Title: Longitudinal Change in Restricted and Repetitive Behaviors from 8-36 months

Authors: Sifre, R., Berry, D., Wolff, J.J., & Elison, J.T.

Corresponding author:
Jed Elison, 51 East River Parkway, Minneapolis, MN, 55455
612-301-6272
jtelison@umn.edu

Robin Sifre, 51 East River Parkway, Minneapolis, MN, 55455
612-624-6373
Sifre002@umn.edu

Abstract (350 word max)
Background. Restricted and repetitive behaviors (RRBs) are core features of autism spectrum disorder (ASD) and one of the earliest behavioral signs of ASD. However, RRBs are also present in typically developing (TD) infants, toddlers, and preschool aged children. Past work suggests that examining change in these behaviors over time is essential to distinguish between normative manifestations of these behaviors and behaviors that denote risk for a neurodevelopmental disorder. One challenge in examining changes in these behaviors over time is that most measures of RRBs have not established longitudinal measurement invariance. The aims of this study were to 1) establish measurement invariance in the Repetitive Behavior Scales for Early Childhood (RBS-EC), a parent-report questionnaire of RRBs, and 2) model developmental change in RRBs from 8- to 36-months.

Methods. We collected RBS-EC responses from parents of TD infants (n=180) from 8- to 36-months (n=606 responses, with participants contributing an average of 3 time points). We leverage a novel methodological approach to measurement invariance testing (Bauer, 2017), moderated nonlinear factor analysis (MNLFA), to determine whether the RBS-EC was invariant across age and sex. We then generated adjusted factor score estimates for each subscale of the RBS-EC (repetitive motor, self-directed, and higher-order behaviors), and used linear mixed effects models to estimate between- and within-person changes in the RBS-EC over time.

Results. The RBS-EC showed some non-invariance as a function of age. We were able to adjust for this non-invariance in order to more accurately model changes in the RBS-EC over time. Repetitive motor and self-directed behaviors showed a linear decline from 8- to 36-months, while higher-order behaviors showed a quadratic trajectory such that they began to decline later in development at around 18 months. Using adjusted factor scores as opposed to unadjusted raw mean scores provided a number of benefits, including increased within-person variability and precision.

Conclusions. The RBS-EC is sensitive enough to measure the presence of RRBs in a TD sample, as well as their decline with age. Using factor score estimates of each subscale adjusted for non-invariance allowed us to more precisely estimate change in these behaviors over time.

List of abbreviations

ASD	Autism spectrum disorders
DIF	Differential item functioning
HR-ASD	Individuals at high-familial risk for ASD who later received a diagnosis
HR-Neg	Individuals at high-familial risk for ASD who do not receive a diagnosis
MGA	Multiple groups analysis
MI	Measurement Invariance
MIMIC	Multiple Indicators, Multiple Causes
MNLFA	Moderated nonlinear factor analysis
RBS-EC	Repetitive Behavior Scales for Early Childhood
RBQ-2	Repetitive Behaviors Questionnaire-2
RRBs	Restricted and Repetitive Behaviors
TD	Typically developing

Introduction

Restricted and Repetitive Behaviors

The DSM-5 defines autism as a disorder marked by social deficits, as well "restricted, repetitive patterns of behavior, interests, or activities" (American Psychiatric Association, 2013). These restricted and repetitive behaviors (RRBs) include repetitive motor movements, ritualistic behaviors, repetitive self-injury, inflexibility, and circumscribed and intense interests. RRBs are among the earliest detectable behavioral markers of ASD (Ozonoff et al., 2008), with evidence from direct observation indicating that elevated frequencies of RRBs in children with ASD can be identified as early as 12 months (Elison et al., 2014; Harrop, McConachie, Emsley, Leadbitter, & Green, 2014; Ozonoff et al., 2008). Parent report of RRBs, as measured by the Repetitive Behavior Scale-Revised (RBS-R; Bodfish, Symons, Parker, & Lewis, 2000), indicate that 12-month-old infants at high-familial risk for ASD who later received a diagnosis (HR-ASD) have elevated repetitive behaviors relative to high-risk infants who do not receive an ASD diagnosis (HR-Neg) and low-risk children (Wolff et al., 2014), raising the possibility of using parent reports of RRBs as an inexpensive early screening tool for ASD.

