Research in Developmental Disabilities 34 (2013) 1804–1814

Contents lists available at SciVerse ScienceDirect

Research in Developmental Disabilities

Validity and reliability of the *Behavior Problems Inventory*, the *Aberrant Behavior Checklist*, and the *Repetitive Behavior Scale – Revised* among infants and toddlers at risk for intellectual or developmental disabilities: A multi-method assessment approach☆

Johannes Rojahn [a,*], Stephen R. Schroeder [b], Liliana Mayo-Ortega [b,c], Rosao Oyama-Ganiko [c], Judith LeBlanc [b,c], Janet Marquis [b], Elizabeth Berke [a]

[a] George Mason University, United States
[b] University of Kansas, United States
[c] Centro Ann Sullivan del Peru, Peru

ARTICLE INFO

Article history:
Received 17 December 2012
Received in revised form 18 February 2013
Accepted 22 February 2013
Available online 16 March 2013

Keywords:
Aberrant Behavior Checklist
Behavior Problems Inventory
Repetitive Behavior Scale – Revised
Psychometric properties
Reliability, Validity
Multitrait–multimethod
Neurodevelopmental disorders
At risk
Infants
Toddlers
Children

ABSTRACT

Reliable and valid assessment of aberrant behaviors is essential in empirically verifying prevention and intervention for individuals with intellectual or developmental disabilities (IDD). Few instruments exist which assess behavior problems in infants. The current longitudinal study examined the performance of three behavior-rating scales for individuals with IDD that have been proven psychometrically sound in older populations: the *Aberrant Behavior Checklist (ABC)*, the *Behavior Problems Inventory (BPI-01)*, and the *Repetitive Behavior Scale – Revised (RBS-R)*. Data were analyzed for 180 between six and 36 months old children at risk for IDD. Internal consistency (Cronbach's α) across the subscales of the three instruments was variable. Test–retest reliability of the three BPI-01 subscales ranged from .68 to .77 for frequency ratings and from .65 to .80 for severity ratings (intraclass correlation coefficients). Using a multitrait–multimethod matrix approach high levels of convergent and discriminant validity across the three instruments was found. As anticipated, there was considerable overlap in the information produced by the three instruments; however, each behavior-rating instrument also contributed unique information. Our findings support using all three scales in conjunction if possible.

Individuals with intellectual or developmental disabilities (IDD) are at a heightened risk for developing chronic and severe behavior problems during the course of their lives. Reliable and valid assessment of such behaviors is an important element in empirically verifying successful prevention and intervention.

In the past decade, there has been an increased interest in early identification and preventive intervention of behavior problems, including aggression, self-injurious behavior (SIB), and stereotyped behavior, among infants and toddlers at risk for IDD (Sigafoos, Lancioni, Didden, & O'Reilly, in press; Schroeder & Courtemanche, 2012). Yet, there are only a few existing

☆ This study is part of an ongoing international collaborative research project between the University of Kansas, Centro Ann Sullivan del Peru, and George Mason University which was funded by the Fogarty International Research grant no. HD 060500. The authors wish to gratefully acknowledge the effort of the participating families and the pro bono work of many professions in Peru.

* Corresponding author at: Department of Psychology, George Mason University, 10340 Democracy Lane, Suite 202, Fairfax, VA 22033, United States. Tel.: +1 703 993 4241.
E-mail address: jrojahn@gmu.edu (J. Rojahn).

0891-4222/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ridd.2013.02.024

assessment instruments that were specifically developed and validated for behavior problems in children below the age of three, such as the recently published *Baby and Infant Screen for Children with Autism Traits, Part 3* (BISCUIT-3) (Matson et al., 2009). There are, however, several well-validated behavior-rating scales that were originally developed for and validated in adult populations with IDD which have also been used successfully among younger populations.

The current study is a psychometric analysis involving three behavior-rating scales with a population of infants and toddlers in Peru. In this study, we analyzed several psychometric areas of the BPI-01 using these challenging behavior instruments. Firstly, we examined the test–test reliability of the BPI-01. Secondly, using a multitrait–multimethod approach (Campbell & Fiske, 1959), we examined the confirmatory and discriminant validity of the BPI-01, in comparison to the ABC and the RBS-R. Campbell and Fiske (1959) described the multitrait–multimethod approach as a tool to examine convergent and discriminant validity in order to establish overall validity of an empirical instrument. This approach involves presenting a matrix that includes all possible intercorrelations when measuring specific traits through several methods. Lastly, we examined the sensitivity and specificity of the BPI-01 in comparison to the RBS-R and the ABC.

1. Method

1.1. Participants and recruitment

Participants were recruited by the Centro Ann Sullivan del Peru (CASP) via radio, television, and newspaper announcements in Lima, Peru and the rest of the country for a family-based early intervention program. The advertisements solicited children between 6 and 36 months old showing signs of aberrant behaviors. Around 1000 families called CASP in response to the advertisements. Parents where then connected with a trained CASP Triage and Information Coordinator who further explained the study and inclusion criteria. At this point, parents decided whether their child was appropriate for the study and 341 families agreed to visit CASP with their child for a 15–30 min screening interview.

