A Psychometric Evaluation of the Motor-Behavioral Assessment Scale for Use as an Outcome Measure in Rett Syndrome Clinical Trials

Melissa Raspa, Carla M. Bann, Angela Gwaltney, Timothy A. Benke, Cary Fu, Daniel G. Glaze, Richard Haas, Peter Heydemann, Mary Jones, Walter E. Kaufmann, David Lieberman, Eric Marsh, Sarika Peters, Robin Ryther, Shannon Standridge, Steven A. Skinner, Alan K. Percy, and Jeffrey L. Neul

Abstract

Rett syndrome (RTT) is a neurodevelopmental disorder that primarily affects females. Recent work indicates the potential for disease modifying therapies. However, there remains a need to develop outcome measures for use in clinical trials. Using data from a natural history study ($n = 1,075$), we examined the factor structure, internal consistency, and validity of the clinician-reported Motor Behavior Assessment scale (MBA). The analysis resulted in a five-factor model: (1) motor dysfunction, (2) functional skills, (3) social skills, (4) aberrant behavior, and (5) respiratory behaviors. Item Response Theory (IRT) analyses demonstrated that all items had acceptable discrimination. The revised MBA subscales showed a positive relationship with parent reported items, age, and a commonly used measure of clinical severity in RTT, and mutation type. Further work is needed to evaluate this measure longitudinally and to add items related to the RTT phenotype.

Keywords: Rett syndrome, psychometrics, outcome measure, clinical trial

Background

Rett syndrome (RTT, MIM312750), a rare neurodevelopmental disorder occurring predominantly in females, is usually caused by mutations in the *Methyl-CpG-binding Protein 2* (*MECP2*) gene at Xq28 (Amir et al., 1999). Initially described in 1966 by Andreas Rett, and later detailed in the landmark paper of Hagberg and colleagues (Hagberg et al., 1983), RTT remains a clinical diagnosis. Classic RTT is indicated by a wide range of impairments; after a period of normal development, people with classic RTT experience a regression of spoken communication and fine motor skills, and develop hand stereotypies and gait abnormalities (Neul et al., 2010). Other co-occurring conditions in RTT include seizures, growth deceleration, breathing abnormalities, gastrointestinal problems, scoliosis, sleep disturbances, mood disorders or anxiety, impaired functional skills, and intellectual disability (Gold et al., 2017; Leonard, et al., 2017; Schultz & Glaze, 2017).

Traditionally, the management of RTT has used a multidisciplinary approach (e.g., a team consisting of a pediatrician, neurologist, gastroenterologist, speech-language pathologist, physical therapist, and occupational therapist) to address symptoms. Symptomatic pharmacological treatments have been used to address medical comorbidities and specialized interventions have been implemented to improve physiological, behavior, or functional abilities. Currently no disease modifying therapies have been approved. However, work in mouse models of RTT have shown that restoration of the gene product, even after symptom onset, can modify or reverse the phenotypes, leading to hope that true disease modifying therapies might be developed for people with RTT (Guy et al., 2007). Adding to this optimism, initial preclinical and clinical

M. Raspa et al.

493

evaluations of several new therapeutics in the last decade have shown promise to potentially modify the course of RTT (Djukic et al., 2016; Galdalla et al., 2011; Glaze et al., 2017; Glaze et al., 2019; Leonard et al., 2017; Moretti & Zoghbi, 2006; Wang et al., 2015).

Although the rise in clinical trials gives hope to individuals and their families, many challenges remain to realize the phenotypic reversal that has been reported in Mecp2 animal studies (Katz et al., 2016). In order to translate these successes into clinical improvement, establishment of accurate, robust outcome measures that are sensitive to treatment effects is essential. This has proved to be challenging in many neurodevelopmental disabilities (Berry-Kravis et al., 2013; Jeste & Geschwind, 2016). One of the major translational obstacles has been how to define and measure outcomes for people with neurodevelopmental disabilities, including RTT, who often have a heterogeneous phenotype. A crucial issue is the lack of specificity of measures used to describe people with neurodevelopmental disabilities. In fragile X syndrome, for example, results of a recent clinical trial found that the primary endpoint measure, the Aberrant Behavior Checklist, was not able to detect treatment effects (Berry-Kravis et al., 2016). Another measurement challenge in clinical trials for those with neurodevelopmental disabilities is the difficulty in assessing cognition using standardized assessments due to low functioning levels resulting in floor effects (Sansone et al., 2014).

Despite these challenges, some measurement development work has been conducted in RTT in order to identify an adequate outcome measure for use in clinical trials. The Rett Syndrome Behavior Scale (RSBQ) is a well-known parent-reported measure in RTT and to date has been used as the primary outcome measure in most RTT clinical trials. The RSBQ measures behavioral and emotional features as well as movement abnormalities. The reliability and validity of the RSBQ, originally developed as a clinical diagnostic tool prior to the availability of genetic testing for RTT (Mount et al., 2002), has been examined. Although internal consistency, test-retest and inter-rater reliability, and convergent and discriminant validity were adequate for the fear/anxiety subscale, the other seven subscales were not examined in-depth (Barnes et al., 2015). Other measures to assess gait abnormalities using a variety of accelerometer-type devices have shown

promise but may be influenced by stride rate (Downs et al., 2015). Additionally, gait measures are only appropriate for use in those individuals that remain ambulatory (approximately 50%). Modifications to a global developmental measure, the Mullen Scales of Early Learning (MSEL), has also been conducted. Clarkson and colleagues (2017) made adaptations to the MSEL, including allowing more time to respond to items and accepting eye gaze as a response method, in order to more accurately assess the development of individuals with RTT. Other studies have also focused on the use of eye tracking as a measure of cognition in RTT (Ahonniska-Assa et al., 2018; Schwartzman et al., 2015). The Clinical Global Impression Scale, which is commonly used as a primary endpoint in clinical trials, has been modified to create RTT-specific anchors in order to improve specificity (Neul et al., 2015). However, there is a pressing need to develop psychometrically sound and well validated measures that assess the multiple and varied domains that are impacted by RTT.

To aid in the development and psychometric assessment of outcome measures, it is essential to have natural history data on a large cohort of participants. The *Rett Syndrome, MECP2 Duplication, and Rett-related Disorders Natural History Study* (RTT NHS; 3U54HD061222-14) has been gathering data since 2006 across a variety of domains from historical forms, physical examinations, and global measures of clinical severity. One such measure is the Motor Behavior Assessment scale (MBA). The MBA, a clinician-reported measure typically completed by a neurologist, was originally developed to survey movement abnormalities, especially extrapyramidal symptoms, behavioral problems, and abnormal physiological features in individuals with RTT (FitzGerald et al., 1990). However, it has been expanded and refined over the last several years in order to further define clinical features in RTT and provide defined anchors to aide interrater reliability. In addition to assessing gross and fine motor skills, the current version also includes items that measure the severity of orofacial and respiratory abilities, social and communication skills, adaptive behaviors (e.g., feeding difficulties, toiler training), and seizures. The intention of these revisions was to capture the wide phenotypic variability found in individuals with RTT. The MBA is associated with developmental milestone attainment (Neul et al., 2014) and physical aspects of quality of life (Lane

et al., 2011). The MBA was selected for further study as a potential outcome measure because of the extensive amount of natural history data available, it is clinician-reported (as opposed to the parent-reported RSBQ), and it captures a broad range of RTT symptoms, all of which are noted as important characteristics of outcomes measures for rare disease (Benjamin et al., 2017). The MBA could be used in symptom-based clinical trials in RTT or ones that are disease modifying. The MBA, though, has not undergone formal psychometric evaluation, which is a critical component recommended for outcome measure development (Powers et al., 2017).

Given that few measurement-focused studies have been conducted on RTT-specific measures including none that have focused on a clinician-reported tool that captures the breadth of RTT functioning, the goal of this study was to examine the factor structure, reliability, and validity of the MBA to aid in the development of outcome measures for clinical trials in RTT. We answered three research questions:

1. Using factor analysis and item response theory, what is the best fit for the diverse items on the MBA?
2. Are the revised MBA factors unique and do they have strong internal consistency?
3. Are the revised MBA factors associated with similar constructs rated by parents?

