criteria to serve on our first expert panel. First, we emphasized the role of experience and required that all participants either possessed certification as a practicing, certified behavior analyst (i.e., BCBA or BCBA-D level) with at least five years of applied experience or we also accepted individuals without BCBAs provided that they each possessed at least seven years of applied experience. This autism experience-based requirement helped to ensure that the participants recruited onto our panel possessed expert level knowledge on common areas of skill deficit. We also required that participants met a second set of criteria to ensure that they received quality training on how to administer the *ABLLS-R* and that they possessed experience with using it. Specifically, we required that all participants received previous training on how to administer the *ABLLS-R* from a qualified professional (i.e., a behavior analyst or behavioral consultant), interacted with the *ABLLS-R* in some capacity (e.g., using the *ABLLS-R* to select teaching objectives, referencing the *ABLLS-R* when teaching specific skills, administering skills assessments, etc.) for at least five years, and that they independently administered the *ABLLS-R* at least once prior to participating in our study (note that we did *not* require them to regularly use the *ABLLS-R* to assess skills as doing so could potentially bias their responses). Those who met or exceeded both of the above sets of criteria qualified for inclusion onto our first expert panel.

We sought to recruit a second expert panel to examine the inter-rater reliability of their *ABLLS-R* scores. The *ABLLS-R Scoring Instructions and IEP Development Guide* (Partington 2010b) specifically noted that individuals known to the student, including parents and professionals, may administer an *ABLLS-R* assessment. This description implies that individuals without formal training can conduct an *ABLLS-R* assessment. Thus, we found it imperative to require individuals on our second expert panel to possess no previous experience with using or administering the *ABLLS-R*, but whose professional career put them into direct contact with individuals with ASD. We also required that participants on the second expert panel possessed a college level degree or higher and that they either worked in an educational setting or delivered autism treatment-related services. Individual who met these specific criteria qualified for inclusion onto our second expert panel.

We consulted with our professional colleagues to help us identify prospective participants for inclusion onto one of our two expert panels based on the respective, aforementioned criteria. After receiving their feedback, we contacted all prospective participants via email to inform them of the study. We recruited participants throughout the United States from organizations and public schools, including consulting and support personnel in special education departments, and individuals who work in classrooms that used techniques from the field of behavior analysis to teach skills to children with ASD.

We initially emailed 13 prospective participants to inform them of our study and gauge their interest in serving on one of the two expert panels. Specifically, we attempted to recruit eight individuals for inclusion onto the first expert panel and five individuals for inclusion onto the second expert panel. Two of the individuals that we attempted to recruit onto our first expert panel either did not respond to our email or declined our invitation to participate. All of the other prospective participants responded to our email invitation and expressed their willingness to participate in our study. These individuals received and completed a prescreening questionnaire that allowed us to confirm their eligibility for inclusion onto one of our expert panels. We then assigned all eligible participants (N = 11) to either the first expert panel (n = 6; three men and three women; M_{age} = 40.83 years) or the second expert panel (n = 5; one man and four women; M_{age} = 39.40 years) based on their experience and qualifications. For additional information on the demographics and other characteristics of the members comprising the first (see Table 1) and second (see Table 2) expert panels, we refer the reader to the tables provided.

Although we only analyzed data obtained from our two expert panels, we also recruited two other participants to aid in the development of video clips for use in the inter-rater reliability component of the study —an eight-year-old child diagnosed with ASD and his case supervisor (i.e., a BCBA). At the time of the study, the child with ASD possessed some, albeit limited language skills. All participants partook in our research on a voluntary basis and did not receive compensation for their participation.

Materials

All participants in both expert panels received a link to complete a panel-specific survey using an online survey program, Survey Monkey (www.surveymonkey.com). The survey administered to members of the first expert panel consisted of all 544 *ABLLS-R* items, the associated scoring criteria for each *ABLLS-R* item, and a rating scale to determine the extent that they considered the skill measured by each *ABLLS-R* item as “essential.”

Members of the second expert panel received excerpts from the *ABLLS-R Scoring Instructions and IEP Development Guide* that described how to score the assessment. The online survey they received thereafter contained 86 different *ABLLS-R* items and their respective scoring criteria so that the panel members could score the actual performance of the student depicted in each of the 86 video clips ($M_{duration}$ = 105.05 s) that they were sent via email.

