Title

Towards the Improved Characterization of Minimally Verbal Children with Autism: Applications of Item Response Theory and Machine Learning Algorithms to Analyze Measures of Social Communication

Author

Schlink, Andrew Jeremy

Publication Date

2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Towards the Improved Characterization of Minimally Verbal Children with Autism: Applications of Item Response Theory and Machine Learning Algorithms to Analyze Measures of Social Communication

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Education

by

Andrew Jeremy Schlink

2022

ABSTRACT OF THE DISSERTATION

Towards the Improved Characterization of Minimally Verbal Children with Autism: Applications of Item Response Theory and Machine Learning Algorithms to Analyze Measures of Social Communication

by

Andrew Jeremy Schlink

Doctor of Philosophy in Education

University of California, Los Angeles, 2022

Professor Connie L. Kasari, Chair

Minimally verbal children are considered the enigmatic and, unfortunately, the neglected end of the autism spectrum. This subpopulation has likely garnered this title due to their exclusion from research studies, which has inevitably affected their evidence base. The paucity of proper measurement tools that sensitively and accurately assess behaviors has been one limiting factor towards the improved knowledge of these children. Short of creating and validating a new measurement tool for this subpopulation, this study took an alternative and more immediate approach: examine an existing social communication measure (ESCS) with repurposed quantitative methodologies, item response theory (IRT) and machine learning algorithms (CART and random forests). The final sample consisted of 453 minimally verbal children culled from four different intervention studies.

The IRT models analyzed the frequency of social communication gestures from the ESCS and returned an objective difficulty hierarchy regarding initiations of joint attention and behavior regulation gestures. The best-fitting and final model was a zero-inflated negative binomial model (ZINBM) which determined that joint attention gestures were, on average, more difficult than behavior regulation gestures. Joint attentional shows and gives were especially tough, and behavior regulation reaches were the easiest gestures for this sample. The ZINBM separately modeled children with some gestures and children who did not present with any gestures and determined that behavior regulation reaches and gives were likely the first gestures a child will eventually exhibit among children with no gestures.

Classification and regression trees (CART) were used to understand the clinical meaning behind frequencies of social communication gestures. Influential cut points were identified by the recursive partitioning algorithm of CART and determined which frequencies were able to classify children into more or less robust language outcomes at baseline. On average, a single behavior regulation point was sufficient to classify children into more robust language outcomes. For many of the trees, responding to bids joint attention around one-third of the time or more was also predictive of more robust language outcomes. Variable importance was examined with a random forest algorithm, which matched the results from the classification trees.

Overall, this study demonstrated that the use of IRT and CART yielded additional information, beyond traditional scoring and analytic techniques, regarding the presentation of social communication gestures among minimally verbal children. This study also discussed the methodological contributions and potential future applications of IRT and CART within this field.

The dissertation of Andrew Jeremy Schlink is approved.

Li Cai

Catherine Lord Morrison

Peter C. Mundy

Connie L. Kasari, Committee Chair

University of California, Los Angeles

2022

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

The work presented here would not have been possible without the support of many people. First, thank you to my committee chair, Connie Kasari, Ph.D., for her indispensable mentorship over the years and for emphasizing the importance of methodologically sound and actionable autism research. Thank you to my committee members, Li Cai, Ph.D., Catherine Lord, Ph.D., and Peter Mundy, Ph.D., for their feedback and expertise. In addition, thank you to my colleagues for their help in theorization, Yon Soo Suh, who offered her spare time to help troubleshoot these statistical models, and to my close friends, also known as my second family, for their constant encouragement and positivity. To my mom and dad, words cannot describe my gratitude for all the opportunities you have provided me — this one is for you. Lastly, the author would like to thank the children and families who participated in this study; this work would not have been possible without your commitment to research.

Curriculum Vitae

EDUCATION

In progress	**University of California, Los Angeles** Ph.D., Education, Concentration: Human Development and Psychology Emphasis: Research Methods
June 2022	**University of California, Los Angeles** Certificate in Advanced Quantitative Methodology in Educational Research
June 2018	**University of California, Los Angeles** M.A., Education, Concentration: Human Development and Psychology
May 2013	**Johns Hopkins University** B.A., Psychology, Minor: Entrepreneurship and Management

RESEARCH POSITIONS

September 2017-Present **Graduate Student Researcher,** University of California, Los Angeles

PUBLICATIONS

Schlink, A., Williams, J., Pizzano, M., Gulsrud, A., & Kasari, C. (2022). Parenting stress in caregiver-mediated interventions for toddlers with autism: An application of quantile regression mixed models. *Autism Research*, 15(2), 353– 365.

Toolan, C., Holbrook, A., **Schlink, A.,** Shire, S., Brady, N., & Kasari, C. (2022). Using the Clinical Global Impression scale to assess social communication change in minimally verbal children with autism spectrum disorder. *Autism Research*, 15(2), 284– 295.