Methods

Participants

Participants were from the RTT NHS, part of the Rare Disease Clinical Research Network. From 2006–2014 (Study #5201), participants were enrolled at four sites (University of Alabama-Birmingham, Baylor College of Medicine, Greenwood Genetic Center, Boston Children's Hospital), and from 2014 (Study #5211) at an additional ten sites (Oakland Children's Hospital, University of California San Diego, Children's Hospital Colorado, St. Louis Children's Hospital and Washington University School of Medicine, Gillette Children's Hospital, Rush University, Vanderbilt University Medical Center, Cincinnati Children's Hospital, and Children's Hospital of Philadelphia). Inclusion criteria for the psychometric analyses reported here included having a diagnosis of classic Rett syndrome (based on clinical assessment using accepted clinical criteria (Neul, 2010) by an experienced child neurologist, geneticist, or developmental pediatrician) and being between the ages of 1 and 25 years. Only data from the baseline clinical visit was used.

In total, there were 1,075 participants. We randomly selected two-thirds of the participants for the development sample ($n = 713$) which was used to identify the factor structure and the remaining one-third were used as a validation sample ($n = 362$). Virtually all participants were females except for four males (see Table 1). The average age was approximately 8 years. The majority of participants were white and non-Hispanic. The average Clinical Severity Scale score (Neul et al., 2008) was 23.74 ($SD = 7.6$; range 0 to 58, with higher scores indicating greater severity). Most participants (59%) had one of the common point mutations (see Table 2).

Measures

Motor-Behavior Assessment (MBA)

The MBA, a clinician-reported measure, is comprised of 34 items scored using an ordinal scale (0 to 4) with a variety of response options measuring either frequency or severity of skills, with higher scores indicating greater severity. Although the original measure consisted of 37 items, 3 items (oculogyric crisis, masturbation, and hypomimia) were removed prior to the start of the 2014 data collection period as they showed very strong floor effects and thus were present in only a handful of participants with RTT. The items were developed to align with three conceptual domains: behavioral/social (15 items), orofacial/respiratory (7 items), and motor/physical (12 items); however, no previous evaluation on how these items cluster into these domains has been performed. Within the context of the RTT NHS, site investigators were trained in person in the use of the MBA by Drs. Percy, Neul, or Glaze in order to ensure consistency in scoring. The MBA was completed during the clinic visit based on clinician observation with the exception of seven items (e.g., feeding difficulties) that were assessed based on history, which may have included parent input. The MBA was used in the psychometric analyses. Table 3 presents a frequency distribution of the 34 items.

Table 1
Sample Description

	All (n = 1075)	Development Sample (n = 713)	Validation Sample (n = 362)
Female	99%	99%	99%
Age in years, mean (min–max)	8.18 (1.07–24.48)	8.29 (1.07–24.48)	7.96 (1.47–24.00)
Race			
White	83%	83%	83%
Black	5%	4%	5%
Asian	4%	4%	3%
Other	3%	3%	5%
More than 2	5%	6%	5%
Ethnicity (Hispanic Origin)	16%	15%	17%
CSS[a] Total Score, mean (SD)	23.74 (7.60)	23.19 (7.35)	24.01 (7.71)
Total MBA score, mean (SD)	47.49 (14.11)	47.92 (14.23)	46.64 (13.84)
Study			
5201 only[b]	52%	52%	52%
5211 only[c]	20%	22%	19%
Both	28%	26%	29%

[a]CSS = Clinical Severity Score.
[b]5201 = Phase 1 of the Natural History Study (2006–2014).
[c]5211 = Phase 2 of the Natural History Study (2014–2020).

Table 2
Participant Mutation Type

	All (n = 1075)	Development Sample (n = 713)	Validation Sample (n = 362)
Common Point Mutations			
R106W	3%	3%	4%
R133C	6%	6%	5%
T158M	10%	10%	10%
R168X	11%	11%	9%
R255X	9%	10%	9%
R270X	6%	5%	6%
R294X	6%	6%	6%
R306C	8%	9%	8%
Other Point Mutations	6%	6%	6%
C-terminal Truncations	10%	10%	11%
Early Truncating	9%	9%	7%
Large Deletion	9%	8%	10%
Exon1	1%	1%	2%
Splice Site	1%	1%	1%
None	2%	2%	3%
Missing	3%	2%	3%

Interval History Form

During clinic visits, parents were asked to complete the Interval History Form which assessed the child's functioning over the previous 6 months. The questionnaire covered several domains which overlap with the MBA, including communication; hand use; sitting, standing, and walking; mood and abnormal behaviors; and Rett-specific behaviors. Similar to the MBA, parents were asked to rate their child's functioning using an ordinal scale, with a variety of response options that measured either frequency or severity. Higher scores indicate more clinical severity. The majority of items that appear on the Interval History Form had similar wording and response options to those on the MBA.

Clinical Severity Scale

The Clinical Severity Scale was developed as part of the RTT NHS to assess common clinical features, including age at regression, age at stereotypy onset, degree of deceleration of head growth, growth (BMI) status, sitting, walking, hand function, scoliosis, vocalization/ verbalization, eye contact, periodic breathing, hand/foot skin temperature, and seizures. Each of the 13 items are rated on their own ordinal scale (scores of 0 to 4 or 0 to 5). The scale was completed by the clinician at the same visit as the MBA. A total sum scores is calculated, with higher scores indicating more clinical severity. Both this measure and the Interval History Form were used in the validation analyses.

Statistical Analyses

Two sets of psychometric analyses were performed: factor analysis and IRT. The factor analysis was used to determine the underlying conceptual structure of the items on the MBA. To optimize the replicability of the factor structure, we applied a split sample validation approach, with two-thirds in the development sample and one-third in the validation sample. The IRT analysis provided additional information about the individual items, including how well the items can discriminate between individuals with different levels or amounts on the construct being measured as well as the functioning of the response options.

First, we used the development sample to determine an appropriate factor structure for the items. The validation sample was used to test the consistency of the factor model once it was finalized based on data from the development sample. Utilizing data from the development sample, we conducted a series of exploratory factor analyses in SAS PROC FACTOR (SAS Institute, 2012–2017) to determine the best fitting factor structure for the items. We applied an oblique rotation method, Promax, to allow for correlations between the factors and fit models with solutions of two to six factors. The most appropriate factor solution was determined based on the pattern of factor loadings (i.e., demonstration of simple structure), size of the factor loadings (above 0.40), and the percentage of variance accounted for by each factor. In addition to the statistical results, clinical and content considerations informed the selection of the final factor structure. Once we determined the final factor structure in the development sample, we ran a confirmatory factor model in Mplus (Muthén & Muthén, 1998–2017) to test the fit of the factor structure in the validation sample. To account for the categorical data, we used weighted least square mean and variance adjusted (WLSMV) estimation. Model fit was assessed based on fit indices, including the Tucker-Lewis index (TLI), comparative fit index (CFI), and root mean square error of approximation (RMSEA). The TLI and CFI are relative fit indices that compare our model to the null model whereas the RMSEA is an absolute fit index that measures how perfect a fit our model is. Values of 0.90 or greater for the TLI and CFI and values less than 0.08 for the RMSEA indicate acceptable fit (Schreiber et al., 2006).

After the factor structure was established, the psychometric properties of the items and factors were further examined using graded response IRT models conducted with IRTPRO software (Cai et al., 2011). These models estimate two types of parameters, slope (a) and threshold parameters (b1-b4). The slope indicates how well the item can differentiate between those with high versus low levels on the underlying construct (e.g., more or less motor dysfunction). Values of 1 or higher generally indicate acceptable discrimination; the higher the slope (i.e., the steeper the line), the better able that item is at discriminating individuals with different abilities. The threshold parameters locate the response options for each item along the continuum of the underlying construct at the point where a respondent would have 50% probability of endorsing the response option. In other words, higher threshold parameters indicate that an individual would need to be higher on the