Children were chosen for clinical trials based on their signs of SIB, aggression, and stereotyped behavior and their parents' answers on the *Parental Concerns Questionnaire* (PCQ). The PCQ is a 15-item yes/no questionnaire based upon risk factors of aberrant behaviors that were developed by our study (Mayo-Ortega et al., 2012). Trained veteran parents administered the PCQ. Each of these parents had had a child enrolled at CASP and had received around 185 h of training per year for several years. We feel that this is particular strength of the screening process because the veteran parents could empathize with the families and encourage them. Participants had to meet several inclusion criteria gleaned from the research literature (Dawson, 1996; Dunlap et al., 2006; Rojahn, Schroeder, & Hoch, 2008) that would classify them as at-risk for behavior problems in IDD (Mayo-Ortega et al., 2012). The inclusion criteria involved genetic disorders associated with IDD (i.e., Down, Smith-Magenis, Prader-Willi, Rett, and Fragile-X syndrome), family history of brain disorders, common comorbid medical conditions associated with IDD (i.e., congenital rubella, tuberous sclerosis, brain trauma, and stroke), pre-and perinatal disorders causing serve or profound ID, psychiatric factors (i.e., family history of autism, mood disorders, compulsive disorders, anxiety disorders, and BPD), neurochemical metabolic factors (i.e., Hyperserotonemia, dopamine depletion in the nigrostriatal pathways of the basal ganglia, dysfunctional opioid peptide system, elevated ammonia or lactic acid levels, organic aciduria, fatty acid disorder, aminoacidopathy, and mitochondrial disorder), neuropsychological factors (i.e., lack of orientation to social stimuli, joint attention, motor coordination, position, and/or motor imitation, and presence of stereotyped behavior), and communication deficits in either receptive or expressive communication (Mayo-Ortega et al., 2012).

Of these 341 families that responded and visited the CASP facility for a screening interview, 262 met the inclusion criteria, and 234 were administered the first round of in-depth interdisciplinary assessment measures. This first round involved developmental pediatric exams with consultations in neurology, genetics, and nutrition, questions about vision, hearing and dental health, cognitive (Bayley, 2006) and communication (Wetherby & Prizant, 2002) assessments, and an autism screening with the *Child Autism Rating Scale* (CARS; Schopler, Reichler, & Renner, 1988). The CARS was part of the interdisciplinary evaluation after screening and their scores did not affect inclusion in the study. Of the 180 participants who completed the study, 74 had CARS scores of 15–53 ($M = 35$). These children were then reassessed six months later and one year later.

In the end, 180 participants completed the BPI-01 (Rojahn, Matson, Lott, Esbensen, & Smalls, 2001) for all three time points and were included in this study. Participants were excluded because they either failed to meet inclusion criteria or did not return to CASP to complete subsequent rounds of data collection. These participants consisted of 110 boys and 68 girls (2 missing data), ranging in age from 4 to 48 months ($M = 27.4$; $SD = 10.1$).

None of the children in this study had a formal diagnosis of an Autism Spectrum Disorder (ASD) but they were all identified as at risk for ASD. Of the 180 to complete the study, 21 were un-testable on the Bayley (2006) due to severity of their disability or noncompliance. A total of 100 children were more than one standard deviation below the mean and 59 had scores that were below average but within one SD below the mean. We were hesitant to label any of these children as having an IDD because they were so young. Instead they were classified as at risk for IDD.

1.2. Measures of behavior problems

The key behavioral assessments utilized in this study were the ABC, the BPI-01, and the RBS-R.

1.2.1. Aberrant Behavior Checklist (ABC)

The ABC (Aman & Singh, 1986; Aman, Singh, Stewart, & Field, 1985a, 1985b) is a 58-item informant-based behavior rating scale that measures the severity of five subscales: *Irritability* (15 items), *Lethargy* (16 items), *Stereotypy* (7 items), *Hyperactivity* (16 items), and *Excessive Speech* (4 items). Karabekiroglu and Aman (2009) found moderate to high internal consistency for each subscale (α = .90, .81, .83, .89, and .68, respectively). Each item is scored on a 4-point Likert scale for severity (0 = not a problem at all, 1 = the behavior is a problem but slight in degree, 2 = the problem is moderately serious, and 3 = the problem is severe in degree). As the subscale labels indicate, this is a broad-band assessment instrument that captures a wide variety of behavior problems and includes generic items of SIB and aggressive behavior. It is one of the most widely used assessment instruments in the IDD literature and has been repeatedly shown to have excellent psychometric properties (e.g., Aman et al., 1985a; Bihm & Poindexter, 1991; Paclawskyi, Matson, Bamburg, & Baglio, 1997). The ABC was used as a dependent variable in this study to capture a broad range of behavior problems.

1.2.2. Behavior Problems Inventory (BPI-01)

The BPI-01 (Rojahn et al., 2001, 2012a, 2012b) is a 52-item informant-based behavior rating scale that assesses the frequency and severity of problem behaviors in individuals with IDD. The BPI-01 was the main dependent variable in this study. The items on the BPI-01 fall into three subscales: *Self-Injurious Behavior* (14 items), *Stereotyped Behavior* (25 items), and *Aggressive/Destructive Behavior* (11 items). Each item is scored on a three-point Likert scale for severity (1 = slight, 2 = moderate, and 3 = severe) and a five-point Likert scale for frequency (0 = never, 1 = monthly, 2 = weekly, 3 = daily, and 4 = hourly). Several previous studies have examined the psychometric properties of the BPI-01. Recently, Rojahn et al. (2012a, 2012b) conducted a comprehensive psychometric analysis on a data set of 1122 individuals collated from several different sites with a BPI-01 total score > 0. Internal consistency (Cronbach's α) of the *Self-Injurious Behavior* subscale was .74 and .73 (for the frequency and severity scales respectively), .89 and .90 for the *Stereotyped Behavior* subscale, and .92 and .90 for the *Aggressive/Destructive Behavior* subscale. Rojahn et al. (2001) reported intraclass correlation coefficient (ICC) test–retest reliabilities of .71, .76, and .64 respectively. Evidence for strong convergent and discriminant validity was also found, and the factor structure was endorsed by a confirmatory factor analysis (Rojahn et al., 2012b).