One challenge with characterizing RRBs, and especially early in development, is that they are not specific to ASD. Rather, they manifest in samples of typically developing children (Arnott et al., 2010; DeLoache, Simcock, & Macari, 2007; Evans et al., 1997; Evans, Uljarević, Lusk, Loth, & Frazier, 2017; Leekam et al., 2007; Thelen, 1979, 1981; Wolff, Boyd, & Elison, 2016) and across a range of neurodevelopmental disorders and monogenic syndromes (Evans, Gray, & Leckman, 1998; Hoch et al., 2016; Lewis & Bodfish, 1998; Moss, Oliver, Arron, Burbidge, & Berg, 2009; Wolff et al., 2012). Because these behaviors variably manifest across age and across the typical-to-atypical continuum (Cicchetti, 1993), quantifying variability in selected topographies requires consideration of normative patterns of development and establishing psychometric integrity on samples that include children who are typically developing. While some behaviors may be specific to ASD at certain ages and levels of functioning (e.g., lateral glances and/or unusual visual inspection in a 10-year-old), most topographies of repetitive behavior are observed in typically developing children at some point over the course of development. Therefore, establishing the psychometric integrity of measures that can capture meaningful variability across the typical-to-atypical continuum is essential to advance our understanding of the constellation of features that defines ASD (Constantino, 2011, 2018).

Further, to distinguish between normative RRBs and those that are predictive of risk for an emerging neurodevelopmental disorder, it is critical to look at change in RRBs over time. One study examining longitudinal change in RRBs in TD children as measured through parent report (the Repetitive Behaviors Questionnaire-2; RBQ-2; Uljarević et al., 2017) found that lower-order behaviors (e.g., motor stereotypies) decreased from 15- to 77 months, whereas higher-order behaviors (e.g., insistence on sameness) peaked around 26 months of age before declining. These

findings suggest that while the presence of RRBs early in development may not be atypical, their persistence over time may be.

In order to effectively measure changes in RRBs across development, it is imperative that investigators are confident that their measures reflect the same underlying construct whether parents are reporting on a 9-month-old or a 36-month-old. One previous study examining change in two RRB sub-types showed that the underlying two-factor structure was stable across three time points (15, 26, and 77 months), suggesting that the underlying structure of the RBQ-2 was stable across development (Uljarević et al., 2017). Referred to as "configural invariance" in the psychometrics literature, establishing a common factor structure over time is a crucial first step to assuring the developmental interpretability of a measure. That is, if the nature of the underlying construct changes qualitatively over time, then quantitative estimates of substantive developmental change (i.e., apples to apples) become inextricably confounded with changes in the meaning of the measure (i.e., apples to oranges). However, as is discussed under the broader heading of 'measurement invariance' (MI) testing, configural invariance is a necessary—though not sufficient—step to assuring meaningful comparisons across groups or within-individuals over time.

What is measurement invariance and why should we care?

MI for a scale exists when "a scale or construct provides the same results across several different samples or populations" (APA, 2014, p. 211; see also Putnick & Bornstein, 2016). In other words, a scale is invariant if the distribution of responses obtained from a group of individuals depends only on their responses which the measurement is intended to reflect, and not on other demographic characteristics such as age or diagnostic group (Mellenbergh, 1989). Without MI, one cannot validly compare scores on a given scale between groups or within

individuals over time. Differential item functioning (DIF) between male and female respondents on measures of depression is a prototypic example. Due to social norms, girls may tend to endorse the item "cries easily" more often than do boys (Steinberg & Thissen, 2006), leading to inflated estimated depression scores for girls because of item bias. To accurately compare rates of depression between boys and girls, one must account for items that function differently due to sex as opposed to true differences in the underlying construct of interest (depression).

Similarly, items on any questionnaire measuring RRBs may function differently at different ages. Thus, to model substantively meaningful change in the underlying constructs we must first evaluate whether the measure is invariant across time. Though this issue has not yet been explicitly discussed in the context of RRBs, researchers have noted how challenging it is to find a measure that is sensitive to both age and cognitive development and differences across diagnostic groups (Honey, Rodgers, & McConachie, 2012). Indeed, the notable variability in factor solutions that have been proposed in the literature is likely due to both 1) substantively meaningful differences in RRBs across the populations and ages studied, and 2) measurement sensitivity differences due to age and diagnostic status. In order to distinguish between these two sources of variability, it is critical that the field test MI and correct for sources of non-invariance.

Factor analysis commonly considers three degrees of MI. A measure has *configural invariance* when it has the same factor structure across time (e.g., items load on to the same latent factors across time) (Muthén & Asparouhov, 2018). Researchers must also test for the possibility that a latent factor *means* something slightly different across groups or ages. This is typically referred to as *metric* non-invariance, and indicates that there are age differences in the factor loadings (or that the items are not reflecting the same construct over time). For example, in typically developing infants the relation between a latent construct like *repetitive motor*

behaviors and the item "repeatedly mouths objects" (i.e., the factor loading) may differ between young infants and four-year-olds because repetitive mouthing may be more meaningful for the latent construct in infancy, when mouthing is developmentally normative. By failing to account explicitly for these difference in the measurement model, we would be essentially comparing apples and oranges.

