Published in final edited form as:
Dev Psychol. 2011 November ; 47(6): 1565–1578. doi:10.1037/a0025418.

Imitation from 12 to 24 months in autism and typical development: A longitudinal Rasch analysis

Gregory S. Young,
M.I.N.D. Institute, Department of Psychiatry and Behavioral Sciences, University of California, Davis, CA

Sally J. Rogers,
M.I.N.D. Institute, Department of Psychiatry and Behavioral Sciences, University of California, Davis, CA

Ted Hutman,
Semel Institute for Neuroscience and Human Behavior, University of California, Los Angeles, CA

Agata Rozga,
School of Interactive Computing, Georgia Institute of Technology, Atlanta, GA

Marian Sigman, and
Semel Institute for Neuroscience and Human Behavior, University of California, Los Angeles, CA

Sally Ozonoff
M.I.N.D. Institute, Department of Psychiatry and Behavioral Sciences, University of California, Davis, CA

Abstract

The development of imitation during the second year of life plays an important role in domains of socio-cognitive development such as language and social learning. Deficits in imitation ability in persons with autism spectrum disorder (ASD) have also been repeatedly documented from toddlerhood into adulthood, raising the possibility that early disruptions in imitation contribute to the onset of ASD and the deficits in language and social interaction that define the disorder. This study prospectively examined the development of imitation between 12 and 24 months of age in 154 infants at familial risk for ASD and 78 typically developing infants who were all later assessed at 36 months for ASD or other developmental delays. The study established a developmental measure of imitation ability, and examined group differences over time, using an analytic Rasch measurement model. Results revealed a unidimensional latent construct of imitation and verified a reliable sequence of imitation skills that was invariant over time for all outcome groups. Results also showed that all groups displayed similar significant linear increases in imitation ability between 12 and 24 months and that these increases were related to individual growth in both expressive language and ratings of social engagement, but not fine motor development. The group of children who developed ASD by age 3 years exhibited delayed imitation development compared to the low-risk typical outcome group across all time-points, but were indistinguishable from other high-risk infants who showed other cognitive delays not related to ASD.

Correspondence concerning this article should be addressed to Gregory S. Young, M.I.N.D. Institute, University of California, Davis, 2825 50th Street, Sacramento, CA 95817. gregorys.young@ucdmc.ucdavis.edu.

Keywords

imitation; autism; Rasch analysis; early identification; HGLM

Introduction

Imitation has long been considered a critical component of the profound social and intellectual development that occurs over the first several years of life. During the second year in particular, a child's ability to imitate sounds, gestures, and actions increases dramatically as part of the continuous process of learning about the world and negotiating complex social relationships. From research on typical development, the natural history of imitative development has been described in some detail. Observational work conducted by a number of researchers (e.g., Abravanel, Levan-Goldschmidt, & Stevenson, 1976; Jones, 2007; Killen & Uzgiris, 1981; Masur & Rodemaker, 1999; McCall, Parke, & Kavanaugh, 1977) has documented the relative explosion of imitative behavior over the first 2 years of life, which occurs not only in vocal behavior – a presumably critical avenue for learning language – but also in gesture and in actions on objects. Moreover, there has been consistent evidence from both cross-sectional and longitudinal research that imitation develops progressively, from the imitation of simple, easily self-observable actions on objects with salient effects (e.g., banging on a noisemaker) to the imitation of complicated, unseen and relatively meaningless gestures (e.g., Abravanel, Levan-Goldschmidt, & Stevenson, 1976; Elsner, 2007; Jones, 2007; Uzgiris & Hunt, 1975).

As an early, critical developmental skill with implications for intellectual and social development, imitation has also received a great deal of attention in research on autism (e.g., Rogers & Pennington, 1991; Williams, Whiten, & Singh, 2004). In 1991, Rogers and Pennington suggested that early deficits in imitation among children with autism may be a universal, primary symptom that disrupts early social interaction and ultimately leads to a cascade of social and communication deficits, and this hypothesis is echoed in the more recent "mirror neuron hypothesis" of autism (e.g., Dapretto et al., 2006; Williams et al., 2006). Employing a variety of gestures and actions on objects (Dunst, 1980; Uzgiris & Hunt, 1975), prior research has demonstrated imitation deficits in toddlers with autism as young as 24 months (McDuffie et al., 2007; Rogers, Hepburn, Stackhouse, & Wehner, 2003; Stone, Ousely, & Littleford, 1997). Moreover, these studies found that imitation deficits were significantly related to concurrent deficits in play, joint attention, and language ability. This appears consistent with the idea that early disruptions in imitation could be partly responsible for shaping the early behavioral phenotype of autism. Nevertheless, relatively little research has examined whether imitation deficits are in fact present before the age when autism can be reliably diagnosed (i.e., before 24 months of age), although evidence converges on this possibility.

Some studies using retrospective parent report methodology have documented specific imitation deficits within the first 2 years of life among those later diagnosed with autism (Dahlgren & Gillberg, 1989; Ornitz, Guthrie, & Farley, 1977). Prospective screening studies, also using parent report, have likewise documented imitation deficits early in the 2nd year among children who later develop autism or ASD (Robins, Fein, Barton, & Green, 2001; Watson et al., 2007). Studies using direct behavioral observation have also revealed apparent imitation deficits between 12 and 30 months (Charman et al., 1997; Mars, Mauk, & Dowrick, 1998; Zwaigenbaum et al., 2005). Despite this convergent evidence for early imitation deficits in autism, however, a number of important methodological and theoretical issues remain.

One issue is the need to collect prospective, longitudinal data as a way to examine individual and group differences in change in imitation over time. Although research on autism has revealed cross-sectional group differences at specific time-points, there has been no longitudinal research on the developmental trajectories of imitation over the second year of life, and only one study documenting such change over time after age 2 (Stone et al., 1997). Examining early developmental trajectories of imitation between 12 and 24 months, when imitation increases so dramatically in typical development, could illuminate the process of imitative development in ASD and could reveal important relationships with motor development, language, and other social behaviors. Thus, one of the primary aims of the current study was to collect prospective longitudinal data on imitation skills from 12 to 24 months in children who are later diagnosed with autism spectrum disorder at 36 months.

A second theoretical and methodological issue to be addressed in the study of imitation is the specificity of early imitation deficits to autism. The use of comparison groups in longitudinal data is particularly important for addressing questions about specificity, since the groups may differ in patterns of change *over* time while not necessarily differing at a specific point *in* time. Although the one existing longitudinal study by Stone et al. (1997) found evidence for significant increases in imitation in autism between 30 and 46 months of age, no comparison groups were included to assess whether the observed rate of development in the autism group differed in any meaningful way. Ideally, the specificity of an imitation deficit in autism would be addressed by the inclusion not only of typical children, but of children with developmental delays as a way to determine whether the observed imitation deficit is simply associated with some non-specific delay rather than something about autism itself. Indeed, the hypothesis proffered by Rogers and Pennington (1991) that an early imitation deficit plays a causal role in the development of autism predicts that early imitation deficits in autism and their trajectories over time would be significantly different from other early childhood disorders. The current study addresses this need for assessing the specificity of early imitation deficits by measuring prospective imitative development in 4 groups of infants: 1) infant siblings of children with autism who develop autism by 36 months of age; 2) infant siblings of children with autism who exhibit developmental delays or other clinical concerns at 36 months of age; 3) infant siblings of children with autism who develop typically; and (4) infant siblings without a family history of autism who are developing typically. The use of a comparison group of infants who are at similar genetic risk as those who later develop autism but who experience other delays instead of autism was expected to provide a more stringent test of specificity relative to the heterogeneous samples of developmentally delayed children typically used in autism research (Jarrold & Brock, 2004; Tager-Flusberg, 2004).

A third, and perhaps the most important, issue brought up by prior research in both typical development and in autism is the need for careful definition and measurement of imitation itself. Imitation abilities have been measured in a variety of ways in prior research, from vocal imitation (Mars et al., 1998; Ornitz et al., 1977) to imitation of movements (Dahlgren & Gillberg, 1989) to imitation of both conventional and novel actions on objects (Charman et al., 1997; Rogers et al., 2003; Rogers, Young, Cook, Giolzetti, & Ozonoff, 2010). Moreover, a variety of measurement methods have been used, ranging from questionnaire items about spontaneous facial imitation occurring during a social exchange with the parent (e.g., Robins et al., 2001) to observable prompted imitation occurring during a laboratory visit with an unfamiliar adult (e.g., Zwaigenbaum et al., 2005). From past research on the development of imitation, it is clear that a variety of things impact imitative performance, including the meaningfulness of the actions (e.g., McGuigan, Whiten, Flynn & Horner, 2007), the saliency of effects produced by the acts (e.g., Hauf, Elsner, & Aschersleben, 2004), the ability to visually self-monitor one's actions, and the use of objects (Abravanel et al., 1976; Masur, 2008). It seems to be generally assumed in much of the literature that these

various task characteristics reflect actual distinct dimensions of imitation, perhaps each influenced by separate cognitive and motivational mechanisms. Although there is a degree of face validity to this assumption, the existence of discrete dimensions of imitation is an important empirical question that has not yet been clearly established.