To develop the video clips that we distributed to members of the second expert panel, the first author contacted the clinical director from a private company that serves children with ASD and received permission from one of their clients—the parents of a

Table 1 Demographics and Related Information Pertaining to the Members of the First Expert Panel

	Degree	Occupation	Race	Age	Gender	Applied Experience (in yrs.)	Geographic Location
Participant 1	M.S., BCBA	Public School Program Director	Hispanic	46	Male	9	California
Participant 2	M.A., BCBA	Public School Program Coordinator	White	33	Female	5	California
Participant 3	Ph.D., BCBA	Independent Executive Program Director	White	43	Male	19	Georgia
Participant 4	M.S., BCBA	Senior Behavior Analyst	White	35	Male	9	Georgia
Participant 5	B.S.	Applied Behavior Therapist (worked in ABA classroom)	White	39	Female	9	Georgia
Participant 6	None	Applied Behavior Therapist (worked in ABA classroom)	White	49	Female	10	Georgia

The "Applied Experience" column specifically refers to a combination of the participants' experience with using applied behavior analysis and teaching skills to individuals with autism

child with ASD—and the acting BCBA that supervised the child, to videotape the assessment of specific skills (i.e., *ABLLS-R* items) during an *ABLLS-R* assessment. The research team selected a combination of 86 items in which the child either scored a zero (i.e., did not meet the lowest scoring criteria), obtained some, but not all possible points, and those that the child scored the maximum number of points allotted for the given item.

After selecting the 86 items to include in the survey, we filmed the BCBA assessing the true skills of the child with ASD on each of these skills within a clinical setting. We then added labels to each clip to denote the number in the survey that the video clip corresponded to as well as the *ABLLS-R* item being assessed (e.g., "Question 1, B8").

Procedure

After developing all of the relevant materials for the study, the first author uploaded materials for both surveys to the Survey Monkey website and recruited our participant sample. The final participant sample included individuals who responded to the email,

Table 2 Demographics and Related Information Pertaining to the Members of the Second Expert Panel

	Degree	Occupation	Years in Occupation	Race	Age	Gender	Geographic Location
Participant 1	M.A.	Elementary School Teacher	8	White	32	Female	Georgia
Participant 2	M.S.	Speech and Language Pathologist	9	White	44	Female	Georgia
Participant 3	M.A.	Elementary School Teacher	8	White	36	Male	Georgia
Participant 4	M.S.	Occupational Therapist	17	White	46	Female	Georgia
Participant 5	Ed.S.	School Psychologist	5	White	39	Female	Georgia

provided their informed consent, and possessed the aforementioned characteristics required for inclusion onto one of the two expert panels. After confirming our final participant sample, the first author emailed each participant with directions for the study and a link to their online survey.

The online survey administered to the first expert panel consisted of each *ABLLS-R* item and its related scoring criteria. The researchers used the technique described by Lawshe (1975) and asked individuals in this expert panel to rate each of the 544 *ABLLS-R* items on a 3-point Likert scale (i.e., 3 = Essential, 2 = Useful, but not essential, or 1 = Not necessary) on the extent that they perceived the skill measured by each *ABLLS-R* item as "essential."

The second expert panel received the scoring instructions from the *ABLLS-R Scoring Instructions and IEP Development Guide* prior to completing their online survey. We asked the panel members to read the material until they achieved a full understanding of how to score the *ABLLS-R*. Following this preliminary task, members of the second expert panel received video clips via email and a link to their online survey. The survey administered to the second expert panel required the experts to watch each of the 86 video clips and use the *ABLLS-R* scoring criteria provided in the online survey to score the actual performance of the student.

Members of both expert panels received three weeks to complete their respective surveys. At the end of the first week, the first author emailed all expert panel members and reminded them to complete their survey within the next two weeks. All participants responded to every question in their panel specific survey and they completed it within the allotted three-week timeframe (i.e., no missing data were observed).

Data Analysis

Validity Researchers commonly reference and employ the techniques described by Lawshe (1975) to evaluate the content validity of a measure. His methods include the calculation of the content validity ratio (CVR) for each item to measure the extent that members of the expert panel considered the skill measured by the particular item as "essential." In his article, Lawshe provided the following formula to calculate the CVR:

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

In this formula, n_e represents the number of judges that rated the item as "essential" and N equals the number of judges in our expert panel. The expert panel from the present study contained six judges which means that a CVR of 1.00 corresponds to an item rated as "essential" by all six judges from our panel, a CVR of .67 corresponds to an item rated as "essential" by five judges, a CVR of .33 corresponds to an item rated as "essential" by four judges, a CVR of zero corresponds to an item rated as "essential" by three judges, a CVR of −.33 corresponds to an item rated as "essential" by two judges, and a CVR of −.67 corresponds to an item rated as "essential" by only one judge.