Assuming invariant factor loadings across groups (metric invariance), it remains possible that, despite showing identical levels of *latent* repetitive behaviors in infancy and at age four, the same "repetitive mouthing" item may be rated as higher in infancy, on average, because of normative developmental differences in object exploration. This is typically referred to as *scalar* non-invariance. This would mean that—without explicit adjustment in the measurement model—this difference would sneak into our estimate of the construct mean and lead to the spurious conclusion that repetitive behavior decreases between infancy and early childhood. Adjusting for these biases is crucial for any meaningful interpretation of development.

Traditionally, applied researchers have leveraged two approaches for measuring and adjusting MI: Multiple Groups analysis (MGA; Jöreskog, 1971) and Multiple Indicators, Multiple Causes (MIMIC) modeling (Jöreskog & Goldberger, 1975). Each has specific strengths and weaknesses. MGA is perhaps the most commonly used approach in MI testing and is especially useful when one is concerned about MI across discrete groupings of individuals (e.g., sex, race). Briefly, MGA works by fitting taxonomies of confirmatory factor analytic models (CFAs) jointly to multiple covariance matrices—one for each group—and systematically testing the extent to which the parameters of interest (e.g., factor loadings, indicator intercepts, indicator residual (co) variances, factor (co)variances) can be constrained to equality across groups without significantly undermining overall model fit (i.e., aggregate model fit across groups). To

the extent to which the parameters can be constrained to equality, the latent variables are said to be invariant. In cases in which some (but not all) of the parameters can be constrained to equality, one invokes 'partial' MI (Byrne, Muthèn, & Shavelson, 1989). Although there is debate about just *how* non-invariant groups can be before the substantive meaning of the construct differs across groups (see Byrne & Watkins, 2003), the key strength of testing MI is that it allows one to quarantine these differences to the measurement model. This approach essentially removes sources of potential bias from the latent construct and subsequent substantive analyses.

A core advantage to MGA is that it allows one to model heterogeneity in all of the parameters of typical psychometric interest (e.g., factor loadings; item intercepts; factor (co)variances, item residual (co)variances). Specifically, MGA allows one to test and adjust for the possibility of both metric and scalar invariance. MGA can be a powerful tool toward this end. However, it also has some non-trivial weaknesses. MGA becomes intractable quickly, as the number of discrete groupings extends beyond a few categories. Also, in order to apply MGA to continuous moderators, such as age or income, one has to discretize continuously distributed scales in typically arbitrary ways.

MIMIC models take a slightly different approach to MI testing that can help to address some of these weaknesses. Specifically, rather than fitting simultaneous measurement models to separate covariance matrices and testing systematic equality constraints across the discrete groups, MIMIC models address MI in a manner akin to regression adjustment (Error! Reference source not found.). Here, an individual's level on a given indicator (e.g., repeated mouthing of object) is considered to be due to the underlying latent construct (i.e., repetitive behavior) (Path a), as well as one or more continuous (e.g., age) or categorical (e.g., sex) moderators (Path b). Similar to multiple regression, by simultaneously regressing the indicator and the latent construct

on the moderator(s), the factor loading to the indicator (Path b) would represent the unique relation between the construct and the indictor that remains—*after adjusting for the moderator* (path c). It could also be considered a variant of a mediational model, such that the moderator is thought to have a direct effect on the indicator, as well as indirect effect on the indicator by virtue of its effect on the latent construct (paths a*b). In principle, as long as the latent variable model is identified, one could include any number or predictors, along with any potentially meaningful interactions between them—an advantage over MGA. A notable downside to traditional MIMIC models, however, is that they typically adjust only for scalar invariance. Modeling systematic difference in any of the other parameters of typical interest (i.e., factor loadings, factor and/or residual measurement (co)variances) is a far more involved endeavor (see Bauer, 2017).

In sum, it is clear that testing and adjusting for MI is critical to establishing valid inferences about group and developmental differences. However, traditional approaches to MI testing have some weaknesses. Indeed, these weaknesses are particularly problematic for longitudinal studies, in which measurement non-invariance may exist simultaneously between groups (e.g., diagnostic group, sex) and time (Curran et al., 2014).

Fortunately, the last several years have seen a wealth of psychometric innovation in the detection and adjustment of MI (e.g., Approximate Measurement Invariance and Alignment (Asparouhov & Muthén, 2014; Muthén & Asparouhov, 2014; Muthén & Asparouhov, 2018); Random effects (Muthén & Asparouhov, 2018); see Davidov, Muthén, & Schmidt (2018) for review). In the present study, we leverage a novel methodological approach to MI testing developed by Curran and Bauer (Bauer, 2017; Curran et al., 2014) that is particularly well-suited for accelerated longitudinal study designs—moderated nonlinear factor analysis (MNLFA).

Conceptually, MNLFA combines the best aspects of MGA and MIMIC into a single approach. Like MGA, MNLFA allows one to adjust for non-invariance across all parameters of typical psychometric interest. However, like MIMIC models, MNLFA allows one to readily extend these tests to multiple, simultaneous moderators—either continuous or categorical.