1.2.3. Repetitive Behavior Scale – Revised (RBS-R)

The RBS-R (Bodfish, Symons, Parker, & Lewis, 2000; Boyd, McDonough, & Bodfish, 2012) is a 43-item informant-based behavior rating scale that measures the severity of a variety of behaviors among individuals with IDD. It contains six subscales: *Stereotyped Behavior* (6 items), *Self-Injurious Behavior* (8 items), *Compulsive Behavior* (8 items), *Ritualistic Behavior* (6 items), *Sameness Behavior* (11 items), and *Restricted Behavior* (4 items). These subscales are each scored on a four point Likert scale (0 = behavior does not occur, 1 = behavior occurs and is a mild problem, 2 = behavior occurs and is a moderate problem, 3 = behavior occurs and is a severe problem). Using data from 320 caregivers, Lam and Aman (2007) validated a five-factor solution for the RBS-R and they also conducted an independent validation study of the RBS-R and found high internal consistency for each subscale (ranging from .78 to .91) and inter-rater reliability using ICCs ranging from .57 to.73.

The SIB items of the RBS-R and the BPI-01 are contrasted in Table 1. The two instruments each have three items that are more or less equivalent (i.e., biting, scratching, and hair pulling), three RBS-R "hitting" items correspond with two BPI-01 "hitting" items, and three BPI-01 "inserting" items correspond with one RBS-R "inserting" item. The BPI-01 has seven SIB items that are not represented in the RBS-R, while the RBS-R has one item ("skin picking") that is not represented in the BPI-01. The stereotyped behavior items of the ABC, the RBS-R and the BPI-01 are compared in Table 2.

Table 1
SIB Items (abbreviated).

RBS-R	BPI-01
Bites self	Self-biting
Hits with body part	Hitting head with hand or body part
Hits against surface or object	Hitting body (not head)
Hits with object	
Rubs or scratches	Self-scratching
Inserts finger or object	Inserting objects
Inserting fingers/objects	Inserting fingers
	Inserting objects
Pulls hair or skin	Hair pulling
	Self-pinching
	Pica
	Pulling nails
	Air swallowing
	Extreme drinking
	Teeth grinding
	Vomiting and rumination
Skin picking	

Table 2
Stereotyped Behavior Items (abbreviated).

RBS-R	BPI-01	ABC
Body rocking	Rocking back and forth	Body movements
	Repetitive body movements	Rocks body back and forth
Rolls, nods, turns head	Rolling head	Moves or rolls head
Locomotion (turns, jumps)	Bursts of running	
	Pacing	
	Bouncing around	
	Spinning own body	
	Whirling, turning around	
Sensory (gazes, sniffs, ...)	Sniffing objects	
	Gazing at hands or objects	
	Sniffing own body	
Hand/finger	Complex hand/finger movements	
	Sustained finger movements	
	Repetitive hand movements	
	Clapping hands	
Object use (spins, twirls, ...)	Twirling things	
	Waving or shaking arms	
	Spinning objects	
	Manipulating objects repeatedly	
	Waving hands	Waves/shakes extremities
	Yelling/screaming	
	Rubbing self	
	Maintaining bizarre postures	
	Grimacing	
		Stereotyped behavior
		Odd, bizarre behavior
		Repetitive movements

1.3. Procedure

The BPI-01 was administered to all 180 children at all three time points, while the ABC and the RBS-R were completed for 97 participants at the second time only. The time between measurements was approximately six months. To avoid literacy and comprehension issues, ten well-trained interviewers administered all assessment instruments by questioning the parents or caregivers and filling out all of the forms. An effort was made to keep the same interviewers with the same parents over all three data collection periods. For quality insurance, the interviewers were periodically monitored.

The intervention, which is not the focus of the present paper, consisted of six 3-h bi-monthly teaching workshops at CASP, follow-up procedures with bi-monthly parent teaching workshops given at CASP, and monthly telephone follow-ups with each family. This treatment lasted throughout the 12-month study period and its effects will be published elsewhere (Oyama, Mayo, Schroeder, & LeBlanc, in preparation).

2. Results

2.1. Descriptive statistics

Table 3 presents descriptive statistics of the three instruments. The values for the BPI-01 were comparable to "clinical norms" of the BPI-01 for children from birth to ten years old published by Rojahn et al. (2012a), which were as follows: *Self-Injurious Behavior* frequency, $M = 5.84$ ($SD = 5.29$) and severity $M = 4.39$ ($SD = 4.01$); *Stereotyped Behavior* frequency $M = 18.39$ (SD = 17.42) and severity $M = 10.5$ ($SD = 11.2$); *Aggressive/Destructive Behavior* frequency $M = 6.81$ ($SD = 7.9$) and severity $M = 5.38$ ($SD = 1.15$). This suggests that the behavior problems exhibited by this sample of children were consistent with other groups of peers with IDD and behavioral concerns.

2.2. Internal consistency

A summary of the internal consistency statistics (Cronbach's α) can be found in Table 4, and interpretations were derived from recommendations by Cicchetti and Sparrow (1981).

Internal consistency of the BPI-01 total score across the three times of measurement was excellent, ranging from $\alpha = .85$ to .90. For the subscale scores, internal consistency ranged from poor to acceptable for the SIB subscale ($\alpha = .42$–.69); excellent for the stereotyped behavior subscale ($\alpha = .82$–.86); and good to excellent for the aggressive/destructive behavior subscale

Table 3
Descriptive statistics for the BPI-01, the ABC, and the RBS-R.

	Scales	Time 1			Time 2			Time 3		
		M	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*
BPI-01										
Self-Injurious Behavior	f[a]	6.8	6.1	180	5.8	4.9	180	5.1	4.1	180
	s[b]	4.7	4.5	180	3.8	3.5	180	3.2	2.7	180
Stereotyped Behavior	f	13.7	13.1	180	12.5	11.5	180	10.8	11.8	180
	s	9.4	10.2	180	7.3	7.6	180	6.1	6.8	180
Aggressive/Destructive	f	7.7	7.4	180	7.2	6.6	180	7.0	7.1	180
	s	6.1	6.5	180	5.5	5.6	180	5.2	5.4	180
Total score	f	28.2	22.0	180	25.5	17.7	180	23	17.3	180
	s	20.1	17.9	180	16.6	13.8	180	14.5	11.6	180
ABC										
Irritability	s				12.6	9.4	97			
Lethargy	s				8.4	8.7	97			
Stereotypy	s				3.6	4.6	97			
Hyperactivity	s				18.1	10.9	97			
Excessive Speech	s				1.9	2.8	97			
Total score	s				44.7	30.3	97			
RBS-R										
Stereotypy	s				2.8	3.1	97			
Self-Injurious Behavior	s				2.0	2.5	97			
RBS	s				3.5	4.0	97			
Rituals	s				3.3	3.5	97			
RBS	s				6.0	5.0	97			
Restricted	s				2.0	2.1	97			
Total score	s				19.5	15.1	97			

[a] Frequency ratings.
[b] Severity ratings.