Lawshe (1975) and Lynn (1986) provided some recommendations to help guide the process of interpreting CVRs values. Lawshe noted that items contain evidence of content validity if they receive the rating of "essential" by more than half of the expert panel members and that a higher CVR value reflects stronger evidence of content validity. Per his guidelines, at least four out of the six expert judges (i.e., a CVR of .33) would need to rate an *ABLLS-R* item as "essential" for it to contain evidence of content validity. Lynn provided a more stringent cutoff for establishing evidence of content validity and suggested that five out of six, or 83% of the expert panel members, needed to agree on the relevancy of the item. This percentage of agreement (i.e., 83%), when applied to the mathematical equation provided by Lawshe, corresponds to a CVR value of .67. In order to achieve a higher level of confidence in our findings, we elected to use the more stringent criterion set forth by Lynn and considered *ABLLS-R* items with a CVR of at least .67 (i.e., 83% of the expert judges considered the item as "essential") as containing evidence of content validity.

Reliability The present study also examined the inter-rater reliability of the *ABLLS-R* scores obtained from our second panel of experts. We obtained an intra-class correlation coefficient (ICC) to measure the reliability of their ratings. Prior to calculating the ICC, we established some parameters that influenced how we conducted our statistical analysis.

The literature cites three different models of ICCs based on the purpose and design of the study and the type of measurements obtained (Shrout and Fleiss 1979). We recruited our participant sample from a larger population and each expert panel member rated the same number of *ABLLS-R* items (i.e., the second model). In addition, we determined a priori that we would use and interpret the average ICC rather than the single measures ICC since the purpose of our study included an examination of the inter-rater reliability of a set of scores and not that of a single rater. Taken together, our research team measured the inter-rater reliability of *ABLLS-R* scores obtained using ICC (2,*k*) whereby the number two specifies the model used and the letter *k* represents the average reliability of the ratings (in this case, scores) obtained from the individuals that comprised our expert panel (Landers 2011). We analyzed our data in *SPSS 22.0* using a two-way random effects model and interpreted our results using the frequently cited, general guidelines set forth for by Cicchetti (1994) who characterized the inter-rater reliability of ICCs under .40 as "poor," between .40 and .59 as "fair," between .60 and .74 as "good," and between .75 and 1.00 as "excellent."

Results

Validity

We used the ratings obtained from members of our first expert panel to calculate a CVR value for each of the 544 items from the *ABLLS-R* (see Table 3). We found that 441 *ABLLS-R* items (i.e., 81% of the items) either met or exceeded our CVR cutoff of .67 for containing evidence of content validity. Specifically, 304 *ABLLS-R* items contained a CVR value of 1.00 and 137 items contained a CVR value of .67. The remaining 103 *ABLLS-R* items contained a CVR value of less than .67—a finding that reflects insufficient evidence of content validity.