Table 3
MBA Item Frequencies

Item	Response options				
	0	1	2	3	4
	N (%)	*N* (%)	*N* (%)	*N* (%)	*N* (%)
Motor skills[a]	38 (0.6)	741 (11.8)	1510 (24.0)	2170 (34.5)	1827 (29.1)
Communication skills[b]	117 (1.9)	3451 (54.9)	1412 (22.5)	652 (10.4)	658 (10.5)
Poor eye/social contact[c]	1661 (26.4)	2229 (35.4)	1598 (25.4)	615 (9.8)	189 (3.0)
Lack of sustained interest[c]	1711 (27.2)	1829 (29.1)	1737 (27.6)	740 (11.8)	274 (4.4)
Irritability/crying[c]	4957 (78.8)	995 (15.8)	223 (3.5)	71 (1.1)	46 (0.7)
Overactive/passive[c]	4179 (66.4)	624 (9.9)	620 (9.9)	477 (7.6)	392 (6.2)
Does not reach for objects or people[c]	1773 (28.2)	615 (9.8)	807 (12.8)	722 (11.5)	2373 (37.7)
Does not respond to spoken words/acts deaf[c]	1262 (20.1)	1894 (30.1)	1741 (27.7)	754 (12.0)	639 (10.2)
Feeding difficulties[d]	1838 (29.2)	1364 (21.7)	1405 (22.3)	1118 (17.8)	563 (9.0)
Chewing difficulties[e]	689 (11.0)	2427 (38.6)	1247 (19.8)	689 (11.0)	1235 (19.6)
Toilet training[f]	61 (1.0)	227 (3.6)	1280 (20.4)	1199 (19.1)	3520 (56.0)
Self-mutilation/pulling hair or ears/scratching[c]	5577 (88.7)	483 (7.7)	141 (2.2)	55 (0.9)	34 (0.5)
Aggressive behavior[c]	5713 (90.8)	418 (6.6)	110 (1.8)	30 (0.5)	20 (0.3)
Seizures[g]	2833 (45.1)	1344 (21.4)	947 (15.1)	697 (11.1)	466 (7.4)
Insensitivity to pain[h]	725 (11.5)	3238 (51.5)	1383 (22.0)	814 (13.0)	127 (2.0)
Speech disturbances[i]	8 (0.1)	59 (0.9)	611 (9.7)	4296 (68.3)	1318 (21.0)
Bruxism[c]	3389 (53.9)	1490 (23.7)	725 (11.5)	409 (6.5)	277 (4.4)
Breath holding[c]	2037 (32.4)	2169 (34.5)	1417(22.5)	452 (7.2)	216 (3.4)
Hyperventilation[c]	3134 (49.8)	1627 (25.9)	957 (15.2)	378 (6.0)	195 (3.1)
Air-saliva expulsion/drooling[c]	2125 (33.8)	1894 (30.1)	1238 (19.7)	530 (8.4)	504 (8.0)
Mouthing hands/objects[c]	3441 (54.7)	1072 (17.0)	705 (11.2)	560 (8.9)	512 (8.1)
Biting self/others[c]	5447 (86.6)	564 (9.0)	163 (2.6)	59 (1.0)	55 (1.0)
Hand clumsiness[j]	71 (1.1)	780 (12.4)	1045 (16.6)	1243 (19.8)	3152 (50.1)
Stereotypic hand activities[c]	144 (2.3)	300 (4.8)	577 (9.2)	1069 (17.0)	4200 (66.8)
Ataxia/apraxia[c]	87 (1.4)	181 (2.9)	165 (2.6)	214 (3.44)	5643 (89.7)
Truncal rocking[c]	3436 (54.7)	1153 (18.3)	1142 (18.2)	354 (5.6)	202 (3.2)
Bradykinesia[k]	4361 (69.3)	452 (7.2)	228 (10.6)	542 (8.6)	268 (4.3)
Dystonia[l]	2792 (44.4)	869 (13.8)	1715 (27.3)	593 (9.4)	319 (5.1)
Scoliosis[m]	2634 (41.9)	1562 (24.8)	749 (11.9)	381 (6.1)	963 (15.3)
Myoclonus[c]	5295 (84.2)	713 (11.3)	191 (3.0)	54 (0.9)	37 (0.6)
Dyskinesia[c]	4921 (78.3)	867 (13.8)	308 (4.9)	116 (1.8)	77 (1.2)

(Table 3 continued)

Table 3
Continued

Item	Response options				
	0	1	2	3	4
	N (%)	N (%)	N (%)	N (%)	N (%)
Hypertonia/rigidity[n]	3132 (49.8)	1021 (16.2)	568 (9.0)	759 (12.1)	810 (12.9)
Hyperreflexia[o]	4170 (66.3)	883 (14.0)	412 (6.6)	588 (9.4)	237 (3.8)
Vasomotor disturbances[p]	2616 (41.6)	1919 (30.5)	1082 (17.2)	610 (9.7)	59 (0.9)

[a] 0 = No regression, 1 = Dyspraxia of gait and hand use including bilateral pincer grasp, 2 = Able to walk and use one or both hands, 3 = Able to walk independently or with support or use one or both hands, 4 = No motor skills.

[b] 0 = Effective communication, 1 = Consistently makes choices (> 50% of time), 2 = Sometimes makes choices (10–50% of time), 3 = rarely makes choices (<10% of time), 4 = No communication.

[c] 0 = None, 1 = 25% of time, 2 = 50% of time, 3 = 75% of time, 4 = 100% of time.

[d] 0 = None, 1 = Occasional chocking/gagging, 2 = More than 30 min to feed, 3 = Oral and gastrostomy feeding, 4 = Gastrostomy only.

[e] 0 = None, 1 = Course chopped, 2 = Fine chopped, 3 = Pureed or mashed, 4 = Gastrostomy.

[f] 0 = Purposeful bowel and bladder control, continent at all times, 1 = Continent during the day, 2 = Time trained, both urine and stool, 3 = Time trained, urine or stool, 4 = Totally incontinent.

[g] 0 = None, 1 = None (last 6 months), with medications, 2 = Monthly, with or without medications, 3 = Weekly, with or without medications, 4 = Daily, with or without medications.

[h] 0 = Normal or immediate response to pain, 1 = delayed response (> 5 sec) to minor pain (shots/blood draw), 2 = No response to minor pain/delayed response to moderate pain (head bump/finger pinch/small laceration/small burn), 3 = No response to moderate pain/delayed response to major pain (major laceration/bone fracture/large burn), 4 = No response to pain of any type.

[i] 0 = Fluent, 1 = Phrases/sentences, 2 = Words with meaning or intention, 3 = Vocalizations, no words, 4 = No utterances

[j] 0 = Purposeful hand use, 1 = Plays with toys or activates switches purposefully, 2 = Uses utensils/cup, may be adaptive, 3 = Finger feed only, 4 = No purposeful hand use.

[k] 0 = None, 1 = Occasional paucity of limb movement (< 10% of time), 2 = Some limb movement (<50% of time), 3 = Occasional limb movement (<10% of time), 4 = Severe lack of limb movement (95–100% of time).

[l] 0 = None, 1 = Focal dystonia, one joint, 2 = Focal dystonia, more than one joint, 3 = Generalized dystonia, >2 extremities, 4 = Fixed positional deformity.

[m] 0 = None, 1 = 1–<20°, 2 = 20 -<40°, 3 = ≥40°, 4 = Surgery.

[n] 0 = None, 1 = Ankle hypertonia/rigidity, 2 = Upper or lower limb hypertonia/rigidity, 3 = Generalized hypertonia without contractures, 4 = Generalized hypertonia with contractures.

[o] 0 = Normal muscle stretch reflexes, 1 = 4+ ankles or knees, 2 = 4+ ankle and knees with clonus, 4+ all extremities with spread or clonus, 4 = 4+ all extremities with spread and clonus.

[p] 0 = Normal temperature and color, 1 = Temperature slightly off (cool or warm), normal color, 2 = Temperature extremely off (cold or hot), mottled, 3 = Temperature off in hands or feet, color changed (blue or red), 4 = Temperature off in hands and feet, color changed.

construct (e.g., have greater motor dysfunction) before endorsing that response option.

Next, we assessed the reliability and validity to determine if the MBA functions better as individual subscales or an overall scale. Internal consistency, which is one aspect of reliability, was assessed using Cronbach's alphas. Pearson correlations were calculated to examine the relationships among the subscales. To assess construct validity, we examined the relationship between the revised MBA subscales and similar items from the Interval History Form, a parent-reported measurement. A mean MBA subscale score was calculated for each item and response option on the parent-report measure. Because of small samples, some response options for items on the Interval History Form were collapsed. A continuous variable (degree of scoliosis) was categorized as none, mild (< 25 degrees), moderate (26–40 degrees), and severe (> 40 degrees). An ANOVA test was conducted to determine if there were statistically significant differences in MBA subscale scores by response category. A second validity comparison examined the correlation between the total revised MBA score and the Clinical Severity Scale, a commonly used measure of overall severity in RTT. A correlation was calculated to examine the relationship between age and the total revised MBA score. Finally, as a preliminary assessment of genotype/phenotype relationship using the total

revised MBA score, we grouped *MECP2* mutations into three groups: "Mild" (exon 1, R133C, R294X, R306C, and carboxy-terminal truncations); "Intermediate" (T158M); and "Severe" (early truncation mutations, R106W, R168X, R255X, R270X, large deletions) based on previously published genotype/phenotype correlation studies in RTT (Neul et al., 2008). An ANOVA test was conducted to determine if there were differences in the total revised MBA between these *MECP2* mutation groupings, with pair-wise post-hoc testing conducted using Bonferroni correction for multiple testing.