(α = .76–.81). The ABC total score (α = .96) and the subscales (α = .81–.92) both, had excellent internal consistency. The RSB-R total score had excellent internal consistency (α = .89), while its subscales ranged from questionable to good (α = .55–.75).

2.3. Test–retest reliability

Test–retest reliability for the BPI-01 was estimated in two ways. First, Spearman's ρ correlations were computed comparing rounds 1 and 2, 1 and 3, and 2 and 3. All correlations were statistically significant at the .01 level, ranging from

Table 4
Internal consistency (Cronbach's α).

Instruments and subscales	Number of items	Scales	Time 1	Time 2	Time 3
BPI-01					
Self-Injurious Behavior	14	f[a]	.69	.55	.42
		s[b]	.69	.58	.45
Stereotyped Behavior	24	f	.86	.82	.86
		s	.88	.84	.86
Aggressive/Destructive	11	f	.81	.76	.80
		s	.84	.80	.80
Total score	49	f	.90	.85	.86
		s	.92	.88	.87
ABC					
Irritability	15	s		.90	
Lethargy	16	s		.91	
Stereotypy	7	s		.87	
Hyperactivity	16	s		.92	
Excessive Speech	4	s		.81	
Total score	58	s		.96	
RBS-R					
Stereotyped Behavior	6	s		.75	
Self-Injurious Behavior	8	s		.55	
Compulsive Behavior	8	s		.72	
Ritualistic Behavior	6	s		.68	
Sameness Behavior	11	s		.74	
Restricted Behavior	4	s		.30	
Total score	43	s		.89	

[a] Frequency ratings.
[b] Severity ratings.
Clinical significance of reliability levels: <.40 = poor agreement; .40–.59 = fair; .60–.74 = good; .75 to 1.00 = excellent (Cicchetti & Sparrow, 1981).

Table 5
BPI-01 test–retest reliability (ICC correlation coefficients) across three assessments.

	Ratings	
	Frequency	Severity
Self-Injurious Behavior	.69	.65
Stereotyped Behavior	.70	.70
Aggressive/Destructive Behavior	.77	.80
Total score	.78	.74

ρ = .41 to .64 (M = .53). The mean test–retest coefficients were .50 and .46 (for frequency and severity scores respectively) for the SIB subscale, .58 and .53 for the stereotyped behavior subscale, and .57 and .59 for the Aggressive/Destructive Behavior subscale. Second, we computed intraclass correlations using a two way mixed effects model with three ratings and each participant representing a level of the random person factor (McGraw & Wong, 1996). Intraclass correlation coefficients ranged from .68 to .80 (see Table 5).

2.4. Consistency of BPI-01 subscale scores over time

To investigate whether systematic changes occurred in BPI-01 subscale scores across the three assessment periods, three MANOVAs with repeated measures were computed, one for each subscale. In each case the frequency and the severity scores were used as the two multiple dependent variables (n = 180). The multivariate test for subscale *Self-Injurious Behavior* was significant, *Wilks*' Λ = .97, F[4,1072] = 3.8, $p < 0.05$. Tests between subjects effects showed that the frequency scores (F[2,537] = 4.67; $p < .05$) and the severity scores changed significantly across time. More specifically, frequency scores and severity scores declined between assessment periods 1 and 3 ($p < .01$ and $p < .001$ respectively). The multivariate test for subscale *Stereotyped Behavior* was significant, *Wilks*' Λ = .95, F[4,1072] = 7.44, $p < 0.001$. Tests between subjects effects showed that the frequency scores did not change over time. The severity scores, however did change, F[2, 537] = 7.16, $p < .01$. They declined between assessment periods 1 and 2 ($p < 05$) and 1 and 3 ($p < .01$). The multivariate test for subscale *Aggressive/Destructive Behavior* was not significant, meaning that there were no statistically significant differences in either

Table 6
Bonferroni adjusted post hoc multiple comparisons (LSD) for the MANOVAs for the BPI-01 subscale scores and total score.

Dependent variables	Time points		Mean difference	Std. error	Sig.
Self-Injurious Behavior					
Frequency	1	2	0.99	0.54	
	1	3	1.63	0.54	*
	2	3	0.64	0.54	
Severity	1	2	0.91	0.38	
	1	3	1.46	0.38	**
	2	3	0.55	0.38	
Stereotypic Behavior					
Frequency	1	2	1.21	1.28	
	1	3	2.86	1.28	
	2	3	1.66	1.28	
Severity	1	2	2.06	0.88	
	1	3	3.28	0.88	**
	2	3	1.22	0.88	
Aggressive/Destructive Behavior					
Frequency	1	2	0.45	0.74	
	1	3	0.64	0.74	
	2	3	0.19	0.74	
Severity	1	2	0.57	0.62	
	1	3	0.88	0.62	
	2	3	0.31	0.62	
BPI-01 total score					
Frequency	1	2	2.64	2.01	
	1	3	5.13	2.01	
	2	3	2.49	2.01	
Severity	1	2	3.53	1.54	
	1	3	5.62	1.54	**
	2	3	2.08	1.54	

* Bonferroni adjusted p of .05 = .0083.
** p of .01 = .0016.

the frequency scores or the severity scores. The multivariate test for the BPI-01 total scores was also significant, *Wilks*' $\Lambda = .96$, $F[4,1072] = 5.19$, $p < 0.001$. Tests between subjects effects showed that the frequency scores did not change over time. However, The severity scores changed, $F[2, 537] = 6.7$, $p < .001$. They declined between assessment periods 1 and 2 ($p < 05$) and 1 and 3 ($p < .01$) (Table 6).