In research on children with autism, studies by Stone et al. (1997) and Rogers et al. (2003) have suggested specific autism related deficits for certain types of imitative tasks, with the interpretation that such specific areas of deficiency are unique to autism. Indeed, research documenting relatively poorer performance on gesture relative to action on object tasks, or significant group differences on only one type of imitation has been cited as evidence that imitation is not a unitary skill (e.g., DeMeyer et al., 1972; Hobson & Lee, 1999; Stone et al., 1997). However, such "dissociations" are still entirely consistent with the possibility that these putative dimensions of imitation simply reflect different levels of difficulty along an underlying single continuum of imitation. Indeed, research on typical development, using both cross-sectional and longitudinal samples (e.g., Abravanel et al., 1976; McCall et al., 1977) has regularly found that younger children are less proficient at imitating gestures than actions on objects but that both types of imitation nevertheless steadily increase over time, a finding that is likewise consistent with an underlying single dimension of imitation, despite claims to the contrary. As such, autism deficits on a type of imitation such as gestures might instead reflect an overall general imitation *delay* rather than a specific deficit in a dissociable dimension of imitation; items reliably failed by children with autism may simply be more difficult items measuring the same general imitation skill, and those more difficult items may be mostly of a similar type such as gestural imitation items.

A second argument for the multidimensionality of imitation is evidence for differential relationships between other developmental skills such as play or language and presumed types of imitation. For instance, Stone et al. (1997), Rogers et al. (2003), and McDuffie et al. (2007) all reported varying degrees of relatedness between domains of imitation (e.g., oral, object, gesture, etc.) and other developmental skills such as language, play, and fine-motor development. Such patterns within correlation tables have been interpreted as evidence for the multidimensionality of imitation. Unfortunately, although direct tests of such differing correlation patterns were not explored in any of these papers, an examination of the reported correlations in each of these papers reveals that virtually none of these coefficients are statistically different from each other (using Fisher's z transformation), suggesting that such relationships between various developmental constructs and presumed types of imitation are more similar than not – a result that actually supports the notion that imitation may best be conceptualized and measured as a unitary skill. Similarly, the correlations between types of imitation themselves may often be fairly high (e.g., Rogers et al., 2003), again suggesting that imitation as measured in such studies may best be conceptualized as a unitary phenomenon.

In addition to building upon the prior literature on imitation in autism with an early, longitudinal sample and the use of multiple comparison groups, the current study was also an attempt to address this third issue of measurement. Using a 10-item battery of imitation including actions on objects, manual gestures, and oral facial imitation items, we attempted to assess the dimensionality of the battery for evidence of discrete, statistically separable dimensions that exist invariant across time and between groups.

Method

Participants

Families with an older child with ASD or typical development (the proband) and an infant under 18 months were recruited as part of a larger longitudinal study examining infants at

risk for autism at two separate research sites (UCLA and UC Davis). A total of 325 families enrolled (UCLA = 164, UCD = 161), 203 of whom were "high-risk" families with at least one older child diagnosed with an autism spectrum disorder (ASD). A comparison group of 122 "low-risk" infant siblings was also enrolled in which there was no family history of autism or ASD in any 1st, 2nd or 3rd degree relatives and no older siblings had any signs of behavioral, emotional, or developmental disorders. ASD diagnoses of probands were confirmed by medical record review, supplemented with additional formal diagnostic testing using the Autism Diagnostic Observation Schedule (ADOS; Lord, Rutter, DiLavore & Risi, 1999) in cases where such records were equivocal or lacking, and scores above the ASD cutoff on the Social Communication Questionnaire (SCQ). Fifty-seven percent of probands met criteria for full autism, and the remaining 43% met criteria for ASD. Infant siblings were enrolled between 1 and 18 months of age, with 64.6% enrolled by 6 months, and 86.2% enrolled by 12 months of age.

For the primary imitation measure used in this study (described below), valid data was available for 248 of the 325 infants for at least one of the three measurement points (12, 18, or 24 months). Infants without usable imitation data either refused to cooperate with testing or left the study prior to diagnostic outcome testing at age 3 (described below). There were no differences between infants with and without imitation data at the time of attrition on any demographic measures such as minority status, income level, gender, risk-group, or site, as well as behavioral variables like IQ or language ability. Missing data points among infants included in the sample were likewise not a function of any of these demographic variables. Seventy-three infants had usable imitation data from only 1 visit, 70 had usable data from 2 visits, and the remaining 105 had usable data from all three visits. There was no relationship between number of visits with usable data and risk-group or outcome status. Of the 248 infant siblings in the final sample, 154 were high-risk infants and 94 were low-risk infants.

Family history and diagnostic assessments carried out at 36 months were used to further classify infants into distinct outcome groups for purposes of analysis, using the standardized measures described below and algorithms developed by the Baby Siblings Research Consortium (presented in Table 1). Three of the 94 children in the low-risk group were classified with autism/ASD, and 16 of the 94 children in the low-risk group were classified as having other developmental concerns. The 3 low-risk children with autism/ASD were retained, whereas the 16 low-risk children with other developmental concerns were removed from the sample so that the other developmental concerns group would represent a more meaningful comparison group of high-risk children with subclinical symptoms such as speech language delays (although results reported below did not differ when such low-risk delayed subjects were included). The final sample consisted of 232 infants in one of 4 categories: (1) autism/ASD (n=24), (2) other developmental delays (n=43), (3) high-risk typical children (n=90), and (4) low-risk typical children (n=75). Sample characteristics at the 36 month outcome time point are shown in Table 2.

Measures

Imitation Battery—The imitation battery was based on that reported by Rogers et al. (2003) and consisted of 10 items that involved performing relatively simple actions such as clapping, banging a block with a stick, or making a raspberry sound. It was administered well into a larger test battery, after the infant had developed a comfortable, friendly relationship with the examiner. Infants were typically either seated in their parent's lap (for the younger infants) or in a high-chair with the mother beside the child (for older toddlers). Each item was administered by the examiner seated across from the infant at a table. Items were administered in a set order according to the Uzgiris-Hunt scales. For each item, the examiner modeled the action three times in quick succession and then invited the infant to

imitate by smiling, gesturing to the infant, looking expectantly, and saying "Now you do it." The examiner waited for the infant to imitate. If the infant did not imitate with at least a partial performance, as defined below, the examiner provided up to two more opportunities to imitate by repeating the procedure. The next item was presented as soon as the child produced a partial imitation or failed across all three opportunities. Items were not modeled by the examiner unless the examiner clearly had the infant's attention. In the few cases where an item was modeled without the infant's full attention, it was not counted in the scoring. Each item was scored on a 3-point scale: (1) Fail, where the child did not imitate despite being engaged, or responded with an unrelated action; (2) Partial-pass, where the child approximated the examiner's demonstration with error; and (3) Perfect-pass, where the child imitated the examiner's demonstration with a high degree of accuracy. Table 3 presents a list of each action and the respective scoring criteria.

All imitation sessions were either scored live by examiners (23.9%, n=168) or were recorded to DVD and scored from video (76.1%, n=539). All coders were trained in the scoring criteria using a manual and multi-media training materials. All examiners and coders were blind to group membership. Examiners and coders were required initially to code video examples from a prior study and were required to establish reliability on at least 10 examples per item, with weighted kappas above .8 for each item. For any given items that coders failed to achieve reliability on, the coder was required to code additional sets of 10 video examples per item until reliability criteria were met. All coders and examiners met reliability criteria for each item prior to coding actual data live or from video. There were no significant differences between raw imitation scores from live vs. from video scoring. During the course of the study, reliability was maintained by double coding 10% of sessions. Reliability estimates for maintenance coding remained high, with a mean weighted kappa = . 84 (range .72 to .91).

Mullen Scales of Early Learning (MSEL; Mullen, 1995)—The MSEL is a normed, standardized developmental measure of language, cognitive and motor functioning that provides age equivalent and standard scores (M=50, SD=10) from birth to 68 months of age on four separate subscales: visual reception, fine motor, expressive language, and receptive language (gross motor functioning was not assessed). It also provides an overall standardized score of developmental functioning, the Early Learning Composite (M=100, SD=15). The MSEL was administered at ages 12, 18, 24, and 36 months.

Autism Diagnostic Observation Schedule-Generic (ADOS; Lord et al., 1999)—The ADOS is a standardized play-based behavioral observation measure of autism symptoms consisting of 25 items across four domains: social interaction, communication, repetitive and stereotyped behaviors, and play. The ADOS yields scores summarizing the number and severity of symptoms in each domain and provides clinical cut-off scores for use in diagnosis of autism spectrum disorders and autistic disorder. Standardized severity scores were also calculated following procedures outlined in Gotham, Pickles, & Lord (2009). All examiners were required to meet reliability criteria of greater than 80% exact agreement in scoring and administration as part of initial and ongoing training. All reliability scoring and training was conducted by licensed psychologists with expertise in autism diagnosis and treatment. The ADOS was administered at 18, 24, and 36 months; however, diagnostic status was based only on the 36 month data.