Results

Psychometric Analyses

A series of exploratory factor analyses were conducted to determine the best factor structure for the MBA scale. Our first exploratory factor analysis yielded a 5-factor solution. However, nine items did not load onto any factor because of low factor loadings. These included irritability/crying, overactive or over passive, aggressive behaviors, lack of toilet training, insensitivity to pain, mouthing hands/objects, ataxia/apraxia, myoclonus, and dyskinesia. Additionally, there were several items that did not fit conceptually with others on the same factor and had borderline factor loadings (0.41 to 0.46). Therefore, we ran a second exploratory factor with the items that did load onto factors as well as 2 of the above items (aggressive behaviors and mouthing hands/objects which was combined with stereotypic hand behaviors to improve item distribution) which did fit conceptually with other items but initially had low factor loadings. This factor analysis resulted in a similar but different 5-factor model. In this iteration, hyperreflexia and vasomotor disturbances were removed because of low factor loadings and did not fit conceptually with the new factors. In addition, seizures, stereotypic hand behaviors/mouthing hands/objects, and truncal rocking had low factor loadings. However, because of their clinical relevance in RTT, they were retained as single items and were included in a total score. For all exploratory analyses, there were no items that loaded on multiple factors. We labeled the 5 revised MBA (R-MBA) subscales (1) motor dysfunction, (2) functional skills, (3) social skills, (4) aberrant behavior, and (5) respiratory behaviors. The subscales contained 21 items with the 3 additional items included when calculating a total R-MBA score.

A confirmatory factor analysis was then conducted. Factor loadings from the confirmatory factor analyses of the development and validation samples are shown in Table 4. Only the 21 items within the 5 subscales were included in these analyses. Both the development sample (CFI = 0.94, TLI = 0.93, and RMSEA = 0.06) and validation sample (CFI = 0.93, TLI = 0.92, RMSEA = 0.06) had acceptable model fit, supporting the generalizability of the factor structure. All items had factor loadings of 0.40 or greater except for bruxism (loading = 0.32) and biting self or others (0.27) in the validation sample but this was likely due to the smaller sample size due to the one-third random split.

The IRT parameters for the items are provided in Table 5. The results are generally consistent across the development and validation samples. Items with high slope values (parameter estimate a) indicate that they are better at discriminating individuals who have more versus less of the characteristic being measured. All items demonstrated acceptable discrimination, with the majority of slopes having a value over 1 across both samples except for communication skills, does not respond to spoken words/acts deaf, bruxism, and air saliva expulsion/drooling. The threshold parameters (b1–b4) provide information on amount of ability needed to be scored at each response option. The lower the threshold parameter, the less of the trait or construct that is needed to be rated at that response option level. For each item, the expected ordering of values (from low to high) is found across the response options; less of a trait (i.e., severity of a given symptom) is needed at the b1 threshold whereas more of a trait is needed at the b4 threshold. Comparisons of threshold parameters also can be made across items. For example, the motor skills item has a low negative threshold for b1 (–2.55 in the development sample and –2.47 in the validation sample) which indicates very few individuals had a score of 0 (no motor skills dysfunction). Conversely, the aggressive behavior item has a positive threshold for b1 (1.59 in the development sample and 2.34 in the validation sample) indicating that more individuals scored a 0 (no aggressive behavior). Of note, the b4 threshold parameter was not able to be calculated for speech disturbances because there were no individuals in development sample who were rated at this response option.

Table 4
Factor Loadings of MBA Items by Sample

Factor/Item	Development Sample	Validation Sample
Factor 1: Motor Dysfunction		
Bradykinesia	0.86	0.83
Dystonia	0.67	0.59
Scoliosis	0.85	0.78
Hypertonia/rigidity	0.62	0.73
Factor 2: Functional Skills		
Hand clumsiness	0.87	0.87
Does not reach for objects or people	0.75	0.74
Motor skills	0.85	0.85
Speech disturbance	0.54	0.55
Communication skills	0.41	0.42
Feeding difficulties	0.58	0.55
Chewing difficulties	0.62	0.53
Factor 3: Social Skills		
Does not respond to spoken words/acts deaf	0.60	0.62
Poor eye/social contact	0.46	0.62
Lack of sustained interest	0.52	0.41
Factor 4: Aberrant Behavior		
Self-mutilation/pulling hair or ears/scratching	0.55	0.45
Aggressive behavior	0.98	0.94
Biting self and others	0.53	0.27
Factor 5: Respiratory Behaviors		
Bruxism	0.44	0.32
Breath holding	0.85	0.89
Hyperventilation	0.59	0.60
Air-saliva expulsion/drooling	0.44	0.47
Model Fit Indices		
CFI[a]	0.94	0.93
TLI[b]	0.93	0.92
RMSEA[c]	0.06	0.06

[a]CFI = comparative fit index.
[b]TLI = Tucker-Lewis index.
[c]RMSEA = root mean square error of approximation.

Reliability and Validity Analyses

To determine whether the subscales and total scale displayed internal consistency, Cronbach's alpha was calculated for each. Subscale scores were calculated as the mean for all items. A total R-MBA score was calculated as the sum of all the items in the subscales as well as the three additional items: seizures, truncal rocking, and a derived item to assess overall hand stereotypies (sum of stereotypic hand movements and mouthing hands/objects then rescaled to 5-point response option: 0 = combined score of 0; 1 = combined score of 1 or 2; 2 = combined score of 3 or 4; 3 = combined score of 5 or 6; 4 = combined score of 7 or 8).

Alphas indicated acceptable levels for the motor dysfunction (0.80 for development sample and 0.77 for the validation sample) and functional skills (0.80 and 0.78) subscales. However, alphas were lower for the remaining subscales, possibly due to the small number of items (3–4 items) on each scale: social skills (0.59 and 0.56), aberrant behavior (0.63 and 0.45), and respiratory behaviors (0.60 and 0.55). Cronbach's alpha was also calculated for the total scale, which included the three additional items that did not load onto the 5 factors. The alpha for the R-MBA scale was 0.77 for the development sample and 0.78 for the validation sample, indicating that together the 24 items had moderately high internal consistency as a measure of RTT severity.

Although the factors displayed clear internal clustering, there were some statistically significant correlations among the subscales (see Figure 1). The motor dysfunction, functional skills, and social skills subscales all had significant positive correlation with one another. Motor dysfunction and functional skills also had significant positive correlations with the respiratory behaviors subscale. Functional skills and aberrant behavior had a small negative correlation trend although it was not statistically significant at $p < 0.05$.

Table 6 presents the means and standard deviations of the MBA subscale scores by response option on the associated parent-reported items from the Interval History Form. For the majority of items, there is a positive relationship between the MBA subscales and parent report items. Individuals who were reported as having more severe symptoms by parents had significantly higher MBA subscales, supporting the validity of the subscales. As a second measure of validity, we compared the total R-MBA score to the Clinical