2.5. Convergent validity

Convergent validity was determined by computing correlations between subscales across the instruments intended to measure the same construct. Since none of the subscales had normally distributed data, non-parametric Spearman's ρ coefficients were computed (see Table 7). [Skewed distributions in behavior problems are common among populations with IDD.] Coefficients in bold font represent correlations between subscales across the instruments that were designed to measure the same construct. For instance, SIB was assessed by the identical label subscales of the RBS-R and the BPI-01. Regular-font coefficients also show expected significant correlations estimates for subscales that measure related constructs. For instance, the ABC *Irritability* subscale was seen as partly analogous with the SIB subscales of the RBS-R and the BPI-01 because three of its 15 items explicitly capture SIBs. All of these subscales were significantly correlated with one another at the .01 level (see Table 7). Similarly, the RBS-R *Stereotypy* subscale, the BPI-01 *Stereotyped Behavior Subscale*, and the ABC *Stereotypy* subscale were all significantly correlated ($\rho = .36, .50, .41.$ and $.54$). Aggressive behavior was measured by the BPI-01 *Aggressive/Destructive Behavior* subscale and in part by the ABC *Irritability* subscale which contained three items related to aggressive behavior. They were all significantly correlated at the .01 level.

2.6. Discriminant validity

Discriminant validity is reflected by weak, non-significant correlations between the scores of theoretically independent subscales, or by negative correlations between subscales measuring theoretically incompatible constructs. In Table 7, underlined numbers indicate a relationship between unrelated constructs. Identifying evidence for discriminant validity is more difficult than finding evidence for convergent validity because of the complex relationships between aberrant behavior and psychopathology which are mostly not entirely independent. Therefore, it was difficult to find subscales that measured theoretically independent or incompatible constructs among the three rating instruments. We anticipated that the ABC

Table 7
Multitrait-multimethod matrix for the convergent and discriminant validity: spearman ρ correlations across subscales at time 2.

		BPI-01 (s)			ABC					RBS-S					
		SIB	Stereotypies	Agg/Dest	Irritability	Lethargy	Stereotypy	Hyperactivity	Exc. Speech	Stereotypy	SIB	Compulsions	Rituals	Sameness	Restricted
BPI-01 (f)	SIB	.90**	.43**	.49**	.50**	.20*	.26**	.24**	.17*	.19*	.54**	.07	.22*	.12	.11
	Stereotypy	.46**	.92**	.29**	.39**	.38**	.50**	.33**	.27**	.41**	.36**	.26**	.09	.26**	.34**
	Agg/Dest	.53**	.32**	.94**	.51**	.07	-.16	.39**	.08	-.01	.09	.27**	.15	.21*	.16
ABC	Irritability	.49**	.36**	.52**		.67**	.39**	.77**	.47**	.27**	.38**	.36**	.33**	.35**	.31**
	Lethargy	.15	.32**	.10			.51**	.71**	.61**	.36**	.24**	.24**	.16	.25**	.23*
	Stereotypy	.17	.35**	-.12				.33**	.34**	.55**	.31**	.20*	.17*	.25**	.27**
	Hyperactivity	.22*	.27**	.39**					.52**	.21*	.07	.32**	.17	.25**	.20*
	Exc. Speech	.16	.26**	.14						.23*	.09	.27**	.03	.19*	.11
RBS-R	Stereotypy	.12	.36**	.02							.42**	.36**	.38**	.52**	.51**
	SIB	.49**	.27**	.13								.17*	.27**	.22*	.16
	Compulsions	.11	.28**	.28**									.52**	.64**	.44**
	Rituals	.15	.08	.16										.67**	.53**
	Sameness	.18*	.29**	.25**											.58**

*$p < .05$.
** = $p < .01$.
Note: Numbers in bold boxes represent expected convergent validity for subscales with identical labels; regular boxes show expected convergent validity for subscales that measure related constructs; underlined number signal expected discriminant validity.

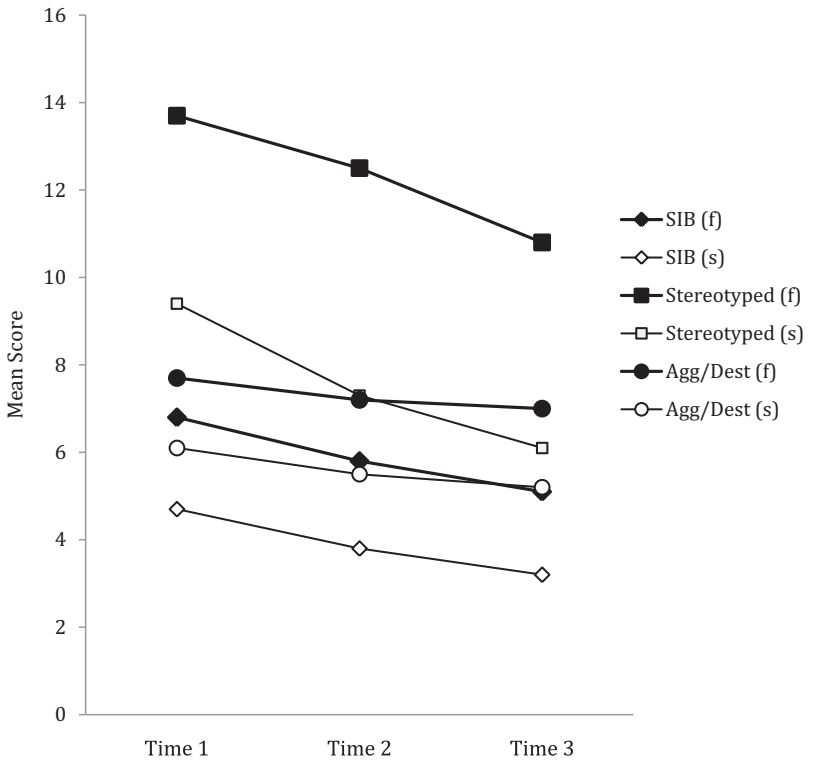

Fig. 1. Treatment sensitivity of the BPI-01 across three assessment times.