The present study

The present study seeks to characterize longitudinal change in RRBs in typically developing toddlers from 8-to 36-months. Distinguishing between meaningful change and measurement bias is a critical first step before modelling change, especially given the challenges associated with distinguishing between typical and atypical early RRBs. In light of these challenges, we (a) leveraged recent advances in latent variable modeling to establish a longitudinally invariant measure of RRB, and (b) used these measures to model normative and individual difference in RRB across infancy and early childhood. By modelling factor score estimates derived from MNLFA and accounting for measurement bias, the observed longitudinal trajectories will reflect what is meaningful change in behavior, as opposed to change in how the instrument measures behaviors.

Methods

RBS-EC responses were pooled from two longitudinal studies on social development. Criteria for study participation were identical for both studies. Infants were required to 1) have no significant medical, genetic, or neurological conditions, 2) have a birthweight > 2,000 grams, 3) a gestational age ≥ 37 weeks, 4) have no first-degree relatives with intellectual disability, psychosis and/or schizophrenia, bipolar disorder, or autism spectrum disorder (ASD), and 5) a caregiver able to communicate in English at a level to provide informed consent. For each study,

parents of infants and toddlers recruited from the University of Minnesota Institute of Child Development participant registry were invited to participate in a study about their child's development. The registry largely reflects the racial/ethnic proportions of the broader Minneapolis- St. Paul metropolitan area but under-represents the socio- economic diversity of this region. Parents of all participants provided informed consent and permission for their child to participate in this research study. All studies involved online questionnaires, as well as in-person visits to the lab.

In total, 612 RBS-EC questionnaires were collected from 181 participants. After excluding questionnaires with 50% or more missing data (n=3), and samples taken outside of our age range (n=3) our final sample comprised 606 RBS-EC questionnaires from 180 participants (88 male). Descriptive information on the final sample can be found in **Table 1**.

Repetitive Behavior Scale for Early Childhood

The Repetitive Behavior Scale for Early Childhood (RBS-EC ; Wolff et al., 2016) is a 34-item parent-report questionnaire that is a downward-extension of the Repetitive Behavior Scale–Revised (RBS-R; Bodfish et al., 2000), with good-to-excellent psychometric properties and evidence of validity and reliability (based on an independent sample of toddlers). The questionnaire is intended to capture normative variation in young children that spans across the typical-to-atypical continuum, and has been used to detect difference in RRBs as a function of birthweight percentile (Sifre et al., 2018), and to detect associations between RRBs and dysregulation and internalizing symptoms (Lasch, Wolff, & Elison, 2019) in toddlers. For distributions of RBS-EC scores in the present sample, see **Figure S1** in the Online Supplement**.** Each item contributes to 2 measures: items endorsed (binary) and frequency score (0-behavior

does not occur, 1-behavior occurs about weekly or less, 2-behavior occurs several times a week, 3-behavior occurs about daily, and 4-behavior occurs many times a day). These measures can be summed into an overall composite measure (scored 0-34) or disaggregated into 4 psychometrically validated subscale scores: Repetitive Motor (scored 0-9), Ritual and Routine (scored 0-10), Restricted Behavior (scored 0-8), and Self-directed behavior (scored 0-7). See http://www.cehd.umn.edu/edpsych/research/ resources/rbs-ec/ for access to the instrument.

Analytic Plan

Factor structure. Visual inspection of distributions of item frequency responses (ranging from 0-4) indicated that responses were highly skewed—typically captured by 2 values of 0-4 scale. We therefore collapsed across empty cells by converting each item response to a binary scale based on median splits (scores ≤ the median were coded as 0, scores > the median were coded as 1). Descriptive statistics on item responses can be found in Supporting information (**Table S1**).

To establish configural longitudinal invariance, we first fit CFAs to two discretized age bins split at 17 months (Mean age of younger group=12.1 months (2.4), mean age of older group=22.6 months (4.9)) (i.e. Bauer, 2017) to confirm that there were separable unitary factors that items from the RBS-EC loaded on to. Repetitive behavior is not typically characterized as a singular, encapsulated construct. Rather, any meaningful characterization of repetitive behavior acknowledges that there are at least two factors that define/represent this category of behavior (e.g., lower-order and higher-order RRBs, see Turner, 1997; 1999) and depending on the sample, measurement/assessment, and analytic decisions, there may be more (Bodfish et al., 2000; Lam, Bodfish, & Piven, 2008; Mandy & Skuse, 2008; Mirenda et al., 2010; Mooney, Gray, Tonge, Sweeney, & Taffe, 2009; Szatmari et al., 2006). Informed by prior validation work with a sample