Social Communication Questionnaire (SCQ; Berument, Rutter, Lord, Pickels, & Bailey, 1999)—The SCQ is a parent report questionnaire with 40 yes/no items about behaviors characteristic of autism. The SCQ was originally developed for use with children age 4 or over, but has been used successfully with younger children as well (Corsello et al.,

2007). The SCQ was used to supplement clinical diagnostic judgments at the time of outcome.

Outcome diagnostic form—A formal clinical diagnosis of autism or PDD-NOS based on symptom criteria outlined in the DSM-IV-TR (APA, 2000) was completed by a clinical psychologist at the 36 month visit. Symptom presence or absence in each of 3 domains (communication, social, repetitive and stereotyped behaviors) was indicated by the clinician using scores from the ADOS, scores from the SCQ, and behavioral observations of the child's behavior during other testing. This clinical rating was used to determine autism spectrum disorder as a final outcome at 36 months.

MacArthur Communicative Development Inventory (CDI; Fenson et al., 1993)—The CDI is a parent questionnaire that assesses a variety of aspects of language development, including vocabulary production, grammar, and sentence construction. The total raw word production score was used, consisting of the number of words endorsed by the parent out of 680 words across 22 categories (e.g., clothing, body parts, action words, etc.). The CDI was administered at ages 12, 18 and 24 months.

Examiner Ratings of Social Engagement—Examiner ratings, described in Ozonoff et al. (2010), were used as a measure of overall social engagement during each testing session. Examiners rated subjects on 3 social behaviors – eye-contact, shared affect, and social responsiveness – using a 3 point scale for each which were then summed together for a total social engagement score. Data using this measure on a number of the same children used in the present study were previously shown to discriminate growth trajectories between 6 and 36 months for children with ASD and those with typical development (Ozonoff et al., 2010). Examiner ratings collected at 12, 18, and 24 months were used for the present study.

Procedures

This study was conducted with the approval of the UC Davis and the UCLA IRBs. Infants were seen longitudinally for standardized testing and the imitation battery at 12, 18, 24 months, with follow-up diagnostic testing at 36 months (plus or minus 2 weeks, with gestational age corrected to 40 weeks when less than 36 weeks). All examiners were blind to infant risk-status and parents were instructed by a third party to assist in keeping experimenters blind by not discussing the infant's older sibling and his or her diagnosis with the examiner.

Analytic Strategy

We employed a statistical approach that allowed us to explore hypothesized longitudinal deficits in imitation specific to autism while simultaneously assessing the measurement properties of a 10-item imitation battery. The measurement model we employed was the Rasch model – as special instance of Item Response Theory – where a child's score for any given imitation item (pass or fail)[1] is assumed to be a logarithmic function of the difference

[1]Although in our imitation battery each item was scored on a 3-point scale (fail = 0, partial-pass = 1, perfect pass = 2), instead of a 2-point scale, each item was represented in the analysis as two dichotomous scale steps, recoding the original single item score set {0,1,2} as: {0,1,1} for the first dichotomous scale step (i.e., fail vs. partial or perfect pass) and {0,0,1} for the second dichotomous scale step (i.e., fail or partial pass vs. perfect pass). Given that each pair of dichotomous scale steps was necessarily correlated per item, this local item dependence was, in turn, modeled as a separate random effect nested within the overall item (see Doran, Bates, Bliese, & Dowling, 2007). This formulation yielded random effects representing the relative difficulty of each scale step (i.e., a scale step between 0 and 1, or between 1 and 2) relative to the overall item difficulty. As such, the scale step random effects correspond to Thurstone thresholds in a partial-credit Rasch model, and were thus added to the overall item fixed effect coefficient to produce difficulty estimates for each scale step of each item in terms of the whole scale. In this way, for the initial 10-item scale, 20 difficulty estimates were calculated across a single continuum of difficulty.

between the particular child's ability and the particular item's difficulty. Kamata (1998; 2001) and others have demonstrated that this basic formulation of the Rasch model can be recast in terms of a hierarchical generalized linear model (HGLM) with maximum likelihood estimation using a binomial distribution for item response and a logit link function to relate model parameters to the response.[2] Item difficulties (D_i) are modeled in logits as fixed effects with a structural level-1 model. Items are modeled as nested within and invariant across persons at level 2, where person abilities (B_j) are modeled as random effects. The anti-log of the difference between any single random effect (ability) and fixed effect (difficulty) is therefore the probability of that particular subject passing that particular item.

A benefit of expressing the Rasch measurement model within the framework of HGLM is that it affords one the ability to include rate of change parameters or additional level-3 person variables, such as diagnosis or IQ as additional predictors of subject scores (Pastor & Beretvas, 2006). Thus, using this HGLM approach, we were able to pursue two primary sets of analyses, corresponding to our two primary aims. The first set of analyses concerned scale evaluation – evaluating the measurement properties of the imitation scale within the Rasch framework. The second set of analyses built upon the final scale model and employed conditional models to examine differences between outcome groups in the development of imitation over time.

Scale Evaluation—In order to evaluate the measurement properties of the imitation battery, we first fit unconditional models to the data with only item and subject effects. We examined both infit and outfit residual statistics as indicators of unidimensionality in the measure, as well as threshold ranges and response category frequencies as indicators of scale step utility and redundancy. These first models allowed us to revise the scale and ensure a fit to the Rasch model by collapsing across redundant scale steps, or separating out scale steps or items that displayed poor fit statistics (see Bond & Fox, 2007). To the degree that individual items show poor fit to the idealized Rasch model – a unidimensional scale model – evidence for separate dimensions, or factors, is obtained. Factor analysis of item residuals (i.e., the degree of item misfit) can then be employed to assess the existence of second or third dimensions (Wright, 1994). The existence of additional factors can then be explicitly modeled within the HGLM as effects in their own right (either correlated or uncorrelated), and group differences or developmental differences between such factors can be assessed (Kamata, 1998)

Following the assessment of item fit and scale dimensionality, HGLM is then used to explore differential item functioning (DIF) as a function of the following higher order variables: site, time, outcome group, and time by outcome group (Williams & Beretvas, 2006). An important assumption in the Rasch model, and of any good unidimensional scale (or factor), is that the relative difficulties of items *within* the scale remain invariant over time and between groups. To the extent that one particular item becomes significantly easier (or more difficult) over time or between groups relative to other items, we can say the item exhibits DIF and needs to be removed from the measure to ensure measurement invariance (Bond & Fox, 2007). Invariance of a measure across such contexts does not preclude overall group differences or even different developmental trajectories between groups with respect to the measured construct itself; rather it necessitates that any such differences are not artifacts of, or confounded by specific items that measure something other than the construct of interest. That is to say, the degree to which items on a given scale or factor all measure

[2]As each scale-step is used as an indicator of the latent ability trait, those items with collapsed scale steps do contribute less to the estimation of the latent trait. However, given the establishment of invariance and unidimensionality of the overall scale, the resultant ability estimates are not biased by such weighted item contributions.

the same thing, they will necessarily be invariant across contexts such as time and group. To explore measurement invariance, we examined interaction terms of each higher order variable with each item and tested for significant interactions (see Luppescu, 2002; Pastor & Beretvas, 2006; Williams & Beretvas, 2006). This step allowed us to evaluate the degree to which the ordering and location of scale item difficulties was invariant over such higher level terms. Items that demonstrated significant interaction effects with time or with group were considered to be biased in that they failed this invariance test and were then removed from the item pool as a further distillation of the scale(s). As a consequence of this process, we were assured of having a scale (or multiple factors) that measured a single construct on a single metric and could then move on to answer questions about how ability in this distilled measure of imitation differs between groups or develops over time.

Conditional Models—In order to examine our hypotheses regarding group differences in imitation ability over time, we used the final HGLM model from the scale evaluation stage (i.e., the final model after collapsing scale steps and/or culling items as necessary) as a framework within which to examine rates of change and additional person variables such as outcome diagnosis and other time-varying covariates that might be associated with differences or changes in imitation skills.

In order to facilitate interpretation of item effects, no intercept term was included in models. This allowed us to generate item difficulty estimates as logistic deviations from 0; all higher-level effects such as time and group parameters remained unchanged as a result. All model effects (e.g., main effects or interaction terms) were tested using the difference between -2log-likelihood values of nested models evaluated as chi-square statistics with the degrees of freedom equivalent to the difference in the number of parameters between models. All analyses were conducted in R, version 2.9.1, using R package lme4 (Bates & Maechler, 2009).

Results

Scale evaluation

Item fit—The first model included all 10 items of the imitation battery as fixed effects with participant intercepts modeled as random effects. Time variables were not included as fixed or random effects in order to estimate unadjusted item parameters. Individual scale step threshold (difficulty) estimates, modeled as random effects (see footnote [1]), were added to each item fixed effect in order to calculate difficulty estimates across the entire 20-point scale. For all items, both outfit and infit mean-square values were well within the acceptable range (.5 to 1.5) indicating that item data fit the unidimensional Rasch model well and suggesting no evidence for additional factors. An examination of ranges of thresholds for item scale-steps, however, suggested that four items had a narrow scale-step difficulty spread of less than 1 logit: clap hands, open-close hands, open-shut mouth, and pat baby. Further examination of the frequency of scale step responses for these four items revealed that most participants received either a score of 2 (perfect pass) or a score of 0 (fail). As such, we decided to collapse partial and perfect pass scores together for these items (with the result that random effect thresholds for these four dichotomized items became essentially zero).[3] We then reanalyzed the model in order to evaluate again item performance after revising these four items. All fit statistics were again well within the acceptable range, with

[3]All Rasch analyses reported here using HGLM techniques were replicated using Winsteps software which is dedicated to Rasch analysis (Linacre, 2009). All item fit statistics, item difficulty estimates, and DIF analyses were essentially the same for both statistical approaches.

good spread between the remaining scale-step thresholds in the 6 unaltered items. This revised scale, of 16 steps within 10 items, was then used for the next analysis phase.