Table 5
Item Response Theory (IRT) Parameters of MBA Items by Sample

Factor/Item	Development Sample					Validation Sample				
	A	b1	b2	b3	b4	a	b1	b2	b3	b4
Motor Dysfunction										
Bradykinesia	1.80	0.96	1.37	1.92	2.76	1.55	1.21	1.71	2.40	3.67
Dystonia	2.29	0.18	0.65	1.63	2.43	1.42	0.26	0.98	2.09	3.07
Scoliosis	2.28	0.22	0.85	1.33	1.58	1.85	0.14	1.00	1.50	1.86
Hypertonia/rigidity	2.42	0.11	0.71	1.09	1.69	3.41	0.13	0.71	1.06	1.63
Functional Skills										
Hand clumsiness	3.66	−2.21	−0.82	−0.34	0.23	3.28	−2.38	−1.12	−0.48	0.23
Does not reach for objects or people	1.83	−0.70	−0.25	0.31	0.82	1.80	−0.72	−0.31	0.28	0.77
Motor skills	3.05	−2.55	−0.91	−0.26	0.85	3.24	−2.47	−1.11	−0.29	0.91
Speech disturbance	1.06	−4.39	−1.90	2.36	N/A	1.00	−6.40	−4.73	−1.87	2.84
Communication skills	0.77	−4.73	0.19	1.69	2.63	0.77	−5.44	−0.10	1.44	2.63
Feeding difficulties	1.20	−0.90	0.18	1.46	2.49	1.20	−0.73	0.24	1.56	2.67
Chewing difficulties	1.51	−1.42	0.14	1.09	1.78	1.22	−1.64	0.29	1.18	1.94
Social Skills										
Does not respond to spoken words/acts deaf	0.78	−2.45	−0.61	1.20	2.63	1.06	−1.55	−0.08	1.38	2.29
Poor eye/social contact	1.84	−0.61	0.61	1.65	2.90	1.59	−0.69	0.45	1.69	3.33
Lack of sustained interest	2.35	−0.65	0.30	1.19	2.25	1.47	−0.67	0.42	1.68	3.01
Aberrant Behavior										
Self-mutilation/pulling hair or ears/scratching	3.40	1.09	1.88	2.26	2.80	3.17	1.13	1.87	2.25	2.91
Aggressive behavior	1.59	1.59	2.53	3.27	3.87	1.04	2.34	3.89	4.59	5.49
Biting self and others	1.89	1.04	1.91	2.55	2.96	1.63	1.22	2.19	3.01	3.49
Respiratory Behaviors										
Bruxism	0.74	−0.62	1.24	2.20	3.37	0.50	−0.54	2.15	3.86	5.86
Breath holding	2.90	−0.47	0.48	1.17	1.68	3.85	−0.41	0.49	1.09	1.43
Hyperventilation	1.51	−0.01	0.94	1.82	2.51	1.35	0.04	0.92	1.79	2.44
Air-saliva expulsion/drooling	0.73	−1.19	0.64	2.00	2.96	0.77	−1.14	0.54	1.86	2.68

Note. a = slope for each item, b1 – b4 = threshold parameters for each response option.
[a]Not applicable: No individuals in the development sample were rated in this response option.

Severity Scale and found a very strong correlation ($r = 0.735$, $p < 0.001$), indicating that the R-MBA score is a good measure of overall severity in RTT. (Figure 2B). Using age as a continuous variable, the correlation with the R-MBA score was 0.37 ($p < 0.001$), indicating that R-MBA scores increase with age in RTT (Figure 2A). Subscale correlations were as follows: (1) motor dysfunction: $r = 0.68$, $p < 0.001$, (2) functional skills: $r = 0.17$, $p < 0.001$, (3) social skills: $r = 0.09$, $p = 0.004$, (4) aberrant behavior: $r = -0.03$, $p = 0.4$, and (5) respiratory behaviors: $r = 0.08$, $p = 0.007$. Finally, we conducted a preliminary genotype/phenotype analysis using the R-MBA as a measure of phenotypic severity, and clustering *MECP2* mutations into previously defined severity groups (as described in the methods). As shown in Figure 3, there was an overall difference between the three *MECP2* mutation groupings ($F(2,957) = 28.242$, $p < 0.0001$), with individuals with mild *MECP2*
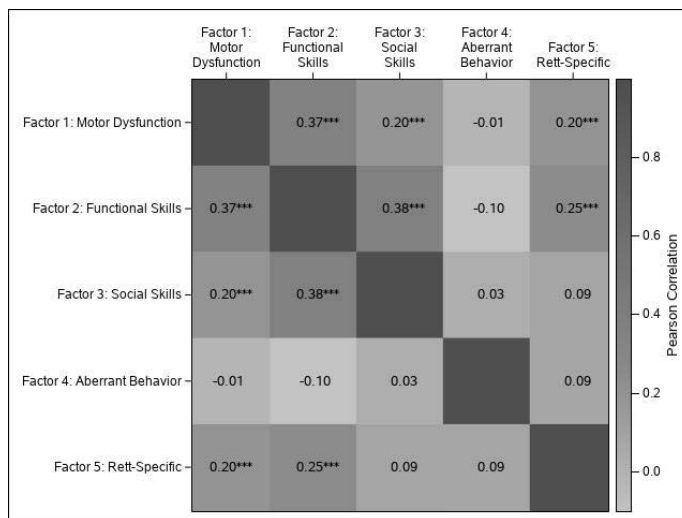
Figure 1. Heat map of correlations among the MBA subscales.

mutations showing significantly lower R-MBA scores than individuals with intermediate or severe *MECP2* mutations (post-hoc pairwise comparisons, Bonferroni corrected, $p < 0.001$ for each).

Discussion

Summary of Results

The goal of this study was to examine the psychometric properties of the MBA scale in order to assess its readiness for inclusion as an outcome measure for clinical trials in RTT. The formal evaluation of the MBA revealed that the originally conceived domains (behavioral/social, orofacial/respiratory, and motor/physical) did not hang together statistically. The five subscales derived from the development sample, although broadly associated with these domains, demonstrated a more fine-grained representation of the items. For example, the items from original behavioral/social domain were split among the functional skills, social skills, and aberrant behavior subscales. Given this, it is not surprising that there was moderate correlation among some of the subscales.

A number of the original items were excluded from the final scale due to measurement issues. These included irritability/crying, overactive or over passive, lack of toilet training, insensitivity to pain, ataxia/apraxia, myoclonus, dyskinesia, seizures, and vasomotor disturbances. Many of these items did not fit conceptually within the five subscales and thus had low factor loadings. Although there were a few that did fit conceptu-ally and had borderline factor loadings, when we added them into the model the fit statistics were not within the acceptable cutoffs. However, three items that did not factor onto one of the five subscales were retained as part of a total score given their clinical importance. In the end, we retained 24 items, 21 of which were on the five subscales. Thus, we recommend using the total R-MBA scale to capture all clinical aspects of RTT.

In the IRT analyses, the slopes demonstrated acceptable item discrimination. Items within the motor dysfunction subscale all had high slopes indicating their ability to differentiate individuals with high and low motor dysfunction. The respiratory behaviors subscale, though, had a range of slopes indicating that some items (e.g., Breath holding) were better able to discriminate than others (e.g., bruxism). As a practical implication, this may indicate that items with higher slopes are easier to score for clinicians than others. As expected, items that had higher slope values also had a narrower range across the threshold parameters. The threshold parameters also provide information about the response options and the level of ability needed to move from one level to the next. For example, the poor eye/social contact item has approximately a 1-unit difference between b_1 and b_2 indicating a higher amount of eye/social contact needed than the approximately 0.50-unit difference between b_1 and b_2 on the dystonia item.

When we examined the relationship between the R-MBA subscale scores, which are clinician-reported, with items from the parent-reported

Table 6

Mean (SD) MBA Subscale Scores by Parent-Reported Items From Interval History Form

Parent-Reported Items by MBA Subscale	Parent-Reported Response Option					
	1	2	3	4	5	
	Mean (SD) MBA score	Mean (SD) MBA score	Mean (SD) MBA score	Mean (SD) MBA score	Mean (SD) MBA score	p-value
Motor Dysfunction						
Stiff arms legs[1]	1.38 (1.51)	1.67 (2.20)	5.04 (4.27)	—	—	<.0001
Ability to sit[2]	2.32 (2.96)	4.88 (3.79)	6.55 (4.80)	—	—	<.0001
Ability to walk[3]	2.03 (2.90)	2.96 (2.98)	4.95 (4.53)	—	—	<.0001
Degree of scoliosis[4]	2.76 (3.46)	2.86 (2.90)	4.67 (3.46)	6.30 (4.40)	—	<.0001
Functional Skills						
Child feeding abilities[5]	12.19 (3.76)	13.74 (4.19)	14.79 (4.42)	16.84 (6.20)	20.29 (4.43)	<.0001
Communicates with body gestures[6]	11.83 (4.53)	12.64 (4.86)	15.92 (4.85)	—	—	<.0001
Communicates With Eye Gaze[6]	13.77 (5.03)	16.20 (5.12)	13.20 (4.04)	—	—	.0096
Social Skills						
Follow with gesture[7]	2.90 (2.12)	3.13 (2.00)	4.23 (2.20)	3.77 (2.14)	—	.0317
Follow without gesture[7]	2.36 (1.85)	4.08 (1.89)	4.14 (2.44)	3.56 (2.03)	—	.0062
Difficulty staying awake[1]	3.24 (2.23)	3.59 (2.15)	4.00 (2.13)	—	—	.2828
Aberrant Behavior						
Aggressive Child[1]	0.38 (0.84)	1.43 (1.60)	1.56 (1.88)	—	—	<.0001
Self-abusive behavior[1]	0.20 (0.52)	0.97 (0.99)	2.14 (2.03)	—	—	<.0001
Irritable/whiny/tantrums[1]	0.24 (0.53)	0.68 (1.06)	1.03 (1.82)	—	—	.0119
Screaming episodes[1]	0.38 (0.76)	0.75 (1.16)	1.04 (1.84)	—	—	.0214
Rapid mood changes[1]	0.48 (0.89)	0.57 (1.03)	0.97 (1.65)	—	—	.1434
Respiratory Behaviors						
Teeth grinding[1]	3.68 (3.11)	4.16 (2.83)	5.15 (2.83)	—	—	.0286
Stopped breathing while awake[1]	3.02 (2.44)	4.46 (3.04)	5.60 (2.72)	—	—	<.0001
Hyperventilation[1]	3.54 (3.24)	4.44 (2.51)	5.45 (2.65)	—	—	.0015
Drooling[1]	3.37 (2.53)	4.48 (3.17)	4.89 (2.78)	—	—	.0690