Table 3 Distribution of the CVR Values Obtained

CVR	# of *ABLLS-R* Items	Specific *ABLLS-R* Items
1.00	304	A6, A7, A15, B3-B6, B8, B13, B16-B18, B20, B25, B26, C7-C17, C20-C24, C27-C31, C32, C34-C40, C42, C43, C45-C49, C52, C55, C56, D1, D3, D4, D5, D7, E18, F2-F6, F8, F9, F11-F13, F15-F22, F24-F29, G1-G8, G11-G29, G31, G33, G34, G37, G39-G41, H4, H5, H7-H25, H27-H33, H35-H40, H43-H45, H49, I7-I9, J1, J2, J4-J14, J17, K8, K9, K13, K14, L1-L4, L8-L10, L12, L13, L19-L22, L24, L25, L32, L33, M1-M12, N1-N10, P1-P6, Q1-Q3, Q5, Q7, Q10, Q12-Q17, R1-R3, R5, R8-R10, R12, R13, R15, R16, R21-R26, R29, S1, S3, S4, S7-S10, T3, T4, T6, T7, U1-U15, V1-V4, V7-V10, W1, W2, X1-X4, X7, X9, X10, Z1-Z4, Z8, Z12, Z13, Z16, Z28
.67	137	A1, A3, A5, A8, A13, A19, B1, B2, B12, B19, B21, C1, C2, C26, C33, C41, C44, C50, C51, C53, C54, C57, D2, D6, D9, D10, D12, D15-D20, D23-D27, E1-E6, E8-E14, E16, E17, F1, F14, F23, G9, G10, G30, G35, G36, G38, G42, G46, G47, H1, H2, H6, H42, H46-H48, I4, J3, J15, J16, J19, K1-K4, K6, K10-K12, L5, L6, L14-L16, L18, L23, L26, L29-L31, L34, Q4, Q6, Q11, R6, R7, R11, R14, R17, R18, R20, R27, R28, S5, S6, T2, T5, V5, V6, W3-W7, X5, X8, Y2, Y6, Z5-Z7, Z11, Z15, Z17, Z19, Z20, Z23-Z27
.33	61	A2, A4, A9, A11, A12, B7, B9, B11, B14, B15, B22, B23, B27, C3-C6, C25, D8, D11, D13, D21, D22, E7, E15, F10, G32, G44, G45, H34, H41, I1-I3, I6, J18, J20, K5, K7, L7, L11, L17, Q8, Q9, R4, R19, S2, T1, X6, Y1, Y3, Y5, Y8, Y13, Y16, Y30, Z9, Z10, Z14, Z18, Z21
0	31	A10, A16-A18, B10, C18, C19, D14, E19, E20, F7, G43, H3, H26, I5, K15, L27, Y4, Y7, Y10, Y11, Y15, Y19, Y20, Y23, Y24, Y26-Y29, Z22
−.33	9	L28, Y9, Y12, Y14, Y17, Y18, Y21, Y22, Y25
−.67	2	A14, B24

Reliability

Participants in our second expert panel scored the performance of a student with ASD on 86 *ABLLS-R* items (see Table 4). We then calculated the average ICC to reflect the inter-rater reliability of their scores. Results from our statistical analysis yielded evidence of excellent inter-rater reliability across the scores obtained by members of our second expert panel with the average ICC (2, 5) = .95 (95% CI: .94–.97; $p < .001$).

Discussion

The current study used data obtained from two separate panels of expert raters to examine the content validity of the *ABLLS-R* and the inter-rater reliability of its scores. Our results reflect evidence of content validity as the majority of our first expert panel rated the skills measured by 441 *ABLLS-R* items (or 81% of the items from the assessment) as "essential." In addition, we found that all experts rated the majority of the *ABLLS-R* items as "essential." The high ratings obtained bear added significance given that the members of our expert panel have worked closely with individuals with ASD for multiple years and are familiar with their common areas of skill deficit.

Table 4 Distribution of the 86 *ABLLS-R* Items Scored by the Second Expert Panel

Repertoire label	*ABLLS-R* items
A Cooperation and Reinforcer Effectiveness	No Items
B Visual Performance	B8, B9, B12, B13, B16-B19
C Receptive Language	C19, C20, C23-C25, C33-C35
D Motor Imitation	D4-D6, D8-D16
E Vocal Imitation	E5, E6, E11, E13-E15
F Requesting	No Items
G Labeling	G5, G7, G8, G12
H Intraverbals	H4-H10
I Spontaneous Vocalizations	No Items
J Syntax and Grammar	J4
K Play and Leisure	No Items
L Social Interaction	No Items
M Group Instruction	No Items
N Classroom Routines	No Items
P Generalized Responding	No Items
Q Reading	Q3-Q9
R Math	R1-R8
S Writing	S3, S4
T Spelling	No Items
U Dressing	No Items
V Eating	No Items
W Grooming	No Items
X Toileting	No Items
Y Gross Motor	Y4, Y6-Y10, Y12, Y13, Y15, Y17-Y19
Z Fine Motor	Z7, Z12, Z14, Z16–18, Z26

We also examined the inter-rater reliability of the *ABLLS-R* scores obtained from a second expert panel comprised of individuals that worked and interacted with children with ASD, but had never previously used the *ABLLS-R*. Since individuals with these characteristics could potentially administer the *ABLLS-R*, we sought to examine the consistency of scoring by individuals with the least amount of training and experience required (i.e., none) to carry out an *ABLLS-R* assessment. Results obtained from the present study yield evidence of excellent inter-rater reliability across the scores obtained from our expert panel. Specifically, the high average ICC obtained demonstrates that the *ABLLS-R* can yield consistent scores, even for those that had no prior experience with using or administering the *ABLLS-R*.