Analysis of Linear and Quadratic time effects—To decide whether to include only a linear or both a linear and quadratic effect for time, we expanded on the final model above by analyzing two separate models: one with a linear fixed effect for time and a random linear slope for participants, both centered at 18 months, and a second model with both linear and quadratic fixed effects for time with both random linear and quadratic slopes for participants, again centered at 18 months to reduce multi-collinearity.[4] A comparison between the two models using the difference between their respective -2log-likelihood values, evaluated using a chi-square distribution with 4 degrees of freedom (the difference between number of model parameters), revealed that the model with both the linear and quadratic terms was preferable (χ^2=191.70, $p < .001$). Thus both linear and quadratic effects for time were included in all subsequent models. The linear effect ($\gamma = 0.165 \pm 0.017$, $z = 9.87$, $p < .001$) yielded an odds-ratio of 7.24 (95% CI = 2.99 to 17.55) from 12 to 24 months, indicating a more than 7-fold increase in the probability of passing any given item at 24 months versus 12 months. The quadratic effect ($\gamma = 0.008 \pm 0.004$, $z = 1.85$, $p = .07$) indicated a slight convex (downward) curvature of the logits of correct item responses over time corresponding to a slight acceleration in growth over time. The correlation between the variance components for centered linear and quadratic effects was 0.34, indicating a relatively low correlation between the terms. The correlation between the variance components for linear time and intercept was also low ($r = .18$), but was moderate for the quadratic effect and intercept ($r = -.67$), suggesting that higher imitation abilities at intercept (i.e., at 18 months) were related to less curvilinear rates of development over time.

Differential Item Functioning (DIF) Analysis—We next investigated DIF as a function of the higher order variables: site, time, and outcome group using both linear and quadratic time effects. Significant interaction effects between individual items and rates of change or other level-3 variables of interest were interpreted as indicative of significant bias in the item.

There were no significant item by site interaction effects (χ^2=12.13, df=9, $p = .21$), indicating that the item difficulty estimates were consistent across sites. Moreover, the main effect for site was not significant, indicating that estimates of item response probabilities overall did not differ as a function of site (χ^2=0.50, df=1, p=.48). As such, site was not included in any additional models.

For rates of change, there was a significant effect for the item by quadratic growth interaction (χ^2=17.54, df=9, $p < .05$). Examination of individual parameter estimates revealed a significant effect for ‘pat cheeks’ as a function of quadratic change ($\gamma = -0.021 \pm 0.009$, $z = -2.47$, $p < .05$), indicating that response probabilities for the pat cheek item showed decelerating growth compared to the rest of the model. This item was removed from the set of items and the analysis was repeated for the set of 9 remaining items with the result that no additional items showed DIF for quadratic growth (χ^2=8.78, df=8, p=.36). There was also a significant effect for item by linear growth interaction (χ^2=62.62, df=8, $p < .001$) wherein the item ‘pat table’ showed significant DIF as a function of linear change ($\gamma = -0.114 \pm 0.028$, $z = -4.12$, $p < .001$), indicating that pat table response probabilities increased at a significantly slower rate than the rest of the scale. The ‘pat table’ item was likewise

[4]Although 3 time-points are generally not sufficient for estimating quadratic effects in growth curve models where subjects are modeled as level-1 random effects, the model used here allowed for this estimation because the available degrees of freedom for each level-2 participant effect consisted of the number of item scale-steps at each age (e.g., 11 scale steps at each of 3 visits in the final model).

removed and reanalysis with the remaining 8 items revealed no other item by time interaction effects, suggesting that the scale without these items met longitudinal invariance requirements.

We next investigated DIF in the resulting 8-item scale as a function of outcome group. For these analyses we set the low-risk typical group as the reference group to model the assumption that any item biases would best be evaluated as deviations from the most normative group. Results revealed a significant overall group by item interaction effect (χ^2=37.28, df=21, $p < .05$). Examination of model parameters revealed a single significant effect for the 'pat baby' by ASD group term (γ = -1.644 ± 0.693, z = -2.37, $p < .05$), indicating that the pat baby item was significantly more difficult for the ASD group than for the low-risk typical group relative to the rest of the scale. Considering the range of difficulty estimates of the rest of the items as seen in Table 4, this difference suggested that, for the ASD group, the pat baby item was one of the most difficult items of the entire scale whereas for the low-risk typical group, it was a moderately easy item. Given this degree of DIF and the likelihood that the item was measuring something quite different for the ASD group than for the low-risk typicals, the pat baby item was removed from the scale and the model was refit to test for additional group DIF among the rest of the items. No other items showed signs of bias against any of the groups when compared to the low-risk typical group.

In order to examine bias in item difficulties over time as a function of group, we next modeled the 3-way interaction of item, time, and group. Results of this analysis revealed no significant 3-way interaction effect (χ^2=16.51, df=18, p = .56), suggesting item invariance over time for each group.

Final Scale—Table 4 presents the final 7-item scale statistics after the process of collapsing scale steps and removing items in response to our scale evaluation analyses. The item difficulty estimates are unadjusted for time or group fixed effects so as to present an average of the overall scale and its item ordering (see Kamata, 2001, for a discussion on the presentation of adjusted vs. unadjusted item estimates).

Overall model summary statistics were calculated from the final 7-item, 11-step scale for both persons and for items. Item reliability, a coefficient representing the reliability of item difficulty estimates, was .99, suggesting that the item ordering and scaling provided by the Rasch analysis was highly reliable. Person reliability, a coefficient representing the reliability of person ability estimates (conceptually equivalent to Chronbach's alpha) was . 63, the smaller magnitude of which reflects the limited number of items included on the scale. Using the Spearman-Brown Prophecy formula, it was determined that increasing person reliability to .80 would require expanding the scale length from 11 scale-steps to at least 26 scale-steps.

Estimated scale scores (proportion correct) were generated for each participant as the sum of the probabilities for passing each item, with such probabilities calculated as the inverse-logit of the difference between the participant's ability and the item difficulty. These estimated scores were then compared to raw data scores derived from the same items (collapsing scale-steps for the 4-items as above), which were also expressed as the proportion correct (i.e., the sum of item raw scores divided by 11). The correlation between the Rasch model estimates and the raw scores was .94 (95% CI = .93 to .95), suggesting that ability estimates were highly consistent with the original raw scale scores.

Group differences in imitation over time

The next phase of analysis examined person-level variables building upon the same HGLM measurement model described above. Demographic variables such as gender and other

variables shown in Table 1 were not associated with imitation in any of these analyses and are not discussed further. The main effects of group and interaction effects between group and time were specifically examined as a way to evaluate our hypotheses of an early imitation deficit in autism and a slower rate of growth compared to other groups. For all analyses, the group with ASD was used as the reference group such that all item parameters reflected item difficulty for those with ASD, and level-3 group effect parameters reflected deviations in overall imitation performance from the referent ASD group. The group main effect was tested with chi-square tests of the difference between -2log likelihood values between the 7-item model with only linear and curvilinear time effects and the 7-item model with both time effects and the group effect. Overall group by time interaction effects (both linear and curvilinear) were similarly assessed using chi-square tests of model improvement between subsequent nested models. Given that time was centered at 18 months to minimize collinearity all simple effects for group reflected intercept differences at 18 months.

Average imitation scores (again calculated as the sum of item probabilities for each subject) are shown for each group at each age in Table 5. Results of the HGLM analyses revealed a significant group main effect (χ^2=283.76, df=3, p < .001), with the ASD group exhibiting significantly lower overall imitation abilities than the low-risk typicals (γ = 0.79, ± 0.353, z = 2.23, p < .05), corresponding to an odds-ratio of 2.20 (95% CI = 1.10 to 4.40) – a greater than two-fold increase in the probability of low-risk typicals passing any given item compared to the ASD group. The ASD group also exhibited marginally lower abilities than the high-risk typicals (γ = 0.60, ± 0.346, z = 1.74, p = .08), with an odds-ratio of 1.82 (95% CI = 0.93 to 3.59). Imitation in the Other delays group was not significantly different from the ASD group (γ = 0.18 ± 0.384, z = 0.46, p = .64). With respect to group differences in rates of change, there were no significant group by time effects for either linear change (χ^2=3.10, df=3, p = .38) or quadratic change (χ^2=4.46, df=3, p = .22). As a result, simple comparisons of group when time was re-centered at 12 months or at 24 months yielded similar significant group main effects. Individual participant ability scores over time (centered for presentation purposes at zero logits for low-risk typicals at 18 months) are shown in Figure 1, with estimated growth trajectories for each group superimposed on the individual ability data.