of 17- to 27-month-olds (Wolff, Boyd, & Elison, 2016), we fitted a model comprising four correlated latent factors tapping: Repetitive Motor, Ritual and Routines, Restricted Interests, and Self-Directed Behavior. Items on the Repetitive Motor latent factor measure repetitive and non-social motor stereotypies; items on the Self-Directed latent factor measure repeated movement directed towards the body, including self-injury and proto-self-injury; items on the Ritual and Routine measure resistance to change and insistence on sameness; and items on the Restricted Interests measure intense or unusual interests or activities (See Online Supplement for a complete list of items). Although model fit was reasonable for the 4-factor model for both age groups (Group 1: X_2=542, p=.0036, RMSEA=0.026, CFI=0.979; Group 2: X_2=512, p=.04; RMSEA=.019; CFI=0.991), there was an indication that two of the factors – Restricted Interests and Rituals and Routines—were so highly correlated (Group 1 φ =0.872, Group 2 φ =0.777) that they were functionally inseparable. As such, we collapsed these subscales into a single factor called "Higher-order," where indicator residual covariances were freely estimated. Both CFAs showed good model fit (Group 1 X_2 = 572, p<.001, RMSEA=.029, CFI=0.972; Group 2 X_2= 572, p<.001, RMSEA=.027, CFI=.981).

Moderated nonlinear factor analysis. We used MNLFA to determine whether the RBS-EC has MI, and to statistically account for any non-invariance. We were mainly concerned with longitudinal MI (i.e., whether the measurement is invariant across different ages), and used MNLFA to simultaneously determine how age as a continuous variable impacts RBS-EC latent scores, as well as its impact on DIF. Because of our two-cohort design, we also tested moderating effects of cohort, and the interaction between cohort and age since the interpretation of a main effect of age rests on the assumption that slopes do not differ between the two cohorts. Finally, we tested for moderating effects of sex, given past findings on sex-related differences in

RRBs (Wolff et al., 2016). Specifically, we conducted our analyses using a slightly modified version of Gottfredson and collegues' (2018) *aMNLFA* package for R. The aMNLFA package functions as a visualization and parameterization pipeline that works with the Mplus (Muthén & Muthén, 2017) statistical program to iteratively estimate the models we detail below.

For longitudinal data, MNLFA first estimates the model parameters on a calibration sample drawn from independent observations (n=180). All significant moderators on the mean and variance of the latent construct (η) and on DIF are combined into a final model, which is then used to estimate factor scores for the full longitudinal sample. All models were run on two independent calibration samples to test for model stability. Independent observations were pseudo-randomly selected, while attempting to ensure a similar age distribution to the full longitudinal sample (sample 1 mean age=18.07, SD=6.37, sample 2 mean age=17.66 months, SD=6.08), resulting in a 36% overlap between the two calibration samples.

Each of the three latent subscales of the RBS-EC (Repetitive Motor, Self-Directed, and Higher-order RRBs) were modelled separately. First, we modelled the effect of moderators (Age, Sex, Cohort, and Age x Cohort) on the latent construct (η). Only Age was tested as a moderator on the latent variance estimates, as suggested by past literature (Curran et al., 2014). For this initial step, model parameters with $p < 0.1$ were retained. Next, we modelled the effect of moderators on DIF (e.g. measurement artifact introduced by predictors) for each indicator intercept and loading, in the presence of moderators on the latent factor mean.

Based on the results of the initial impacts and DIF model, model parameters were then trimmed such that all effects (moderator effects on the latent mean and variance, and on DIF) with $p < .05$ were retained. The surviving impact and DIF effects were then tested simultaneously in one model. At this stage, Benjamini-Hochberg family-wise error correction

was applied to all model parameters to protect against Type I errors. Lastly, a final model was estimated using the parameters that survived the Benjamini-Hochberg correction. That model was then used to estimate the factor score estimates in the full longitudinal sample (n=606), yielding person- and visit-specific estimates of η for each latent subscale which reflect the weighted estimates of each individual's latent score, as well as the effects of significant moderators on DIF and η mean and variance.

Longitudinal Analyses. MNLFA provides individual factor score estimates for each RBS-EC subscale adjusted for individual differences in moderating factors such as age. While MNLFA provides estimates of moderator effects on η, using these scores as outcomes in a multi-level model allows for the estimation of both between- and within-person effects of age. To test normative and individual differences in children's RRB growth rates, we fitted taxonomies of growth models to each of the respective RRB subscales, moving systematically across linear and polynomial specifications of time. We subsequently added sex and cohort to model, to test for interactions with RRB growth rates. Any non-significant effects were dropped from the final model. Complete information on model comparisons for all subscales can be found in **Tables 4-6**. Model comparisons were conducted using chi square log likelihood ratio tests and second-order Akaike Information Criteria (accounting for sample size and model complexity). All models were fitted using the lme4 and LmerTest package (Bates, Mächler, Bolker, & Walker, 2015) in R 3.31.

For descriptive purposes, we fitted these models using both the MNLFA-derived factor scores and the raw-mean scores. Results for raw-mean scores can be found in the Online Supplement. However, it should be noted that because these variables are on different scales, absolute quantitative comparisons are impossible. Indeed, the different scales between the

MNLFA-derived factor scores (i.e., interval scaled) and the raw-mean scores (i.e., proportion scale) required different modeling approaches—general linear mixed models versus generalized mixed models (logistic link), respectively.