Table 8
Sensitivity and specificity estimates of the BPI-01 SIB and Stereotyped Behavior Subscale Scores (treated as test variable) compared to the Analogous Subscales of the RBS-R and the ABC (treated as condition variable).

Subscale	Test	Condition				95% CI	
SIB		Absent[2]	Present[1]	Totals		lower	upper
	BPI-01	RBS-R					
	Positive[1]	28	58	86	Sensitivity =.98	.90	1.00
	Negative[2]	10	1	11	Specificity =.26	.14	.43
	Totals	38	59	97			
Stereotypy	BPI-01	RBS-R					
	Positive	19	65	84	Sensitivity =.92	.82	.97
	Negative	7	6	13	Specificity =.27	.12	.48
	Totals	26	71	97			
Stereotypy	BPI-01	ABC					
	Positive	25	59	84	Sensitivity =.95	.86	.99
	Negative	10	3	13	Specificity =.29	.15	.47
	Totals	35	62	97			

[a]Score of 1 or higher.
[b]Score of 0.
CI = confidence interval.

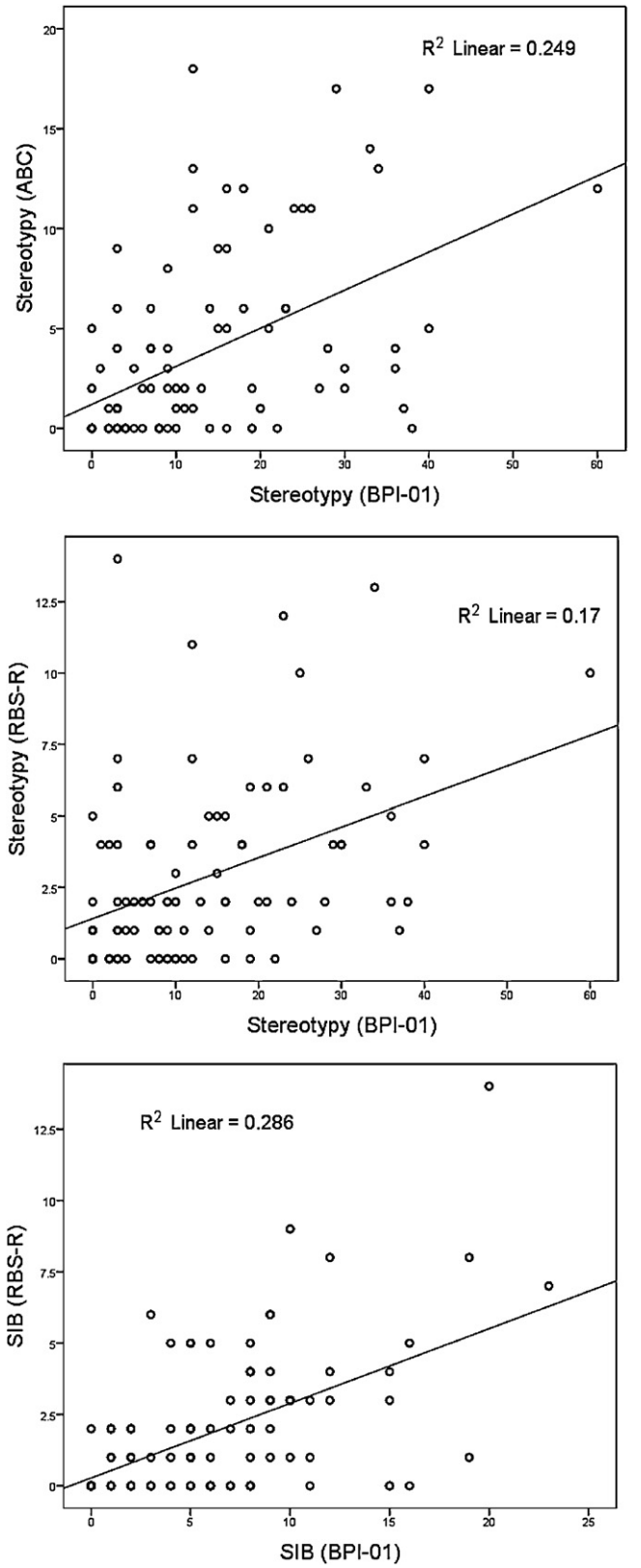

Fig. 2. Scatter plots with linear regression lines between corresponding SIB and stereotypic behavior BPI-01 (frequency ratings), ABC, and RBS-R subscales (as observed at the 2nd point of measurement). *Note*: Dots in the scatter plots can represent one or more cases.

Excessive Speech subscale should be independent of the *Self-Injurious Behavior* and *Aggressive/Destructive Behavior* subscales of the BPI-01, and the RBS-R *Self-Injurious Behavior* subscale. Indeed, none of these subscales were significantly correlated (see Table 7). We found similar evidence for discriminant validity between the ABC *Lethargy* subscale, and the *Self-Injurious Behavior* and *Aggressive/Destructive Behavior* subscales.

However, to our surprise several of the subscales that measured seemingly unrelated constructs were significantly correlated. For example, we expected that the ABC *Lethargy* subscale and the RBS-R *Self-Injurious Behavior* subscale would not be related. However, the correlation was significant ($\rho = .24$) but weak. As mentioned before, this unanticipated correlation may be accounted for by the fact that many of the constructs measured by those subscales are non-independent (Fig. 1).