Analysis of time-varying covariates—We next analyzed the degree to which changes in other measures were related to changes in imitation and whether such relationships differed as a function of group. Means and standard deviations for the variables considered as covariates are also shown in Table 5 as a function of both group and time point. We first considered fine motor ability as indexed at each age by Mullen fine motor age-equivalent scores. Results of the HGLM analyses with fine-motor scores added to the group main effects model reported above revealed no significant effect for changes in fine-motor age equivalent scores in relation to imitation scores (χ^2=0.48, df=1, p = .49), and no significant interaction effects with group (χ^2=2.13, df=3, p = .55).

Analysis of expressive language age-equivalent scores on the Mullen between ages 12 and 24 months revealed a significant main effect for language (χ^2=25.12, df=1, p < .001) when compared to the model with only time and group main effects, with an odds-ratio of 7.32 (95% CI = 2.50 to 7.32) for a 12-month increase in language age equivalent scores. There was no group by language interaction effect and no time by language interaction effect. Inspection of model parameters revealed that with the inclusion of Mullen language scores, simple effects for group differences were no longer significant (p = .25, .58, and .76 for ASD vs. low-risk typical, high-risk typical, and other delays, respectively).

As a validation of the relationship between Mullen expressive language and imitation, a separate but similar analysis was conducted for vocabulary production as reported by

parents on the MacArthur CDI. Given the high correlation between the Mullen expressive language age equivalent scores and CDI vocabulary production (r=.84, 95% CI = .80 to .86), the Mullen expressive language scores were not retained in the model for this analysis to avoid problems with multicollinearity. Consistent with the analyses for the Mullen expressive language data, results revealed a significant effect for parent reported vocabulary over-and-above the baseline time and group main effects model (χ^2=22.63, df=1, p < .001), with an odds-ratio of 4.32 (95% CI = 2.45 to 7.62) for an increase of 300 words. The main effect for group was also again not significant after inclusion of vocabulary in the model. There were no group by vocabulary or time by vocabulary interactions, and no higher-order three-way interactions.

Analyses of examiner ratings of social engagement were conducted with Mullen expressive language age equivalent scores retained in the model. There was a relatively low correlation between social engagement ratings and Mullen expressive language scores (r=.21, 95% CI = .10 to .32). Analyses revealed a significant effect for social engagement ratings compared to the model with age, group, and Mullen expressive language (χ^2=15.50, df=1, p < .001), with an odds-ratio of 1.48 (95% CI = 1.24 to 1.78) for a 1-point difference in social engagement ratings. The main effect for Mullen expressive language after including social engagement ratings was attenuated to marginally significant effect with an odds-ratio of 1.94 (95% CI = 0.97 to 3.89) for a 12 month increase in expressive language age (p = .06). Analyses did not reveal any group by social engagement interactions with respect to the development of imitation ability, and no higher-order three-way interactions.

Discussion

This study had two primary aims: (1) to examine the measurement properties of a behavioral imitation battery involving prompted imitation of simple actions and actions on objects, and (2) to test the hypothesis of an early imitation deficit in autism prior to formal diagnosis. Both research aims were addressed by applying the same analytic framework – a hierarchical generalized linear model (HGLM) – within which to evaluate both the measurement properties of the imitation scale and differences in individual abilities over time as a function of outcome group.

Rasch Analysis of the Imitation Battery

With respect to the measurement properties of the 10-item behavioral imitation measure, results of HGLM analyses initially revealed that all items fit the Rasch model well as indicated by fit statistics. Because the Rasch model is an idealized unidimensional model, the fact that all 10 items fit the model suggests that there was no compelling evidence for multidimensionality among the items and no reason to conduct further analysis of item residuals in pursuit of uncovering additional factors to be included in the measurement model. Given prior literature on imitation and the presumed separate dimensions of imitation such as actions with objects versus manual gestures, this finding was somewhat surprising; the full 10 item scale used in our study contained a variety of types of imitation from actions on objects to manual gestures to oral/facial actions which could have formed separate, independent scales had the data supported it. Because the Rasch model assumes unidimensionality, the degree to which separate, multiple dimensions exist within the scale would be revealed by the extent to which certain items violated this unidimensional assumption, thereby prompting the explicit modeling of such discrete dimensions. According to the fit statistics, however, all 10 items appeared to index a single general imitation construct.

In addition to examining fit statistics as evidence for unidimensionality, however, we also examined differential item functioning (DIF) as evidence for measurement invariance – an

aspect of measurement that should be true for any unitary construct. Specifically, we examined item data for DIF as a function of both group and time with the assumption that our unidimensional measure of imitation should exhibit the same item ordering and the same difficulty scaling regardless of chronological age and of group membership. With respect to longitudinal invariance, two items – pat cheeks and pat table – failed to show adequate stability over time relative to the rest of the scale. Although all items on the scale decreased in difficulty over time for a given ability – as would be expected for any developmental measure – the difficulty estimates for both the pat table and the pat cheeks items changed at different rates than the rest of the scale. Given the implicit assumption that any pure measure of a unitary construct should change in the same way over time (by definition), the significant longitudinal DIF seen for both the pat cheeks and pat table items indicated that these items may have been measuring something else in addition to imitation. As such, these items were removed from the scale in order to achieve a more unidimensional measure of imitation.

DIF analysis was also conducted to test invariance between groups and to identify items that might be specifically biased toward one group or another. The DIF analysis for group revealed one item that was significantly biased against the ASD group: the pat baby item. Indeed, the ASD group appeared to find this item one of the most difficult items. Although the rest of the item difficulties (and thus item ordering) were consistent between groups, it is interesting that the pat baby item in particular was so disordered for the group with ASD. Although reasons for this can only speculative at this point, it could be argued that the pat baby item was the only item in the battery that had a particularly symbolic as well as social aspect to it. Given deficits in ASD in both symbolic play and social cognition (e.g., Hobson et al. 2009), it is perhaps not surprising that this item in particular was specifically more difficult for the participants who developed ASD. Specific deficits in imitation of symbolic actions using symbolic toys may amount to a separate construct to be examined in future research using these analytic techniques and a wider range of imitation items; for purposes of the current study, however, we decided to remove the pat baby item given that it appeared to be specifically biased against the ASD group in relation to the rest of the imitation scale. In this way we were able to ensure that our measure of imitation as a whole was invariant across groups while still allowing for overall differences in imitation ability estimates. This is to say that the final set of scale items functioned as a scale in the same way over time for typically developing children as it did for all the other groups – a critically important measurement assumption for comparing groups on a single construct of interest.

It could be argued that the three items that were removed from the scale all had similar motor demands of “patting” something and thus may have comprised a separate distinct factor of sorts. Although there does seem to be some superficial similarity between these three items, it is important to consider that each item showed a distinct form of bias. Even the pat table and pat cheeks item did not show the same pattern of bias over time given that one showed linear bias (pat table) and the other showed curvilinear bias (pat cheeks). If the three “patting” items that were removed did comprise a single unitary scale unto themselves, we would expect to see a similar bias in all three by definition. That there was no such uniformity to the bias as revealed by DIF analyses suggests that these items were better left out of further analyses altogether.

Regarding the final 7 item, 11-step scale that we used in our analyses of group differences, it is informative to consider the item ordering of the scale as a potential window on the development of imitation in general. As seen in Table 4, the easier items on the scale tended to be object or gestural items with oral-motor items being the more difficult items. It could be that such an ordering reflects aspects of imitation that develop sequentially over time, where relatively more instrumental imitation with objects (i.e., bang block) are

developmentally prior to the imitation of less meaningful, but perhaps more symbolic or socially relevant gestures (e.g., clap hands), which, in turn, are developmentally prior to more sophisticated oral-motor imitation. Indeed, this ordering is roughly consistent with much of prior literature in typical development suggesting that actions on objects are mastered at younger ages than manual gestures or oral/facial actions (e.g., Abravanel et al., 1975; McCall et al., 1976; Uzgiris & Hunt, 1975). For example, in the developmental assessment scale proposed by Uzgiris and Hunt (1975), actions such as "patting an object" (p.182) developmentally preceded imitation of gestures such as "opening and closing the fisted hand" (p.183), with the most developmentally difficult items being unfamiliar actions on which the infant is unable to self-monitor performance such as "opening and closing the mouth" (p. 184). Although the current study was not designed specifically to exhaustively test such a developmental progression, it is clear that the use of a Rasch model for identifying item difficulty provides useful information for validating and exploring such developmental models. To the extent that the ASD group did not exhibit any profound disordering of item difficulty estimates in comparison to the typically developing group (as would have been revealed by DIF analyses), it could be argued that this is evidence that imitation in autism follows the same hierarchy of skill difficulty as it does for typically developing children. Moreover, although much of the prior imitation literature in autism that has suggested children with ASD have specific deficits in discrete types of imitation such as manual gestures (e.g. Stone et al., 1997), in light of the results of the Rasch analyses and the unidimensionality of our scale, a better interpretation seems to be that such specific ASD deficits simply reflect a general imitation delay. If children with ASD had specific deficits in a particular type of imitation, we should have found evidence for significant DIF as a function of group among those items, with consistent and significant bias against the ASD group over time.