FELLOWSHIPS/AWARDS/HONORS

UCLA 2020	**Graduate Summer Research Mentorship award**
UCLA 2020	**Financial Aid Management award**
UCLA 2019-2020	**Graduate Research Mentorship award**
UCLA 2018	**Graduate Summer Research Mentorship award**
UCLA 2017	**Louise Tyler Fellowship award**

TEACHING EXPERIENCE

Winter 2021	**Teaching Assistant,** University of California, Los Angeles
	Course title: ED132: Autism, Mind, Brain, and Education
Winter 2020	Supervisor: Dr. Connie Kasari

SELECT POSTER PRESENTATIONS

A.J. Schlink & C. Kasari. (2021). *Creating Clinically Significant Cut Points within a Social Communication Measure: Application of Machine Learning to the Assessment of Minimally Verbal Children with ASD.* Poster presented at the annual meeting of the International Society for Autism Research.

A.J. Schlink & C. Kasari. (2020). *The Clinical Global Impressions-Severity (CGI-S) Scale: Measuring Severity of Social Communication Impairment Among Minimally Verbal Children with Autism*. Poster presented at the annual meeting of the International Society for Autism Research.

A. J. Schlink, M. Pizzano, L. Chiang, A. Gulsrud & C. Kasari. (2019). *Predictors of Parent Fidelity in a Caregiver-Mediated Intervention for Toddlers with ASD.* Poster presented at the annual meeting of the Gatlinburg Conference on Research and Theory in Intellectual and Developmental Disabilities, San Antonio, TX.

A. J. Schlink, L. M. Baczewski, K. Sterrett, C. Kasari & A. Gulsrud. (2018). *Targeting Joint Engagement in Toddlers with ASD: What Predicts Sustained Engagement at Study Follow-up?* Poster presented at the annual meeting of the International Society for Autism Research. Rotterdam, Netherlands.

A. J. Schlink, A. Sturm, M. Kuhfeld & C. Kasari. (2017). *Exploring the Relationship Between Autism Symptoms, Language Ability, and Externalizing Behaviors in Children with Autism*. Poster presented at the annual meeting of the International Society for Autism Research, San Francisco, CA.

Introduction

Autism Spectrum Disorder

As it is known today, the characteristics of Autism Spectrum Disorder (ASD) were first described by Leo Kanner in his seminal 1943 paper. Within his 11 case studies, Kanner documented the significant variability in impaired *social affective communication* — a hallmark of the disorder still recognized under the current Diagnostic and Statistical Manual of Mental Disorders (DSM–5). ASD is now recognized as a complex neurodevelopmental disorder marked by the diverse phenotypic expression of two core characteristics: social communication difficulties and restricted, repetitive behaviors (American Psychiatric Association, 2013). In addition to variability in core characteristics, individuals with ASD also exhibit a wide range of cognitive levels, ranging from very low in those with intellectual disability to well above the average in savants with unique talents (Maenner et al., 2020; Rimland, 1978).

Although once thought to be a rare neurodevelopmental disorder, ASD has experienced an increase in its population rate in recent years. In the 2018 report within the Autism and Developmental Disabilities Monitoring (ADDM) Network, the Center for Disease Control (CDC) determined that 1 in 59 children had a diagnosis of ASD (Baio et al., 2018); the findings from the most recent report in 2021 indicate the prevalence rate has increased to 1 in 44 (Maenner et al., 2021). At least some of this rise can be attributed to greater public awareness, broadening diagnostic criteria, and earlier detection of symptomatology in young children (Fombonne, 2009; Maenner et al., 2020; Matson & Kozlowski, 2011). These continually growing rates have contributed to increased attention for scientific research into the pathophysiology, etiology, and treatments for ASD.

Although understanding the direct cause of ASD still eludes current researchers, we have nevertheless learned much about how the disorder develops and whom it affects. We know that at some point between 12 and 36 months, the behavioral trajectory of children later diagnosed with ASD diverges from their neurotypical counterparts. During this onset period, deficits in social communication emerge, like reduced eye contact, a decline in smiling, and fewer vocalizations towards others (Ozonoff et al., 2010, 2011). From these departures in social engagement, a reliable diagnosis of ASD is achievable as early as age 2 (Johnson et al., 2007; Landa et al., 2007).

Language Development in ASD

Although not within the ASD diagnostic criteria, language delays are often associated with autism and are a potential byproduct of social communication impairment (Charman, 2003; Dawson et al., 2004; Mundy et al., 1986). These language delays can be detected by parents as early as 18 months (Mitchell et al., 2006) and are typically the basis for parents' seeking a diagnosis (Franchini et al., 2018; Lord, 2000). Like the core characteristics of autism, verbal language ability reflects the heterogeneity of the disorder — some individuals have fluent speech while others do not produce any words (Tager-Flusberg et al., 2005). Historically, it was estimated that up to 50% of children diagnosed with ASD would never develop functional speech (Rutter, 1978). However, we have now estimated that about half of all preschool-aged children with ASD *will* eventually develop age-appropriate language. The other half will experience language delays, with approximately 25-30% remaining nonverbal or minimally verbal (less than 20 spoken words) into late childhood (Anderson et al., 2007; Lord et al., 2004; NIH workshop, 2010; Tager-Flusberg & Kasari, 2013).