2.7. Sensitivity and specificity

Sensitivity and specificity was computed for the BPI-01 in relation to the RBS-R, and the ABC. Only those subscales were chosen that carried the same label. This means the *Self-Injurious Behavior* and the *Stereotyped Behavior* subscales of the BPI-01, the RBS-R, and the ABC. Table 8 shows the sensitivity and specificity estimates of the BPI-01 subscales for *Self-Injurious Behavior* and *Stereotyped Behavior* compared to the analogous subscales of the RBS-R and the ABC. The BPI-01 was treated as the "test" variable and the ABC and the RBS-R were treated as the "condition" variable. Table 8 indicates that the BPI-01 sensitivity scores were very high (ranging between .92 and .98) but that specificity was low (ranging from .26 to .29). This means that the BPI-01 identified most cases that were also identified by the ABC and the RBS-R as having SIB and stereotyped behaviors (i.e., high BPI-01 sensitivity), but that it identified more individuals as having SIB and stereotyped behaviors than the ABC and the RSB-R (i.e., low BPI-01 specificity).

Fig. 2 presents the scatter plots of the subscale raw scores of the BPI-01, the ABC and the RSB-R. Linear regression estimates ranged from .17 to .29. Panel 1 contrasts the SIB subscales of the BPI-01 and the RBS-R. As shown in Table 8, there were many cases ($n = 28$) that received positive frequency scores on the BPI-01 *Self-Injurious Behavior* subscale (up to a score of 16) that received zero ratings on the analog RBS-R subscale. A closer look at the types of SIB topographies that were missed during the RBS-R interviews but that were identified by the BPI-01 shows that there were 16 cases with teeth grinding (frequency ratings ranging from 1 to 3 and severity ratings from 1 to 2), 14 cases of pica (frequency 1–4, severity 1–3), six cases of inserting fingers into body openings (frequency 1–3, severity 1–3), five cases of head hitting (frequency 2–3, severity 1–2), four cases of vomiting (frequency 2–3, severity 1–3), three cases of hair pulling (frequency 1–2, severity 1), three cases of extreme drinking (frequency 1 to 4, severity 1 to 3), two cases of biting (frequency 2, severity 1), two cases of pinching (frequency 1–2, severity 1), two cases of nail pulling (frequency 3–4, severity 2–3), one case with inserting objects (frequency 3, severity 1), and one case with air swallowing (frequency 3, severity 1). Conversely, there was only one case that received a score >0 on the RBS-R and a zero on the BPI-01 (self hitting and inserting fingers).

Panels 2 and 3 compare the number of cases the three instruments identified with regard to stereotyped behavior. While there were a good number of the same cases identified, each instrument identified cases that were not identified by the other instruments.

3. Discussion

The BPI-01 descriptive statistics were comparable to BPI-01 "clinical norms" of the instrument for children from birth to ten years old published by Rojahn et al. (2012a). This suggests that the amount of aberrant behavior in this sample was consistent with that in other peer groups with IDD in the United States and abroad who are at risk for severe behavioral concerns.

As far as this sample of participants is concerned, internal consistency was excellent for all ABC subscales. The RBS-R and the BPI-01 had lower internal validity scores that ranged between fair and excellent. The lower scores on the RBS-R are in part due to the small number of items in those subscales, which inherently lowers Cronbach's alpha. Somewhat surprising was the steady decline of the BPI-01 *Self-Injurious Behavior* subscale reliability across the three assessment times for which we do not have a ready explanation. BPI-01 test–retest reliability was found to be moderate. While this may be indicative of measurement instability (error), it is most likely a reflection of behavioral inter-individual variability across time due to the young age and/or due to idiosyncratic response to intervention. In fact, declines in the rates of behavior as found in the BPI-01 *Self-Injurious Behavior* subscale for frequency and severity, the *Stereotyped behavior* subscale for severity, and the total score for severity, would indicate a potentially successful intervention.

The concurrent validity analysis shows that the three assessment instruments measure a number of similar and dissimilar constructs. Primary convergent validity was examined by comparing subscales across instruments that carried the same label, namely SIB (on the BPI-01 and RBS-R) and stereotyped behavior (on the ABC, BPI-01, and RBS-R). Despite the different item content across instruments as shown in Tables 1 and 2, these subscale scores were also significantly correlated. Evidence for discriminant validity is harder to come by than for confirmatory validity because many aberrant behavior constructs are known not to be entirely independent from one another.

In summary, each one of the three instruments showed decent to excellent psychometric properties to the extent that they were examined in this study. Based on these findings, we suggest that there is merit in using all three scales simultaneously in research studies on behavior problems in children at risk for IDD. The subscales of the measures differ and using them in conjunction may provide a more complete picture of other aberrant, but perhaps less severe, behavior

problems. The identification of these behavior problems is equally important as they often accompany SIB, aggression, and stereotyped behavior.

Another outcome of this study is that these three measures are now further validated and have shown versatility in the assessment of infant populations at risk for IDD. Early identification and intervention are some of the greatest tools we have to treat children with IDD and to prevent symptom exacerbation and consolidation. Problem behaviors are intricately linked to prognosis. By developing well validated assessment tools we can also enhance correct identification rates of early signs of behavior problems while also using resources more effectively. The ultimate goal is to improve the well-being and prognosis for these infants and children. Future studies should aim to test the validity of these measures in older children to broaden their applicability.

References

Aman, M. G., & Singh, N. N. (1986). *Aberrant Behavior Checklist: Manual*. East Aurora, NY: Slosson Educational Publications.

Aman, M. G., Singh, N. N., Stewart, A. W., & Field, C. J. (1985a). Psychometric characteristics of the Aberrant Behavior Checklist. *American Journal on Mental Deficiency, 89*, 492–502.

Aman, M. G., Singh, N. N., Stewart, A. W., & Field, C. J. (1985b). The Aberrant Behavior Checklist: A behavior rating scale for the assessment of treatment effects. *American Journal on Mental Deficiency, 89*, 491–495.

Bayley, N. (2006). *Manual for the Bayley Scales of Infant Development* (3rd ed.). San Antonio, TX: Psychological Corporation.