Group Differences in Imitation and Changes over Time

Our second aim, of evaluating group differences in imitation ability over time, revealed that the group of infants who were identified at 36 months with ASD exhibited significantly poorer imitation skills than low-risk typical infants across all time points, but showed a similar developmental trajectory through 24 months. The finding of an overall imitation deficit in the ASD group as early as 12 months suggests that imitation is disrupted early in ASD. Of particular interest, however, is that the group with ASD did exhibit the same amount of growth as the other groups, showing an increase in ability of over 2 logits between 12 and 24 months. In terms of actual response probabilities, this amounts to an 88% probability at 24 months of passing an item that was passed with only 50% probability at 12 months. As an example using an actual item – the partial-pass scale step for tongue click – the estimated probability increased from 2.39% (SD=2.59%) at 12 months to 20.43% (SD=20.85%) at 24 months in the ASD group. A similarly increasing, albeit higher overall set of probabilities was seen for the same item in the lowrisk typical group over time from 5.02% (SD=6.48%) at 12 months to 33.80% (SD=26.42%) at 24 months. As such, the ASD group appears simply to be delayed in the development of imitation, and does not show quantitatively or qualitatively different developmental trajectories compared to other groups. Indeed, this may be taken as an extension of and parallel to what was found with the DIF analysis: the same developmental progression applies equally to the ASD and comparison groups, despite overall differences in ability.

With respect to the question of specificity for an imitation delay in ASD, it is important to note that the observed delay in the ASD group was only significant when compared to the low-risk typically developing group. When compared to the high-risk typically developing group the observed delay was only marginally significant and when compared to the other delays group of high-risk infants, the delay was not significantly different at all. As such, the

specificity of the imitation delay we observed in the ASD group was poor, at least in relation to the other delays group. This result is in contrast to a number of other cross-sectional results in the autism literature that report significant differences in imitation between those with ASD and those with developmental delays (e.g., Rogers et al., 2003; Stone et al., 1997). Such studies, however, have used comparison groups that were qualitatively different than the one used in our current study in that they did not have the same genetic liability for developing ASD. It could be that to the extent that the other delays group in our own study represents a group with subclinical autism symptoms, the imitative delay observed in our study is part of this broader autism phenotype instead of ASD itself. The use of an additional comparison group of developmentally delayed children who are not at increased risk for developing ASD would help to clarify this issue. Nevertheless, the lack of specificity of an imitation delay in ASD when compared to the high-risk other delays groups suggests that early imitation deficits by themselves do not cause a developmental cascade of symptoms resulting in ASD, even in a group of infants at a presumably similar genetic risk for developing ASD. In fact, our use of a comparison group of infants at similar genetic risk for developing autism who developed other non-ASD delays provides a more stringent test of specificity than would other comparison groups with delays; it controls for genetic liability as well as the presence of a degree of developmental delay. Thus, the presence of an early deficit in imitation does not appear to be sufficient for causing ASD; rather, it may best be thought of as an early associated component of an emerging disorder that simultaneously disrupts a variety of socio-cognitive skills.

Although we found no relationship between the development of fine motor ability over time and the development of imitation ability, we did find significant relationships between two separate measures of language ability and imitation. Increases over time in both the Mullen expressive language age equivalent scores and raw vocabulary scores were significantly related to imitation ability over time, and this relationship was the same for all groups. Moreover, the main effect for group differences in imitation disappeared after inclusion of either language variable which suggests that the group differences may be largely accounted for by language ability, further casting doubt on the specificity of imitation deficits in ASD. A similar relationship was found between examiner ratings of social engagement and imitation over time, again with no group differences in the relationship. These findings are important for two reasons: a) they extend prior research that has shown a link between language and imitation by documenting the longitudinal nature of this relationship, and b) they suggest that interrelationships between socio-cognitive measures and imitation are similar between typically developing children and those who develop ASD – a finding that again parallels and extends our own results from the Rasch model that the developmental sequence *within* imitation is similar between groups. In addition, our analyses suggest that the relationship between imitation and social engagement was relatively independent of the relationship between imitation and language in that expressive language remained associated (albeit marginally) with imitation ability after accounting for social engagement. Unfortunately, however, our findings do not disentangle the directionality of such relationships. The degree to which imitation ability actually drives language development and social engagement could best be answered with larger studies with additional longitudinal measurement points that would allow for the use of models such as cross-lagged structural equation models where various causal pathways could be explored in more detail and directly compared.

Regarding future research on the development of imitation, it is also important to note that although the evaluation of the imitation items in the current study using the Rasch model did yield a very high overall item reliability, the person reliability was much lower. The item reliability reflects the fact that the estimates of item difficulties, and thus the item ordering of the entire scale had a high degree of precision, made with relatively little error. This is

perhaps not surprising given the fact that item estimates were based on several hundred observations over time. The person reliability of the scale, however, was much lower and suggests that ability estimates contained a fair amount of error. This is a direct function of the fact that only 7 items with 11 total scale steps were used. Increasing the scale length to at least 26 scale steps (as suggested by the Spearman-Brown Prophecy formula) would clearly be a critical improvement for future research aimed at examining imitative development over time. Moreover, given the distribution of item difficulties in our 7-item, 11-step measure as seen in Table 4, it would be particularly useful to develop and test items that are relatively easy in order to better differentiate imitation skill at very early ages. The inclusion of a wider range of imitation items, including imitation of symbolic actions or even imitation of failed intentions, would also help to better differentiate the development of imitation ability; indeed, the inclusion of a greater variety of types of imitation might yet reveal important distinct dimensions of imitation ability and associated specific group differences, including divergent trajectories in ASD. Despite these limitations, however, it is instructive to note that our final scale did appear to capture the range of performance between 12 and 24 months in all groups rather well. As seen in Table 5, there were no clear floor or ceiling effects in any of the groups which suggests that the scale was well calibrated for assessing the development of imitation during the second year of life. Moreover, the rates of imitation we found with our imitation scale are commensurate with prior reports in the literature. For instance, Abravanel et al. (1976) reported an average of 55.41% correct in 18 month-olds using a similar 22-item scale. Likewise, Stone et al. (1997) reported an average of 47.5% correct in 18 month old typically developing children on a similar 10-item imitation scale.

In summary, this study has extended prior research on the development of imitation in both typical development and in autism by being the first to use a large longitudinal sample to develop and evaluate a measure of imitation as it changes over time prior to age 24 months. Using a single analytic framework, we were able to accomplish two important aims simultaneously: a) evaluate and establish a measure that is structurally invariant over time and invariant between groups, and b) test specific hypotheses regarding group differences in the overall measured construct of imitation ability and relationships to other time-varying variables. The fact that the *development* of imitation was significantly related to language and social behavior over time, as well as the fact that children with autism exhibited delayed imitation ability as early as 12 months suggests that future research aimed at measuring imitation and other socio-cognitive skills in even greater detail may help to illuminate important developmental dynamics in both the onset of autism as well as typical development in general.

Acknowledgments

This research was supported by NIMH/NIH Studies to Advance Autism Research and Treatment (STAART) program grant number U54-MH-068172 awarded to M. Sigman, P.I, grant number R01 MH068398 from the National Institute of Mental Health awarded to S. Ozonoff, P.I., and grants from the National Association for Autism Research (NAAR), from Cure Autism Now (CAN), and from the Medical Investigation of Neurodevelopmental Disorders Institute (M.I.N.D) awarded to S. J. Rogers.

The authors wish to thank Whitney Mattson for his creation of training materials for coding procedures and reliability, and wish to thank Jeffrey Martinez for his unflagging efforts in transferring and coding hundreds of video recordings of testing sessions.

References

Abravanel E, Levan-Goldschmidt E, Stevenson MB. Action imitation: The early phase of infancy. Child Development. 1976; 47:1032–1044. [PubMed: 1001087]

American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. Fourth Edition. American Psychiatric Association; Washington, DC: 2000. Text Revision

Bates, D.; Maechler, M. [August 31, 2009] Package'lme4'. Available from http://cran.r-project.org

Berument SK, Rutter M, Lord C, Pickles A, Bailey A. Autism screening questionnaire: Diagnostic validity. British Journal of Psychiatry. 1999; 175:444–451. [PubMed: 10789276]

Bond, TG.; Fox, CM. Applying the Rasch model: Fundamental measurement in the human sciences. 2. New Jersey: Lawrence Erlbaum Associates, Inc; 2007.

Charman T, Swettenham J, Baron-Cohen S, Cox A, Baird G, Drew A. Infants with autism: An investigation of empathy, pretend play, joint attention, and imitation. Developmental Psychology. 1997; 33:781–789. [PubMed: 9300211]

Corsello C, Hus V, Pickles A, Risi S, Cook EH, Leventhal BL, et al. Between a ROC and a hard place: decision making and making decisions about using the SCQ. Journal of Child Psychology and Psychiatry. 2007; 48:932–940. [PubMed: 17714378]

Dahlgren SO, Gillberg C. Symptoms in the first two years of life. European Archives of Psychiatry and Neurological Sciences. 1998; 238:169–174. [PubMed: 2721535]

Dapretto M, Davies MS, Pfeifer JH, Scott AA, Sigman M, Bookheimer SY, et al. Understanding emotions in others: mirror neuron dysfunction in children with autism spectrum disorders. Nature Neuroscience. 2006; 9:28–30.