[1]Response options: 1 = Never, 2 = Occasionally, 3 = Frequently/Constantly.
[2]Response options: 1 = Without help, 2 = With some help, 3 = Cannot sit alone.
[3]Response options: 1 = Independently, 2 = Only with support, 3 = Cannot walk.
[4]Response options: 1 = None, 2 = Mild, 3 = Moderate, 4 = Severe.
[5]Response options: 1 = No difficulties, 2 = Occasional difficulties, 3 = Largest meal by mouth >30 minutes, 4 = Both mouth and gastronomy, 5 = Gastronomy only.
[6]Response options: 1 = Normally, 2 = With difficulty, 3 = Not at all.
[7]Response options: 1 = More than half the time, 2 = About half the time, 3 = Less than half the time, 4 = Never.

Interval History form, we found further evidence to support the construct validity of the measure. For example, parents who indicated that their child could walk independently had lower motor dysfunction scores (indicating less severity) than those whose children could not walk. There were, however, a handful of items on the Interval History form that were not as strongly associated with the R-MBA subscales. For example, difficulty staying awake and rapid mood changes were not
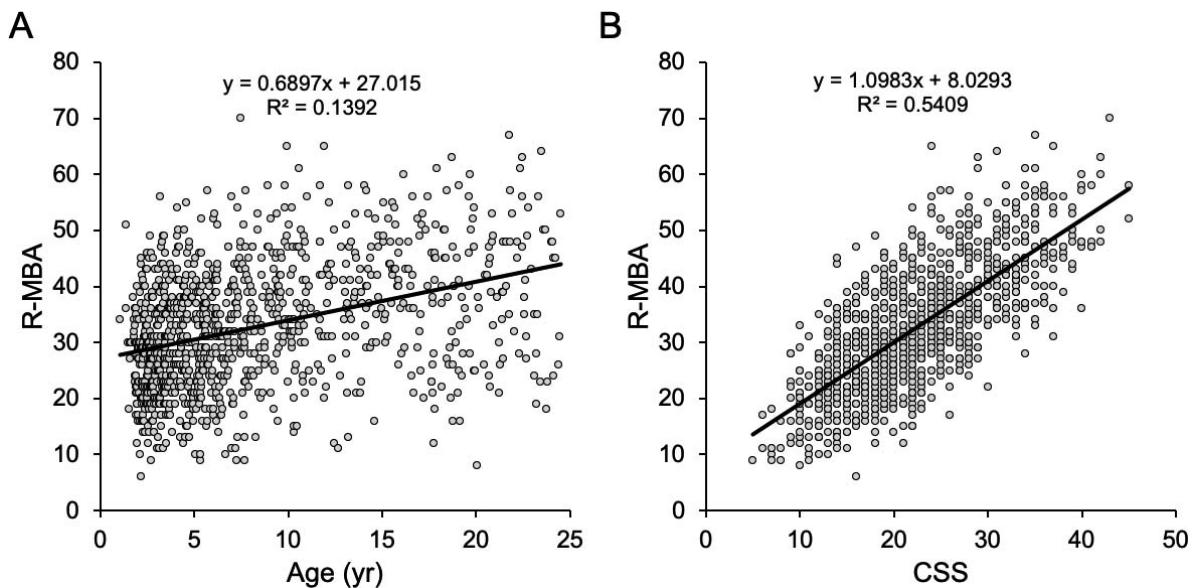
Figure 2. Correlations between the total MBA score and age (A) and the Clinical Severity Scale (B).

statistically related to the mean social skills score and the mean aberrant behavior score, respectively. This likely is due to the fact that the R-MBA does not contain similar items. Drooling, however, also was not statistically related to the mean respiratory behaviors scores despite the fact that the R-MBA has a similar item. This may be due to the differences in response options or because one is clinician-reported and the other parent-reported.

The overall R-MBA score shows age-related increases, which fits with the expected clinical pattern of increasing symptoms with age. Scores on four of the subscales (motor dysfunction, functional skills, social skills, and respiratory behaviors) increase with age, thus indicating increasing sever-

Figure 3. Genotype/phenotype analysis.

ity. However, the aberrant behavior subscale showed a non-significant decrease with age. The overall R-MBA score also showed a very strong correlation with overall clinical severity assessed using the RTT Clinical Severity Scale, and initial genotype/phenotype analysis showed the expected pattern of R-MBA scores compared to *MECP2* mutation severity (Neul et al., 2008). Both of these findings support the idea that the total R-MBA is a useful measure of overall severity in RTT.

Future Directions

These analyses were a critical step in determining whether the R-MBA could be used as an outcome measure in clinical trials in RTT. Regulatory agencies and professional organizations assert that careful attention to the performance of a potential outcome measure is needed prior to its use as an endpoint to assess treatment impact (Patrick et al., 2011; U.S. Department of Health and Human Services, 2009). Clinician-reported outcomes, like all other outcome assessments, must provide a robust assessment of the condition of interest (Walton et al., 2015). The MBA was originally conceptualized as a survey instrument that could provide details on the clinical impact of RTT. However, the scale evolved over the course of the first phase of the NHS; this paralleled the refinement of the clinical diagnosis of RTT over the same period of time. And although the R-MBA measures a variety of areas that are affected
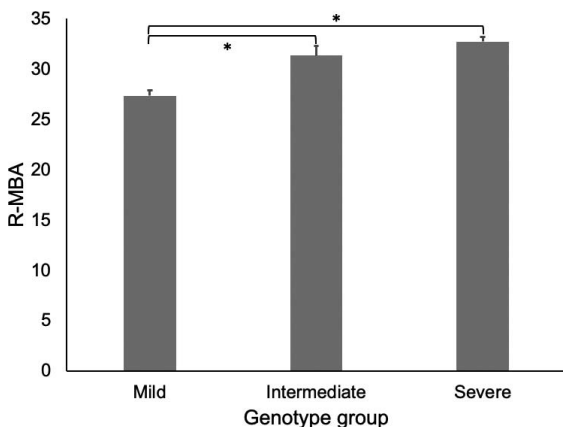
in individuals with RTT, it may not capture all the domains of importance. As a next step in instrument development, it will be important to revisit some of the clinical features that are present in RTT but not well-characterized in the revised version. For example, the R-MBA does not include items that assess sleep dysfunction, constipation, or vasomotor problems which are commonly present in RTT and are frequently troublesome to families. However, some of these items (sleep, constipation) are not readily assessed clinically and rely on parent-report. Another concern is that the subscales are not evenly weighted, with the subscales such as functional skills containing far more items than aberrant behavior. Finally, we included clinically relevant items such as hand stereotypies/hand mouthing that do not cluster into subscales. Thus, there is a need to return to the item generation phase in some of these domains in order to expand the R-MBA in subsequent versions. This could also improve the Cronbach alpha scores for some of the subscales with fewer items. As a result, the R-MBA may have stronger internal consistency, a balance of items across the subscales, and perhaps additional subscales of clinical importance in RTT.

Like many outcome measures in rare diseases (Basch & Bennett, 2014; Slade et al., 2018), the R-MBA could benefit from additional analyses to understand its utility better for clinical trials. Analysis of item performance longitudinally as symptoms progress would provide valuable information about the natural history of RTT and can be used to examine the timing of treatment and expected outcomes. Difference by mutation type would also be helpful. In addition, validation of the revised MBA subscales with other well-validated measures, such as the Vineland Scales of Adaptive Behavior which have been used in other studies of RTT, would provide further data on construct validity. However, this may prove challenging if existing measures do not accurately represent individuals with RTT or assess treatment targets (International Rare Diseases Consortium, 2016). Additional reliability assessments, such as test-retest and inter-rater reliability, should also be conducted.