Bihm, E. M., & Poindexter, A. R. (1991). Cross-validation of the factor structure of the Aberrant Behavior Checklist for persons with mental retardation. *American Journal on Mental Retardation, 96*(2), 209–211.

Bodfish, J. W., Symons, F. J., Parker, D. E., & Lewis, M. H. (2000). Varieties of repetitive behavior in autism: Comparisons to mental retardation. *Journal of Autism and Developmental Disorders, 30*(3), 237–243. http://dx.doi.org/10.1023/A:1005596502855.

Boyd, B. A., McDonough, S. G., & Bodfish, J. W. (2012). Evidence-based behavioral interventions for repetitive behaviors in autism. *Journal of Autism and Developmental Disorders, 42*(6), 1236–1248.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin, 56*(2), 81–105. http://dx.doi.org/10.1037/h0046016.

Cicchetti, D. V., & Sparrow, S. A. (1981). Developing criteria for establishing interrater reliability of specific items: Applications to assessment of adaptive behavior. *American Journal of Mental Deficiency, 86*(2), 127–137.

Dawson, G. (1996). Brief report: Neuropsychology of autism: A report on the state of the science. *Journal of Autism and Developmental Disorders, 26*(2), 179–184.

Dunlap, G., Strain, P. S., Fox, L., Carta, J. J., Conroy, M., Smith, B. J., et al. (2006). Prevention and intervention with young children's challenging behavior: Perspectives regarding current knowledge. *Behavioral Disorders, 32*(1), 29–45.

Lam, K. S. L., & Aman, M. G. (2007). The Repetitive Behavior Scale-Revised: Independent validation in individuals with autism spectrum disorders. *Journal of Autism and Developmental Disorders, 37*(5), 855–866. http://dx.doi.org/10.1007/s10803-006-0213-z.

Karabekiroglu, K., & Aman, M. G. (2009). Validity of the Aberrant Behavior Checklist in a clinical sample of toddlers. *Child Psychiatry and Human Development, 40*(1), 99–110. http://dx.doi.org/10.1007/s10578-008-0108-7.

Matson, J. L., Wilkins, J., Sharp, B., Knight, C., Sevin, J. A., & Boisjoli, J. A. (2009). Sensitivity and specificity of the Baby and Infant Screen for Children with Autism Traits (BISCUIT): Validity and cutoff scores for autism and PDD-NOS in toddlers. *Research in Autism Spectrum Disorders, 3*(4), 924–930. http://dx.doi.org/10.1016/j.rasd.2009.04.001.

Mayo-Ortega, L., Oyama-Ganiko, R., Leblanc, J., Schroeder, S. R., Brady, N., Butler, M. G., et al. (2012). Mass screening for severe problem behavior among infants and toddlers in Peru. *Journal of Mental Health Research in Intellectual Disabilities, 5*(3–4), 246–259. http://dx.doi.org/10.1080/19315864.2011.590626.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30–46. http://dx.doi.org/10.1037/1082-989X.1.1.30 (*Correction, 1*(4), 390. http://dx.doi.org/10.1037/1082-989X.1.4.390).

Oyama, R., Mayo, L., Schroeder, S. R., & LeBlanc, J. M. (2012). Early distance intervention and follow-up for families of infants and toddlers at risk for developmental disabilities and severe behavior problems in Peru, in preparation.

Paclawskyi, T. R., Matson, J. L., Bamburg, J. W., & Baglio, C. S. (1997). A comparison of the Diagnostic Assessment for the Severely Handicapped-II (DASH-II) and the Aberrant Behavior Checklist (ABC). *Research in Developmental Disabilities, 18*(4), 289–298. http://dx.doi.org/10.1016/S0891-4222(97)00010-3.

Rojahn, J., Matson, J. L., Lott, D., Esbensen, A. J., & Smalls, Y. (2001). The Behavior Problems Inventory: An instrument for the assessment of self-injury, stereotyped behavior and aggression/destruction in individuals with developmental disabilities,. *Journal of Autism and Developmental Disorders, 31*(6), 577–588. http://dx.doi.org/10.1023/A:1013299028321.

Rojahn, J., Rowe, E. W., Sharber, A. C., Hastings, R. P., Matson, J. L., Didden, R., et al. (2012a). The Behavior Problems Inventory-short form (BPI-S) for individuals with intellectual disabilities. Part I: Development and provisional clinical reference data. *Journal of Intellectual Disability Research, 56*(5), 527–545. http://dx.doi.org/10.1111/j.1365-2788.2011.01507.x.

Rojahn, J., Rowe, E. W., Sharber, A. C., Hastings, R. P., Matson, J. L., Didden, R., et al. (2012b). The Behavior Problems Inventory-short form (BPI-S) for individuals with intellectual disabilities. Part II: Reliability and validity. *Journal of Intellectual Disability Research, 56*(5), 546–565. http://dx.doi.org/10.1111/j.1365-2788.2011.01506.x.

Rojahn, J., Schroeder, S. R., & Hoch, T. A. (2008). *Self-injurious behavior in intellectual disabilities*. New York, NY: Elsevier.

Schopler, E., Reichler, R. J., & Renner, B. R. (1988). *The Child Autism Rating Scale*. Los Angeles, CA: Western Psychological Services Corporation.

Schroeder, S. R., & Courtemanche, A. (2012). Early prevention of severe neurodevelopmental disorders: An integration. *Journal of Mental Health Research in Intellectual Disabilities, 5*, 203–214.

Sigafoos, J., Lancioni, G. E., Didden, R., & O'Reilly, M. F. Early signs/development of challenging behavior and early intervention. In R. Hastings & J. Rojahn (Eds.), *International Review of Research in Developmental Disabilities – Challenging Behavior*. Waltham, MA: Elsevier, Inc., in press.

Wetherby, A. M., & Prizant, B. M. (2002). *Communication and Symbolic Behavior Scales Developmental Profile*. Baltimore, MD: Paul Brookes Publishing Co Inc.