DeMeyer MK, Alpern GD, Barton S, DeMeyer WE, Churchill DW, Hingtgen JN, et al. Imitation in autistic, early schizophrenic, and nonpsychotic subnormal children. Journal of Autism and Childhood Schizophrenia. 1972; 2:264–287. [PubMed: 4678763]

Doran H, Bates D, Bliese P, Dowling M. Estimating the multilevel Rasch model: With the lme4 Package. Journal of Statistical Software. 2007; 20:1–18.

Dunst, CJ. A Clinical and Educational Manual for Use with the Uzgiris and Hunt Scales of Infant Psychological Development. Baltimore: University Park Press; 1980.

Elsner B. Infants' imitation of goal-directed actions: The role of movements and action effects. Acta Psychologica. 2007; 124:44–59. [PubMed: 17078915]

Fenson, L.; Dale, PS.; Reznick, JS.; Thal, D.; Bates, E.; Hartung, JP.; Pethick, S.; Reilly, JS. The MacArthur Communicative Development Inventories: User's guide and technical manual. San Diego, CA: Singular Publishing Group; 1993.

Gotham K, Pickles A, Lord C. Standardizing ADOS scores for a measure of severity in autism spectrum disorders. Journal of Autism and Developmental Disorders. 2009; 39:693–705. [PubMed: 19082876]

Hauf B, Elsner P, Aschersleben G. The role of action effects in infants' action control. Psychological Research. 2004; 68:115–125. [PubMed: 14652756]

Hobson RP, Lee A. Imitation and identification in autism. Journal of Child Psychology & Psychiatry. 1999; 40:649–659. [PubMed: 10357170]

Hobson RP, Lee A, Hobson JA. Qualities of symbolic play among children with autism: a social-developmental perspective. Journal of Autism and Developmental Disorders. 2009; 39:12–22. [PubMed: 18509752]

Jarrold C, Brock J. To match or not to match? Methodological issues in autism-related research. Journal of Autism and Developmental Disorders. 2004; 34:81–86. [PubMed: 15098961]

Jones S. Imitation in infancy: the development of imitation. Psychological Science. 2007; 18:593–599. [PubMed: 17614867]

Kamata, A. One-parameter hierarchical generalized linear logistic model: An application of HGLM to IRT; Paper presented at the annual meeting of American Educational Research Association; San Diego, CA. 1998 Apr.

Kamata A. Item analysis by the hierarchical generalized linear model. Journal of Educational Measurement. 2001; 38:79–93.

Killen M, Uzgiris IC. Imitation of actions with objects: The role of social meaning. Journal of Genetic Psychology. 1981; 138:219–229.

Leppuscu, S. DIF detection in HLM; Paper presented at the AERA Annual Meeting; April 2002; New Orleans.

Linacre, JM. [August 31, 2009] A user's guide to Winsteps Mimistep Rasch-model computer programs: Program manual 3.68.0. Available from http://www.winsteps.com/winman/index.htm

Lord, C.; Rutter, M.; DiLavore, P.; Risi, S. Autism Diagnostic Ovservation Schedule. Los Angeles, CA: Western Psychological Services; 1999.

Mars AE, Mauk JE, Dowrick PW. Symptoms of pervasive developmental disorders as observed in prediagnostic home videos of infants and toddlers. Journal of Pediatrics. 1998; 132:500–504. [PubMed: 9544908]

Masur, EF. Vocal and action imitation by infants and toddlers during dyadic interactions: Development, causes, and consequences. In: Rogers, SJ.; Williams, JHG., editors. Imitation and the social mind: autism and typical development. New York: Guilford Press; 2008. p. 27-47.

Masur EF, Rodemaker JE. Mothers' and infants' spontaneous vocal, verbal, and action imitation during the second year. Merrill-Palmer Quarterly. 1999; 45:392–412.

McCall RB, Parke RD, Kavanaugh RD. Imitation of live and televised models by children one to three years of age. Monographs of the Society for Research in Child Development. 1977; 42(5, Serial No. 173)

McDuffie A, Turner L, Stone W, Yoder P, Wolery M, Ulman T. Developmental correlates of different types of motor imitation in young children with autism spectrum disorder. Journal of Autism and Developmental Disorders. 2007; 37:401–412. [PubMed: 16900404]

McGuigan N, Whiten A, Flynn E, Horner V. Imitation of causally-opaque versus causally-transparent tool use by 3- and 5-year-old children. Cognitive Development. 2007; 22:353–364.

Mullen, E. Mullen Scales of Early Learning. Circle Pines, MN: America Guidance Service; 1995.

Ornitz EM, Guthrie D, Farley AH. The early development of autistic children. Journal of Autism and Childhood Schizophrenia. 1977; 7:207–229. [PubMed: 71292]

Ozonoff S, Iosif A, Baguio F, Cook IC, Hill MM, Hutman T, et al. A prospective study of the emergence of early behavioral signs of autism. Journal of the American Academy of Child & Adolescent Psychiatry. 2010; 49:258–269.

Pastor DA, Beretvas SN. Longitudinal Rasch modeling in the context of psychotherapy outcomes assessment. Applied Psychological Measurement. 2006; 30:100–120.

Robins DL, Fein D, Barton ML, Green JA. The modified checklist for autism in toddlers: An initial study investigating the early detection of autism and pervasive developmental disorders. Journal of Autism and Developmental Disorders. 2001; 31:131–144. [PubMed: 11450812]

Rogers SJ, Pennington BF. A theoretical approach to the deficits in infantile autism. Development and Psychopathology. 1991; 3:137–162.

Rogers SJ, Hepburn SL, Stackhouse T, Wehner E. Imitation performance in toddlers with autism and those with other developmental disorders. Journal of Child Psychology and Psychiatry. 2003; 44:763–781. [PubMed: 12831120]

Rogers SJ, Young GS, Cook I, Giolzetti A, Ozonoff S. Imitating actions on objects in early-onset and regressive autism: effects and implications of task characeristics on performance. Developmental Psychopathology. 2010; 22:71–85.

Stone WL, Ousely OY, Littleford CD. Motor imitation in young children with autism: What's the object? Journal of Abnormal Child Psychology. 1997; 25:475–485. [PubMed: 9468108]

Tager-Flusberg H. Strategies for conducting research on language in autism. Journal of Autism and Developmental Disorders. 2004; 34:75–80. [PubMed: 15098960]

Uzgiris, IC.; Hunt, JM. Assessment in infancy: Ordinal scales of psychological development. Urbana: University of Illinois Press; 1975.

Watson LR, Barenek GT, Crais ER, Reznick JS, Dykstra J, Perryman T. The first year inventory: Retrospective parent responses to a questionnaire designed to identify one-year olds at risk for autism. Journal of Autism and Developmental Disorders. 2007; 37:49–61. [PubMed: 17219058]

Williams JHG, Whiten A, Singh T. A systematic review of action imitation in autistic spectrum disorder. Journal of Autism and Developmental Disorders. 2004; 34:285–299. [PubMed: 15264497]

Williams JH, Waiter GD, Gilchrist A, Perrett DI, Murray AD, Whiten A. Neural mechanisms of imitation and 'mirror neuron' functioning in autistic spectrum disorder. Neuropsychologia. 2006; 44:610–621. [PubMed: 16140346]

Williams NJ, Beretvas N. DIF identification using HGLM for polytomous items. Applied psychological measurement. 2006; 30:22–42.

Wright BD. Rasch factor analysis. Rasch Measurement Transactions. 1994; 8:348–349.

Zwaigenbaum L, Bryson S, Rogers T, Roberts W, Brian J, Szatmari P. Behavioral manifestations of autism in the first year of life. International Journal of Developmental Neuroscience. 2005; 23:143–152. [PubMed: 15749241]