Findings obtained from the present study add to the growing body of literature on the psychometric properties of the *ABLLS-R*. The results obtained from the current study extend the research by Malkin et al. (2016) by documenting evidence of a second form of validity (i.e., content validity). In addition, we also found evidence of excellent inter-rater reliability across the *ABLLS-R* scores—a finding that extends the research by

Partington et al. (2016) who obtained evidence of both internal consistency and test-retest reliability. Collectively, the evidence obtained from the current study and past research represents a major step toward establishing the *ABLLS-R* as a valid instrument that yields reliable scores—an important component that addresses the requirements of both state and federal law.

While our results further demonstrate the strength of the psychometric properties of the *ABLLS-R*, some limitations exist which warrant further discussion. Despite using the guidelines set forth by Salkind (2013), our sample size of six expert panel members likely represents one such limitation. The guidelines for establishing evidence of content validity provided by Lynn (1986) allowed us to conclude that 81% of the *ABLLS-R* items contain evidence of content validity. However, establishing more concrete evidence of content validity requires both a larger number of expert panel members and an evaluation of their data using a table of critical values (Wilson et al. 2012) to identify which *ABLLS-R* items to retain and delete from the assessment based on their associated CVR values (i.e., which reflect quantitative evidence of content validity). While we recognize that we could conceivably use the critical values to interpret the CVR values obtained from the present study and make recommendations as to which items to retain and delete, doing so would exceed the scope of the present study. Further, the small size of our first expert panel and our use of a more stringent CVR cutoff to establish evidence of content validity (i.e., which would influence which items we would recommend for retention or deletion) would collectively, hinder the accuracy of our results and any subsequent conclusions drawn from our findings. Taken together, we consider findings from the present study as preliminary evidence of content validity that requires future research, which accounts for the aforementioned limitations, to lend further confidence to our findings and to establish more concrete evidence of content validity.

A second limitation includes the distribution of the 86 *ABLLS-R* items used to assess the inter-rater reliability of the scores obtained from our second expert panel. Specifically, we selected 86 items from three out of the four skill sets (i.e., the basic learner skills, academic skills, and motor skills) cited in the *ABLLS-R Scoring Instructions and IEP Development Guide*. Although we obtained very high inter-rater reliability, we did not include items from the self-help skill set (i.e., repertoires U, V, W, and X).

Another related set of limitations involves our criteria for inclusion onto our expert panels and our method of selecting our expert panel members. In the present study, we consulted with our professional colleagues as a means to locate and identify prospective participants. Unfortunately, the inclusion criteria that we used allowed for a wide range of individuals (e.g., those who trained or practiced in different programs or regions, those who worked in different, but related fields, etc.) to qualify for our study as expert panel members—a factor that likely either led to or promoted some bias in the process of identifying and selecting participants for our sample. Further, qualified individuals with specific characteristics (i.e., those noted above) may display variability in their patterns of responding relative to other qualified individuals with different characteristics. Collectively, our sampling technique employed and the consequences of using it (i.e., the extent of the overall representativeness of our participant sample) could well have influenced our findings.

The aforementioned limitations, in addition to some prevailing gaps within this line of research, can guide researchers seeking to further examine the psychometric properties of the *ABLLS-R*, including both the validity of the assessment and the reliability of its scores. Indeed, researchers may seek to take the next step toward establishing more concrete evidence of content validity by using a larger number of expert panel members, recommending specific *ABLLS-R* items for retention or deletion based on their CVR value, and then calculating the content validity index for the entire scale (i.e., using the items retained). A second direction for future research may include the use of items from all four *ABLLS-R* skill sets to yield a more accurate measure of the inter-rater reliability of the *ABLLS-R* scores (i.e., since the current study did not include items from the self-help skill set). Those seeking to pursue this line of research might also consider counterbalancing the order in which the *ABLLS-R* items are presented—a tactic that was not employed in the present study. Lastly, one might also assess the inter-rater reliability of *ABLLS-R* scores obtained from both individuals with extensive training on the administration of the *ABLLS-R* (e.g., professionals) and those with no previous training (e.g., parents of a child with ASD). Examining the *ABLLS-R* scores obtained from individuals that possess varying degrees of experience with using the *ABLLS-R* will confirm whether those with minimal to no training, can also administer the *ABLLS-R* and achieve consistent scoring in relation to trained professionals.