Results

MNLFA Results

The final MNLFA model results on the structural relation between moderators and latent factor means and variances can be found in **Table 2**, and final results on DIF can be found in **Table 3**. Overall, age was the only significant moderator of these structural models, and of item functioning. Items associated with motor ability (e.g. *lines up or arranges toys or other objects)* tended to be endorsed more frequently as infants got older, while items such as "*mouths, bites, licks, or sucks objects"* were endorsed less frequently with age, reflecting changes in the developmental appropriateness of such behaviors. Of note, higher-order items that index intense focus with objects (e.g. *focuses on parts of objects rather than the whole object)* were also endorsed less frequently with age.

MNLFA-derived factor score estimates for the longitudinal sample approached unity across the two separate calibration samples used for all three subscales (Repetitive Motor r=.99, Self-directed r=.99, and Higher-order r=.98, all p's <.001)**,** demonstrating that factor score estimates were not dependent upon the calibration sample used. MNLFA-derived factor scores were highly correlated with raw mean scores for all three subscales (Repetitive Motor (r=0.93), Self-directed (r=0.96), and Higher-order (r=0.96) behaviors, all p's < .001)). However, as illustrated by Error! Reference source not found. and **Figure 3,** the MNLFA-adjusted factor scores provided considerably greater variability than the unadjusted mean scores, because scores adjust for individual-level factors that influence the degree of variation (Curran et al., 2014).

Age-related change in Restricted and Repetitive Behaviors

Repetitive Motor. Likelihood ratio tests of nested models and AIC comparisons indicated that the longitudinal MNLFA-derived factor scores for repetitive motor behavior were best represented by a linear growth function (**Figure 4**). On average, children showed linear declines in their repetitive motor behaviors (B_{Age} = -0.11, p < .0001) between 8 and 37 months of age. Scaling on the within-person variance, this corresponds to an approximate 3.0 SD decrease across this span. Notably, a statistically significant random linear slope also indicated noteworthy individual differences in children's growth rates, around this mean trajectory (Δ-2ll =19.1, Δdf =2, p<.0001). Subsequent models indicated that cohort was not predictive of differences in children's intercepts (B_{Cohort}=.13, p=.2) or growth rates (B =-0.02 , p=.13), so it was not included in the final model. Sex was predictive of differences in children's intercepts (B_{sex}=0.21, p=.04), and of growth rates (B=-.034, p=.001) (**Figure 4**); girls had slightly lower Repetitive Motor behaviors at 18 months, and showed a more rapid decline in these behaviors than did boys. Collectively, the final model including linear Age, Sex and an Age x Sex interaction accounted for approximately 39% of the variance in repetitive motor behavior (R^2 =0.398%).

Self-directed. MNLFA-derived factor scores for self-directed behaviors were best represented by a linear growth function (**Figure 5**). On average, children showed linear declines in their self-directed behaviors (B_{Age}=-0.06, p <.0001) between 8 and 37 months of age. Scaling on the within-person variance, this corresponds to an approximate 2.14 SD decrease across this span. Neither Cohort nor Sex were predictive of differences in children's intercepts (B_{Cohort}=0.07, p=0.4; B_{sex}=0.015, p=0.86) or growth rates ($B_{Age*Cohort}$ =-0.007 , p=0.5; $B_{Age*Sex}$=0.009, p=0.27). Collectively, the final model including linear time accounted for approximately 17.6% of the variance in self-directed behavior (R^2 =.176).

Higher-order repetitive behavior. The longitudinal MNLFA-derived factor scores for higher-order behaviors were best represented by quadratic growth function (**Figure 6**). On average, children showed a slight decline in their higher-order behaviors at around 18 months (B_{age}=-0.002, p=.7; B_{age2}=-0.002 p=.0001). Scaling on within-person variance, this corresponds to a 0.06 SD decrease from 18 to 37 months. A statistically significant random linear slope also indicated noteworthy individual differences in children's growth rates, around this mean trajectory (Δ*-2ll* =11.46, Δ*df* =2, *p*<.0001). Subsequent models indicated that Cohort was not predictive of differences in children's intercepts (B_{cohort}=0.08, *p*=0.5) or growth rates (*B* =0.007, *p* =0.54), so it was not included in the final model. Sex was not predictive of differences in children's intercepts (*B*=0.19, *p*=0.08), but was predictive of linear growth rates (*B*=.024, *p*=.01) (**Figure 6**); while boys and girls had similar higher order behaviors at 18 months, girls showed a slight decline in these behaviors, while boys stayed flat. Collectively, the final model including linear and quadratic Age, Sex and an Age*Sex interaction accounted for approximately 74% of the variance in higher-order behaviors (R_2 =0.74).