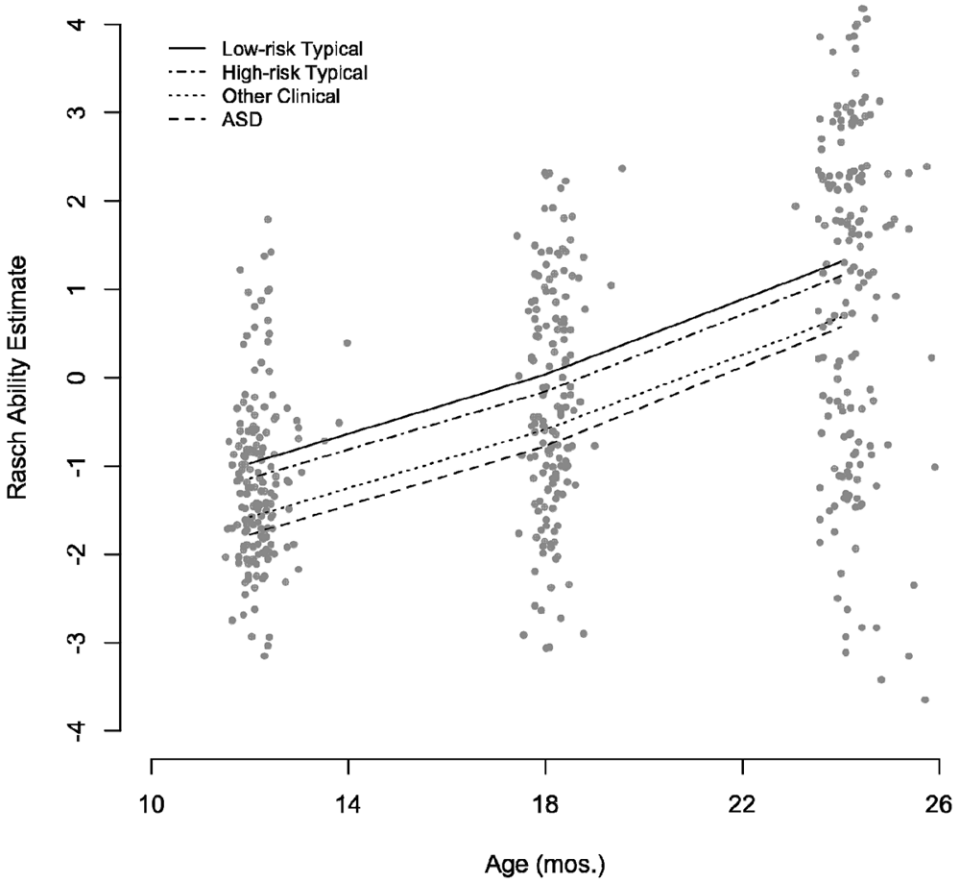

Figure 1.
Longitudinal Rasch estimates of person abilities and growth model trajectories.

Table 1

Diagnostic outcome definitions

ASD	Other delays	Typical
• Above ASD cutoff of ADOS *and* • Meets DSM-IV criteria for Autistic Disorder or PDD-NOS	• Mullen Early Learning Composite < 78 *or* • Multiple Mullen subtests <= 1.5 sd below mean *or* • ADOS communication and social total within 3 points of ASD cutoff *and* • Does not meet DSM-IV criteria for Autistic Disorder or PDD-NOS	• Mullen Early Learning Composite > 78 *and* • No more than one Mullen subtest <= 1.5 sd below mean *and* • No Mullen subtest <= 2 sd below mean *and* • Four or more points below ASD cutoff of ADOS *and* • Does not meet DSM-IV criteria for Autistic Disorder, PDD-NOS, or any other developmental delay

Table 2

Sample Characteristics at 36 month Outcome[a]

	Autism/ASD (n=24)	Other Delays (n=43)	High-Risk Typical (n=90)	Low-Risk Typical (n=75)
Gender ratio (male:female)	7:1	1.5:1	0.7:1	1.3:1
Income level[b]	3.76 (1.79)	4.17 (1.47)	4.12 (1.61)	4.30 (1.67)
Minority status[c]	38.1%	43.8%	35.9%	28.6%
ADOS Severity Score	6.57 (2.02)	1.93 (1.08)	1.20 (0.50)	1.09 (0.41)
SCQ total score	12.50 (6.14)	5.03 (4.58)	3.28 (3.51)	3.94 (2.84)
Mullen Early Learning Composite	74.81 (24.12)	93.63 (17.60)	114.61 (14.23)	114.79 (13.94)
Mullen Expressive Language Age Equivalent	29.04 (9.99)	32.45 (7.73)	38.81 (5.67)	38.69 (5.90)
Mullen Receptive Language Age Equivalent	26.21 (11.22)	30.05 (9.37)	37.56 (6.08)	37.36 (7.20)

[a] All numbers in parentheses are standard deviations.

[b] Income level measured on 6 point scale where 4=$75k to $100k.

[c] Minority status measured as Hispanic and/or non-Caucasian.

Table 3

Imitation Items and scoring criteria

Item	Perfect Pass	Partial Pass
Pat table	Palm side of 2 flat hands hit table surface repeatedly and synchronously	Slaps, pats, or hits table with one or both hands asynchronously or without repetition.
Bang block	Stick held radially, strikes block top repeatedly in a vertical trajectory	Contacts block with mallet with alternative grip, on different side or at an angle without repetition.
Clap hands	Palm side of flat hands contact each other repeatedly with both hands moving toward midline.	Hands clap but not flat, not repeated, or one hand stationary.
Open/close hands	Two hands above table surface, palms facing forward, open and close simultaneously and repeatedly	Hands open and close but asynchronously or without repetition, not facing outward, or only one at a time with arm resting on table.
Pat baby	One hand flat, palm down, pats doll repeatedly.	Pats doll without repetition, uses two hands or fist, or misses doll.
Open/shut mouth	Mouth repeatedly opens and shuts so top lip contacts bottom lip.	Some open and shut motion of mouth, but not repeated, or without complete shutting.
Pat puffed cheeks	Two hands, flat, pat cheeks synchronously and repeatedly.	Only one hand used or with widely splayed or flexed fingers, pats other part of face, or is asynchronous or not repeated.
Wiggle tongue	Tongue protrudes from mouth and moves laterally repeatedly with face relatively relaxed	Protrusion of tongue but without lateral movement or repetition or with extraneous tension and movement in rest of face.
Raspberry	With lips together, tongue not visible and rest of face relatively relaxed, lips vibrate together making the ‘raspberry’ sound.	Audible attempt at sound but with tongue visible, mouth slightly open, or extraneous tension in rest of face.
Tongue click	With mouth open and rest of face relaxed, repeated click sound made by tongue against roof of mouth.	Movement of mouth and jaw without involvement of tongue, extraneous tension in rest of face, or no repetition of click.

Table 4

Final scale item estimates, thresholds, and fit statistics

Item	Scale step	Difficulty	Item Difficulty (fixed effect)[a]	Step Threshold (random effect)	Outfit	Infit
Bang block	0-1	-3.14	-2.12	-1.02	0.67	0.88
Bang block	1-2	-0.68	-2.12	1.48	1.09	1.07
Wiggle tongue	0-1	-0.24	1.29	-1.53	0.90	0.92
Open-shut mouth[b]	0-1	0.32	0.37	0.06	0.84	0.88
Clap hands[b]	0-1	0.40	0.48	0.08	0.78	0.84
Raspberry	0-1	0.61	1.56	-0.95	0.85	0.88
Open-close hands[b]	0-1	0.66	0.75	0.09	0.87	0.91
Tongue click	0-1	0.92	1.93	-1.10	0.75	0.83
Raspberry	1-2	2.23	1.56	0.67	0.85	0.96
Wiggle tongue	1-2	2.57	1.29	1.29	0.80	0.97
Tongue click	1-2	2.61	1.93	0.68	0.68	0.89

[a] Item parameters multiplied by (-1) to reflect difficulty, given that actual parameter estimates reflect probability of correct answer, or ‘easiness’.

[b] Item scores collapsed to pass-fail dichotomy.

Table 5

Imitation scores and Covariates for Groups by Age

		Age		
Measure	**Group**	**12 months**	**18 months**	**24 months**
Imitation Score (% correct)	ASD	21.28% (9.64)[a]	25.18% (10.14)	49.62% (26.67)
	Other Delays	22.07% (12.51)	35.62% (18.73)	50.97% (25.42)
	High-risk Typical	27.39% (11.16)	41.86% (19.17)	58.73% (28.10)
	Low-risk Typical	29.02% (14.62)	44.83% (20.53)	63.27% (25.63)
Fine Motor age equivalent	ASD	13.33 (1.76)	16.58 (1.62)	20.57 (2.38)
	Other Delays	15.07 (1.86)	18.86 (1.81)	23.12 (3.55)
	High-risk Typical	15.28 (1.60)	19.49 (2.20)	24.75 (3.01)
	Low-risk Typical	15.34 (1.31)	19.57 (2.27)	25.63 (2.62)
Expressive Language age equivalent	ASD	10.13 (2.66)	14.08 (5.28)	17.12 (7.41)
	Other Delays	11.87 (2.22)	17.37 (3.83)	22.22 (4.70)
	High-risk Typical	12.79 (2.60)	18.11 (4.08)	25.35 (5.46)
	Low-risk Typical	12.51 (2.19)	18.09 (3.34)	25.52 (4.01)
MacArthur CDI Vocabulary (# words)	ASD	3.83 (7.7)	41.27 (76.8)	45.08 (47.5)
	Other Delays	5.28 (8.6)	52.16 (64.7)	180.75 (136.1)
	High-risk Typical	10.62 (12.0)	80.77 (79.3)	271.85 (170.1)
	Low-risk Typical	8.18 (8.9)	96.83 (110.7)	343.24 (200.0)
Social Engagement rating (out of 9)	ASD	8.09 (1.81)	6.00 (1.49)	6.47 (1.64)
	Other Delays	8.35 (0.93)	7.43 (1.09)	7.94 (1.20)
	High-risk Typical	8.17 (1.26)	8.20 (1.16)	8.33 (1.30)
	Low-risk Typical	8.67 (0.71)	8.69 (0.60)	8.75 (0.50)

[a]Standard deviations shown in parentheses.