Funding This study did not receive any funding.

Compliance with Ethical Standards

Ethical Approval All procedures performed in studies involving human participants 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.

Informed Consent Informed consent was obtained from all individual participants included in the study.

Conflict of Interest The first and second authors declare that they have no conflict of interest. The third author owns stock in the company that publishes the *ABLLS-R*.

References

Aman, M.G., Novotny, S., Samango-Sprouse, C., Lecavalier, L., Leonard, E., Gadow, K.D…Chez, M. (2004). Outcome measures for clinical drug trials in autism. *CNS Spectrums, 9*, 36–47. https://doi.org/10.1017/S1092852900008348.

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

American Medical Association. (2014). *CPT assistant*. Chicago: American Medical Association.

Center for Disease Control and Prevention (2015). *Data and statistics*. Retrieved from http://www.cdc.gov/NCBDDD/autism/data.html

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*, 284–290. https://doi.org/10.1037/1040-3590.6.4.284.

Every Student Succeeds Act, Pub. L. No. 114–95 § 114 Stat. 1177 (2015).

Guldberg, K. (2010). Educating children on the autism spectrum: Preconditions for inclusion and notions of "best autism practice" in the early years. *British Journal of Special Education, 37*, 168–170. https://doi.org/10.1111/j.1467-8578.2010.00482.x.

Individuals with Disabilities Education Act, 20 U.S.C. § 1400 (2004).

Landers, R. N. (2011). *Computing intraclass correlations (ICC) as estimates of inter-rater reliability in SPSS*. Retrieved from http://neoacademic.com/2011/11/16/computing-intraclass-correlations-icc-as-estimates-of-inter-rater-reliability-in-spss/#.U6z9uChhmGl. doi:10.15200/winn.143518.81744.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*, 563–575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x.

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research, 35*, 382–385. https://doi.org/10.1097/00006199-198611000-00017.

Malkin, A., Dixon, M. R., Speelman, R. C., & Luke, N. (2016). Evaluating the relationships between the *PEAK Relations Training System-Direct Training Module, Assessment of Basic Language and Learning Skills-Revised*, and the *Vineland Adaptive Behavior Scales-II. Journal of Developmental and Physical Disabilities*. https://doi.org/10.1007/s10882-016-9527-8.

Partington, J. W. (2010a). *The Assessment of Basic Language and Learning Skills-Revised*. Pleasant Hill: Behavior Analysts, Inc..

Partington, J. W. (2010b). *The Assessment of Basic Language and Learning Skills-Revised: Scoring Instructions and IEP Development Guide*. Pleasant Hill: Behavior Analysts, Inc..

Partington, J. W., Bailey, A., & Partington, S. W. (2016). A pilot study examining the test-retest and internal consistency reliability of the *ABLLS-R*. *Journal of Psychoeducational Assessment*. https://doi.org/10.1177/0734282916678348.

Salkind, N. (2013). *Encyclopedia of research design: Content validity*. Thousand Oaks: SAGE Publications, Inc..

Schwartz, I. S., Boulware, G., McBride, B. J., & Sandall, S. R. (2001). Functional assessment strategies for young children with autism. *Focus on Autism and Other Developmental Disabilities, 16*, 222–227. https://doi.org/10.1177/108835760101600404.

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.

Sparrow, S. S., Cicchetti, D. V., & Balla, D. A. (2005). *Vineland Adaptive Behavior Scales* (2nd ed.). Bloomington, MN: Pearson. https://doi.org/10.1037/t15164-000.

Sundberg, M. (2008). *VB-MAPP Verbal Behavior Milestones Assessment and Placement Program: A language and social skills assessment program for children with autism or other developmental disabilities: Guide*. Concord: AVB Press.

Thompson, T. (2007). *Making Sense of Autism*. Baltimore: Paul H. Brookes Publishing Co..

Thompson, T. (2011). *Individualized Autism Intervention for Young Children*. Baltimore: Paul H. Brookes Publishing Co..

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*, 197–210. https://doi.org/10.1177/0748175612440286.

Wolak, M. E., Fairbairn, D. J., & Paulsen, Y. R. (2012). Guidelines for estimating repeatability. *Methods in Ecology and Evolution, 3*, 129–137.