Value Added

What if we had simply used the unadjusted mean scores, as opposed to the MNLFA-adjusted factor scores? As noted, the correlations between the unadjusted mean scores and MNLFA-adjusted factor scores were very strong. However, failing to account for longitudinal non-invariance led to some non-trivial differences in longitudinal trajectories. Most notably, when using raw unadjusted scores the observed decline in repetitive motor (**Figure 4**) and self-directed behaviors (**Figure 5**) flattened out over time (full model results can be found in the online supplement). This may be because modelling the bounded and right skewed raw-mean

data for these subscale results in a floor-effect for the older toddlers. Further issues arise when modelling raw unadjusted scores that afford limited variance. For example, for the repetitive motor subscale, the model failed to converge when we attempted to test for the effect of sex and cohort on intercept and growth rate, suggesting that with the limited variance afforded by proportion scores, only a limited number of parameters could be included in the model. For the self-directed subscale, non-significant random effects for the instantaneous linear slopes indicated that all children showed statistically identical growth rates.

Similarly, failing to account for longitudinal non-invariance and adopting the bounded and right skewed raw-mean higher-order behavior scale led to some non-trivial differences in children's higher-order growth trajectories, as compared with invariance-adjusted MNLFA scores (**Figure 6**). In this case, the best-fitting model for the raw mean scores included only fixed and random effects of intercept, and we were unable to detect either within- or between-person growth effects.

Discussion

The goal of the present study was to characterize normative rates of RRBs and their subsequent change over time in a large longitudinal community sample of toddlers. We tested longitudinal MI in the RBS-EC in order to examine the extent to which our constructs (repetitive motor, self-directed, and higher-order behaviors) were commensurate across child demographic covariates (age and sex). We found no differences in the meaning of scale (i.e. non-invariance) due to sex. However, there were some differences as a function of age, with the higher-order subscale showing the most evidence of DIF. Interestingly, many of these items describe unusual object and visual exploration (e.g. *focuses on parts of objects rather than the whole objects, closely inspects objects, lines up or arranges toys).* Past work has shown that unusual object and

visual exploration as 12-months is associated with subsequent autism severity scores (Ozonoff et al., 2008), suggesting that these items may be useful for predicting emerging atypicalities during a specific developmental window. However, past work has also demonstrated low rates of similar behaviors in 2-years-olds with ASD and non-spectrum developmental disorders (Richler, Bishop, Kleinke, & Lord, 2007), suggesting that they may be challenging to measure in the first years of life regardless of diagnostic group. Thus, future work comparing the trajectories of these behaviors between ASD and TD children must be careful to adjust for DIF both as a function of age and diagnostic status. Crucially, MNLFA scores adjust for DIF and therefore generate more precise estimates of these constructs.

Our model results indicate that RRB subscales have differing developmental trajectories, which in part validates their utility as separable constructs (Bodfish et al., 2000; Lewis & Bodfish, 1998; Turner, 1997). Self-directed behaviors showed a linear decrease from 3- to 36-months, as did repetitive motor behaviors which showed a slower decline in boys relative to girls. These findings corroborate past longitudinal work by Uljarević et al. (2017) showing a decline in repetitive and sensory motor behaviors beginning at 15-months, as well as higher rates of these behaviors in boys beginning at 15-months that becomes statistically higher than girls by 77 months (Uljarević et al., 2017). We found that higher-order behaviors begin to decline later in development relative to lower-order behaviors, also corroborating past work (Evans et al., 1997; Uljarević et al., 2017).

Conclusions

Why use MNLFA scores to examine rates of RRBs, as opposed to raw mean scores? First, adjusting for DIF due to individual differences in age provided us with more within-person

variability to model within-person changes in RRBs. This has potential implications for measuring behavioral change over time in treatment and intervention studies, beyond RRBs. Many have argued for the importance of outcome measures that are sensitive enough to capture subtle within-person change for assessing treatment outcomes (e.g. Kim, Grzadzinski, Martinez, & Lord, 2019). We argue in addition to the sensitivity of the measure itself, using scores that adjust for DIF confers the additional benefit of increasing within-person variability when measuring individual treatment outcomes or efficacy.

Second, these scores disentangle measurement bias introduced by age and true changes in the latent construct over time, allowing us to more accurately estimate their developmental trajectories. This may be particularly important in follow-up work examining group differences between ASD and TD toddlers in these trajectories, as DIF due to age may interact with risk status. For example, in the present study we found that items associated with routinized play (e.g. *lines up or arranges toys or other objects)* tended to be endorsed more frequently as infants got older. It is likely that in this item also functions differently in populations with neurodevelopmental disorders, in addition to DIF due to age. When comparing behavioral trajectories in TD and ASD infants, we must have a clear understanding of the underlying factor structure and be able to statistically adjust for biases in our measures due to diagnostic group and age for this information to be interpretable. Validity and reliability are not inherent attributes of an instrument, but are inextricably linked to the sample in which they were established. Indeed, establishing MI is essential in order to avoid what Meehl referred to as 'detached validity claims' (Meehl, 1990), in which instruments psychometrically verified in one sample or at one age are assumed to function similarly in other samples or ages of children. The implications of such assumptions are broad and potentially costly. Thankfully, with recent simplifications in the

implementation of MNLFA (Gottfredson et al., 2018), researchers can adjust for non-invariance more easily than ever.