Similarly, there is a need to develop a complementary parent-reported measure of clinical severity. Although we examined a handful of items in the Interval History form, no formal analysis has been conducted to assess its subscales or items. As a clinician-reported outcome measure, the R-MBA needs to be completed by a profes-sional with specific training in Rett syndrome and clinical assessments (Powers et al., 2017). A patient- or observer-reported outcome measure, however, collects data from the patient's or parent's perspective and is often focused on how a condition affects functioning in daily life (Benjamin et al., 2017; Walton et al., 2015). No matter the type of assessment used, the outcome measure needs to appropriately quantify the condition of interest and be able to detect treatment benefit. Ultimately, the R-MBA will need to demonstrate sensitivity to change and positive correlations with quality of life in order to be used in clinical trials. Many existing and upcoming clinical trials are already capturing data on the full set of items in the MBA, making these analyses possible in the near future.

Limitations

The results of this study should be interpreted with the following limitations. First, although the sample of participants was quite large, it was not socioeconomically diverse. Very few individuals from minority or other traditionally underserved populations were included. This may be because individuals had to be seen in-person at a RTT clinic in order to be enrolled in the NHS. Second, our validity analyses using the parent-reported Interval History form was based on items that we hypothesized to be related to the MBA subscales. Ideally, construct validity would be analyzed with an established measure or subscale. Finally, as mentioned as possible future directions, additional validity and sensitivity analyses were not conducted as part of this study.

Implications

Despite these limitations, the findings of this study show promise for the use of the R-MBA in upcoming clinical trials. The current subscales cover the majority of clinical criteria for diagnosing RTT, including loss of purposeful use of hands and spoken language, gait abnormalities, as well as a number of supportive criteria (e.g., scoliosis, sleep disturbances, bruxism) (Neul et al., 2010). With small modifications and improvements, the R-MBA could fill a significant gap in the field. Ensuring that the R-MBA, or other outcome measures in RTT, are aligned with a conceptual model and are psychometrically-sound will increase the ability to assess change for treatment targets.

References

Ahonniska-Assa, J., Polack, O., Saraf, E., Wine, J., Silberg, T., Nissenkorn, A., & Ben-Zeev, B. (2018). Assessing cognitive functioning in females with Rett syndrome by eye-tracking methodology. *European Journal of Paediatric Neurology*, *22*(1), 39–45. https://doi.org/10.1016/j.ejpn.2017.09.010

Amir, R. E., Van den Veyver, I. B., Wan, M., Tran, C. Q., Francke, U., & Zoghbi, H. Y. (1999). Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nature Genetics*, *23*(2), 185. https://doi.org/10.1038/13810

Barnes, K. V., Coughlin, F. R., O'Leary, H. M., Bruck, N., Bazin, G. A., Beinecke, E. B., Walco, A. C., Cantwell, N. G., & & Kaufmann, W. E. (2015). Anxiety-like behavior in Rett syndrome: Characteristics and assessment by anxiety scales. *Journal of Neurodevelopmental Disorders*, *7*(1), 30. https://doi.org/10.1186/s11689-015-9127-4

Basch, E., & Bennett, A. V. (2014). Patient-reported outcomes in clinical trials of rare diseases. *Journal of General Internal Medicine*, *29*(3), 801–803. https://doi.org/10.1007/s11606-014-2892-z

Benjamin, K., Vernon, M. K., Patrick, D. L., Perfetto, E., Nestler-Parr, S., & Burke, L. (2017). Patient-reported outcome and observer-reported outcome assessment in rare disease clinical trials: An ISPOR COA emerging good practices task force report. *Value in Health*, *20*(7), 838–855. https://doi.org/10.1016/j.jval.2017.05.015

Berry-Kravis, E., Des Portes, V., Hagerman, R., Jacquemont, S., Charles, P., Visootsak, J., Brinkman, M., Rerat, K., Koumaras, B., Zhu, L., Barth, G. M., Jaecklin, G. A., von Raison, F. (2016). Mavoglurant in fragile X syndrome: Results of two randomized, double-blind, placebo-controlled trials. *Science Translational Medicine*, *8*(321), 321ra5. https://doi.org/10.1126/scitranslmed.aab4109

Berry-Kravis, E., Hessl, D., Abbeduto, L., Reiss, A. L., Beckel-Mitchener, A., Urv, T. K., & the Outcome Measures Working Group (2013). Outcome measures for clinical trials in fragile X syndrome. *Journal of Developmental and Behavioral Pediatrics*, *34*(7), 508–522. https://doi.org/10.1097/DBP.0b013e31829d1f20

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). *IRTPRO for Windows* [Computer software]. Scientific Software International.

Clarkson, T., LeBlanc, J., DeGregorio, G., Vogel-Farley, V., Barnes, K., Kaufmann, W. E., & Nelson, C. A. (2017). Adapting the Mullen Scales of Early Learning for a standardized measure of development in children with Rett syndrome. *Intellectual and Developmental Disabilities*, *55*(6), 419–431. https://doi.org/10.1352/1934-9556-55.6.419

Downs, J., Leonard, H., Jacoby, P., Brisco, L., Baikie, G., & Hill, K. (2015). Rett syndrome: Establishing a novel outcome measure for walking activity in an era of clinical trials for rare disorders. *Disability and Rehabilitation*, *37*(21), 1992–1996. https://doi.org/10.3109/09638288.2014.993436

Djukic, A., Holtzer, R., Shinnar, S., Muzumdar, H., Rose, S. A., Mowrey, W., Galanopoulou, A. S., Shinnar, R., Jankowski, J. J., Feldman, J. F., Pillai, S., & Moshe, S. L. (2016). Pharmacologic treatment of Rett syndrome with Glatiramer acetate. *Pediatric Neurology*, *61*, 51–57. https://doi.org/10.1016/j.pediatrneurol.2016.05.010

FitzGerald, P. M., Jankovic, J., & Percy, A. K. (1990). Rett syndrome and associated movement disorders. *Movement Disorders: Official Journal of the Movement Disorder Society*, *5*(3), 195–202. https://doi.org/10.1002/mds.870050303

Gadalla, K. K., Bailey, M. E., & Cobb, S. R. (2011). MeCP2 and Rett syndrome: Reversibility and potential avenues for therapy. *Biochemical Journal*, *439*(1), 1–14. https://doi.org/10.1042/BJ20110648

Glaze, D. G., Neul, J. L., Kaufmann, W. E., Berry-Kravis, E., Condon, S., Stoms, G., Oosterholt, S., Della Pasqua, O., Glass, L., Jones, N. E., & Percy, A. K. (2019). Double-blind, randomized, placebo-controlled study of trofinetide in pediatric Rett syndrome. *Neurology*, *92*(16), e1912–e1925. https://doi.org/10.1212/WNL.0000000000007316

Glaze, D. G., Neul, J. L., Percy, A., Feyma, T., Beisang, A., Yaroshinsky, A., Stoms, G., Zuchero, D., Horrigan, J., Glass, L., & Jones, N. E. (2017). A double-blind, randomized, placebo-controlled clinical study of trofinetide in the treatment of Rett syndrome. *Pediatric Neurology*, *76*, 37–46. https://doi.org/10.1016/j.pediatrneurol.2017.07.002

Gold, W. A., Krishnarajy, R., Ellaway, C., & Christodoulou, J. (2017). Rett syndrome: A genetic update and clinical review focusing on comorbidities. *ACS Chemical Neuroscience*, *9*(2), 167–176. https://doi.org/10.1021/acschemneuro.7b00346

Guy, J., Gan, J., Selfridge, J., Cobb, S., & Bird, A. (2007). Reversal of neurological defects in a mouse model of Rett syndrome. *Science*, *315*(5815), 1143–1147. https://doi.org/10.1126/science.1138389

Hagberg, B., Aicardi, J., Dias, K., & Ramos, O. (1983). A progressive syndrome of autism, dementia, ataxia, and loss of purposeful hand use in girls: Rett's syndrome: Report of 35 cases. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, *14*(4), 471–479. https://doi.org/10.1002/ana.410140412

International Rare Diseases Consortium (IRDiRC). *Patient-centered outcome measures initiatives in the field of rare diseases*. February 2016. http://www.irdirc.org/wp-content/uploads/2017/12/PCOM_Post-Workshop_Report_Final.pdf.

Jeste, S. S., & Geschwind, D. H. (2016). Clinical trials for neurodevelopmental disorders: At a therapeutic frontier. *Science Translational Medicine*, *8*(321), 321fs1. https://doi.org/10.1126/scitranslmed.aad9874

Katz, D. M., Bird, A., Coenraads, M., Gray, S. J., Menon, D. U., Philpot, B. D., & Tarquinio, D. C. (2016). Rett syndrome: Crossing the threshold to clinical translation. *Trends in Neurosciences*, *39*(2), 100–113. https://doi.org/10.1016/j.tins.2015.12.008

Lane, J. B., Lee, H. S., Smith, L. W., Cheng, P., Percy, A. K., Glaze, D. G., Neul, K. J., Motil, J. O., Barrish, S. A., Skinner, F., Annese, F., McNair, L., Graham, J., Khwaja, O., Barnes, K., & Krischer, J. P. (2011). Clinical severity and quality of life in children and adolescents with Rett syndrome. *Neurology*, *77*(20), 1812–1818. https://doi.org/10.1212/WNL.0b013e3182377dd2

Leonard, H., Cobb, S., & Downs, J. (2017). Clinical and biological progress over 50 years in Rett syndrome. *Nature Reviews Neurology*, *13*(1), 37. https://doi.org/10.1038/nrneurol.2016.186

Moretti, P., & Zoghbi, H. Y. (2006). MeCP2 dysfunction in Rett syndrome and related disorders. *Current Opinion in Genetics & Development*, *16*(3), 276–281. https://doi.org/10.1016/j.gde.2006.04.009

Mount, R. H., Charman, T., Hastings, R. P., Reilly, S., & Cass, H. (2002). The Rett Syndrome Behaviour Questionnaire (RSBQ): Refining the behavioural phenotype of Rett syndrome. *Journal of Child Psychology and Psychiatry*, *43*(8), 1099–1110. https://doi.org/10.1111/1469-7610.00236

Muthén, L. K. and Muthén, B. O. (1998–2017). *Mplus user's guide* (8th ed.). Muthén & Muthén.

Neul, J. L., Fang, P., Barrish, J., Lane, J., Caeg, E. B., Smith, E. O., Zoghbi, H., Percy, A. & Glaze, D. G. (2008). Specific mutations in methyl-CpG-binding protein 2 confer different severity in Rett syndrome. *Neurology*, *70*(16), 1313–1321. https://doi.org/10.1212/01.wnl.0000291011.54508.aa

Neul, J. L., Glaze, D. G., Percy, A. K., Feyma, T., Beisang, A., Dinh, T., Suter, B., Anagnostou, E., Snape, M., Horrigan, J. & Jones, N. E. (2015). Improving treatment trial outcomes for Rett syndrome: The development of Rett-specific anchors for the Clinical Global Impression Scale. *Journal of Child Neurology*, *30*(13), 1743–1748. https://doi.org/10.1177/0883073815579707

Neul, J. L., Kaufmann, W. E., Glaze, D. G., Christodoulou, J., Clarke, A. J., Bahi-Buisson, N., Leonard, H., Bailey, M. E. S., Schanen, N. C., Zappella, M., Renieri, A., Huppke, P., & Percy, A. K. (2010). Rett syndrome: Revised diagnostic criteria and nomenclature. *Annals of Neurology*, *68*(6), 944–950. https://doi.org/10.1002/ana.22124

Neul, J. L., Lane, J. B., Lee, H.-S., Geerts, S., Barrish, J. O., Annese, F., Baggett, L. M., Barnes, K., Skinner, S. A., Motil, K. J., Glaze, D. J., Kaufmann, W. E., & Percy, A. K. (2014). Developmental delay in Rett syndrome: Data from the natural history study. *Journal of Neurodevelopmental Disorders*, *6*(1), 20. https://doi.org/10.1186/1866-1955-6-20

Patrick, D. L., Burke, L. B., Gwaltney, C. J., Leidy, N. K., Martin, M. L., Molsen, E., & Ring, L. (2011). Content validity−establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO Good Research Practices Task Force report: Part 2−assessing respondent understanding. *Value in Health*, *14*(8), 978–988. https://doi.org/10.1016/j.jval.2011.06.013

Powers, J. H., Patrick, D. L., Walton, M. K., Marquis, P., Cano, S., Hobart, J., Isaac, M.,

Vamvakas, S. Slagle, A., Molsen, E., & Burke, L. B. (2017). Clinician-reported outcome assessments of treatment benefit: Report of the ISPOR Clinical Outcome Assessment Emerging Good Practices Task Force. *Value in Health*, *20*(1), 2–14. https://doi.org/10.1016/j.jval.2016.11.005

Sansone, S. M., Schneider, A., Bickel, E., Berry-Kravis, E., Prescott, C., & Hessl, D. (2014). Improving IQ measurement in intellectual disabilities using true deviation from population norms. *Journal of Neurodevelopmental Disorders*, *6*(1), 16. https://doi.org/10.1186/1866-1955-6-16

SAS Institute, Inc. (2012-2017). *SAS version 9.4* [Computer software]. Cary, NC: SAS Institute, Inc.

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of Educational Research*, *99*(6), 323–338. https://doi.org/10.3200/JOER.99.6.323-338

Schultz, R. J., & Glaze, C. D. G. (2017). *Rett syndrome: Genetics, clinical features, and diagnosis*. UpToDate. https://www.uptodate.com/contents/rett-syndrome-genetics-clinical-features-and-diagnosis

Schwartzman, J. S., Velloso, R. D. L., D'Antino, M. E. F., & Santos, S. (2015). The eye-tracking of social stimuli in patients with Rett syndrome and autism spectrum disorders: A pilot study. *Arquivos de Neuro-Psiquiatria*, *73*(5), 402–407. https://doi.org/10.1590/0004-282X20150033

Slade, A., Isa, F., Kyte, D., Pankhurst, T., Kerecuk, L., Ferguson, J., Lipkin, G., & Calvert, M. (2018). Patient reported outcome measures in rare diseases: A narrative review. *Orphanet Journal of Rare Diseases*, *13*(1), 61. https://doi.org/10.1186/s13023-018-0810-x

U. S. Department of Health and Human Services, Food and Drug Administration. *Guidance for industry*: *Patient-reported outcomes measures: Use in medical product development to support labeling claims*. December 2009. http://www.fda.gov/downloads/Drugs/GuidanceCompliance RegulatoryInformation/Guidances/UCM193282.pdf

Walton, M. K., Powers, J. H., Hobart, J., Patrick, D., Marquis, P., Vamvakas, S., Isaac, M., Molsen, E., Cano, S., & Burke, L. B. (2015). Clinical outcome assessments: conceptual foundation–Report of the ISPOR clinical outcomes assessment–emerging good practices for outcomes research task force. *Value in Health*, *18*(6), 741–752. https://doi.org/10.1016/j.jval.2015.08.006

Wang, H., Pati, S., Pozzo-Miller, L., & Doering, L. C. (2015). Targeted pharmacological treatment of autism spectrum disorders: fragile X and Rett syndromes. *Frontiers in Cellular Neuroscience*, *9*, 55. https://doi.org/10.3389/fncel.2015.00055

Authors:
Melissa Raspa, Carla M. Bann, and **Angela Gwaltney,** RTI International; **Timothy A. Benke,** University of Colorado School of Medicine; **Cary Fu,** Vanderbilt Kennedy Center; **Daniel G. Glaze,** Baylor College of Medicine; **Richard Haas,** University of California San Diego; **Peter Heydemann,** Rush University Medical Center; **Mary Jones,** Benioff Children's Hospital; **Walter E. Kaufmann,** Greenwood Genetic Center; **David Lieberman** and **Eric Marsh,** Children's Hospital Boston; **Sarika Peters,** Vanderbilt Kennedy Center, **Robin Ryther,** Washington University School of Medicine; **Shannon Standridge,** Cincinnati Children's Hospital; **Steven A. Skinner,** Greenwood Genetic Center; **Alan K. Percy,** University of Alabama at Birmingham; and **Jeffrey L. Neul,** Vanderbilt Kennedy Center.

Correspondence concerning this article should be addressed to Melissa Raspa, RTI International, 3040 East Cornwallis Road, P. O. Box 12194, Research Triangle Park, NC 27709 (e-mail: mraspa@rti.org) or Jeffrey Neul, Vanderbilt Kennedy Center, Vandergbilt University Medical Center, Nashville, TN (e-mail: Jeffrey.l.neul@vumc.org).