In sum, this paper provides foundational evidence of the developmental trajectories of RRB sub-types among typically developing children. Future work will test the invariance of these metrics in a sample enriched for ASD risk, and will consider the role of language and cognition in RRBs across the typical-to-atypical continuum.

Declarations

Ethics Approval and Consent to Participate

This study was approved by the Institutional Review Board at the University of Minnesota. Parental permission was provided for all infant participants. This study was carried out in accordance with The Code of Ethics of the World Medical Association (Declaration of Helsinki).

Consent for publication

Not applicable.

Availability of Data and Materials

The datasets analyzed during the current study and the analysis scripts are available in the https://github.com/rrobinn/invariance-repetitive-bx repository.

Competing Interests

The authors have no financial competing interests. JJW led the creation/development of the RBS-EC, which is freely available

Funding

RS was supported by the National Science Foundation Graduate Research Fellowship Program and JJW was funded by NIMH R01 MH116961. This study was made possible by an NIMH award (R01 MH104324) to JTE. The funders had no role in study design, data collection, analysis, data interpretation, or the writing of the report.

Authors' contributions

RS: Contributions to design of analytic plan, data analysis and interpretation, drafting of manuscript.
DB: Contributions to design of analytic plan, data analysis and interpretation, drafting of introduction.
JW: Contributions to study conception and design, substantive revisions of manuscript.
JE: Contributions to study conception and design, drafting of introduction, substantive revisions of manuscript.

Acknowledgements.

The authors gratefully acknowledge all of the infants and families for their participation. The authors also acknowledge Carolyn Lasch, Elayne Teska, Rachel Roisum, and Kristen Gault for their involvement in data collection.

Tables

Table 1. Descriptive information on study sample. Ages are reported in months.

Table 2. Results from final MNLFA model testing covariate effects on factor mean and variance.

Table 3. Items for which there was a significant loading DIF as a function of the covariate Age.

Table 4. Table of Model evidence for predicting Repetitive Motor MNLFA scores. The best-fitting model (**Model 4)** included linear fixed and random effects of slope (Age), as well as a significant effect of sex of on the intercept and slope.

Table 5. Table of Model evidence for predicting Self-directed MNLFA scores. The best-fitting model (**Model 4)** included a linear fixed effect of slope (Age). Neither Sex nor cohort were included in the final model.

Table 6. Table of Model evidence for predicting Higher-order MNLFA scores. The best fitting model (**Model 4)** included a quadratic fixed effect and linear random effect for slope (Age), and a significant effect of Sex on intercept and slope.

Figures

Figure 1. Example of MIMIC model. An individual's response on Item 1 is due to the underlying latent construct of Repetitive Motor Behaviors (path a) as well as an age moderator (path b). Path c represents the factor loading on Item 1 after adjusting for the moderator.

Figure 2. Correlation between raw mean scores and MNLFA factor scores (Eta) for **a.** Repetitive Motor (r=0.94), b. Self-directed (r=0.96), and **c.** Higher-order (r=0.96) behaviors (all p's < .001).

Figure 3. Distributions of MNLFA-derived factor scores (top) and raw-mean scores (bottom) for a) Repetitive Motor, b) Self-directed, and c) Higher-order behaviors.

Figure 4. Best-fitting model estimates for Repetitive Motor scores plotted over raw data for **a)** MNLFA-derived factor score estimates and **b)** raw mean scores. The model outcome for the raw mean score is converted to probability, as a logistic linking function was used. Y-axes for both plots are scaled with limits of mean outcome +/- 2.5 SDs. **c)** Impact of Sex on intercept and slope. Parents of females reported fewer repetitive motor behaviors at 18 months, and their rate of decline was more rapid relative to males.

Figure 5. Best-fitting model estimates for Self-directed scores plotted over raw data for a) MNLFA-derived factor score estimates and b) raw mean scores. The model outcome for the raw mean score is converted to probability, as a logistic linking function was used. Y-axes for both plots are scaled with limits of mean outcome +/- 3 SDs.

Figure 6. Best-fitting model estimates for Higher-order scores plotted over raw data for **a)** MNLFA-derived factor score estimates and **b)** raw mean scores. The model outcome for the raw mean score is converted to probability, as a logistic linking function was used. Y-axes for both plots are scaled with limits of mean outcome +/- 4 SDs. **c)** Effect of Sex on Slope. The decline in Higher-order scores was driven by females.