Minimally Verbal Children: What Do We Know?

Current estimates indicate that roughly 30% of individuals diagnosed with ASD will remain minimally verbal throughout their lifespan despite intensive intervention (National Research Council, 2001; Norrelgen et al., 2015; Pickles et al., 2014; Rose et al., 2016). Although their low expressive language output classifies minimally verbal children, they still exhibit a wide range of heterogeneity in other developmental domains (Tager-Flusberg & Kasari, 2013; Kasari et al., 2013).

We know that verbal IQ and nonverbal IQ levels are not identical with minimally verbal children. Some children with very low verbal IQ can have higher levels of nonverbal IQ than what their expressive language would suggest (Munson et al., 2008). Similarly, other minimally verbal children can have cognitive abilities within the average range of intelligence (Bal et al., 2016). Receptive language skills vary as well. Some minimally verbal children have very low receptive language skills, whereas others can maintain average levels of receptive language ability (Rapin et al., 2009). Thus, the heterogeneity of phenotypic expression that is the hallmark of ASD is still present within this subgroup of minimally verbal children.

Preverbal vs. Minimally Verbal

A distinction must be made between preverbal children and those who are minimally verbal. Most young children are considered 'preverbal'; that is, they are not talking yet (delayed in their development) but are expected to talk by school age. Those who remain limited in their verbal ability, defined as having no language or a repertoire of fewer than 20 words by age 5 or 6, are classified as "minimally verbal" and are at greater risk for poorer long-term outcomes and quality of life (Howlin et al., 2000; NIH workshop, 2010; Tager-Flusberg & Kasari, 2013). Differentiating these children at a young age is complex, which raises a pertinent issue in the field of knowing when and how to determine if a child is on a trajectory of language

development that will result in an optimal outcome — defined as speaking by school age, or on a slower trajectory that will reclassify the child as 'minimally verbal' (Ellis Weismer & Kover, 2015; Georgiades & Kasari, 2018; Kasari et al., 2018). We believe that to recognize these differential language trajectories earlier in a child's development, it is imperative first to understand children's skills thoroughly before an intervention.

Gains in Minimally Verbal Children

Previous data on minimally verbal children at school-age suggest that few develop fluent speech in the next several years (Koegel, 2000). However, there is growing evidence that language development is possible after age 5. Two follow-up studies indicated that a small fraction of children (7 out of 63 and 4 out of 120) learned to use functional language after five years of age (DeMyer et al., 1973; Rutter et al., 1967). Similarly, a review by Pickett and colleagues (2009) found that 167 individuals developed language between 5 and 7 years of age, with roughly one-third developing phrase speech. Again, in a larger review of 535 children who did not have functional speech at age 4, 70% developed phrase speech, and 47% developed fluent speech by age 8 (Wodka et al., 2013). Blended or adaptive interventions have also been proven to be efficacious with minimally verbal children. Among a sample of 61 minimally verbal children between 5 and 8 years of age, clinically and statistically significant improvements were made in spontaneous communicative utterances and novel words after 24 weeks if the intervention protocol began with a speech-generating device and a naturalistic developmental behavioral intervention (Kasari et al., 2014).

Therefore, although five years of age has typically been a cutoff for sufficient language gains, research suggests that the window for growth extends beyond this age and warrants additional research for this population. Not only is it of substantive importance to support

children in their language goals, but it has been shown that difficulty in expressive language is associated with maladaptive outcomes like aggression, inattention, and self-injury, which further emphasizes the importance of this research (Dominick et al. 2007; Hartley et al. 2008).

Hindrances in Minimally Verbal Research

Despite the justification for additional social communication research among minimally verbal children, some inherent barriers have hindered its progress.

Exclusion from Studies

First, less is known about this minimally verbal subgroup due to the systematic exclusion from research studies (Tager-Flusberg & Kasari, 2013; Kasari et al., 2013). In one review, only 50 out of 301 studies (17%) reported data on minimally verbal children (Russell et al., 2019). In another review, 105 out of 964 studies (11%) targeted these individuals (Jack & Pelphrey, 2017). In a large meta-analysis of 100,245 individuals from original autism research published in 2016, only 2% were considered non or minimally verbal, indicating a gap in the descriptive and behavior intervention literature for those with limited language ability (IACC, 2017).

Consequently, the Interagency Autism Coordinating Committee (IACC) has recognized the lack of representation of minimally verbal children in the literature and has highlighted the pressing need to study children with ASD with limited language abilities; specifically, those children who have not yet developed language into later childhood (IACC 2011, 2017; Tager-Flusberg & Kasari 2013).

Insufficient Assessment and Measurement

Another factor that has hindered the research of minimally verbal children has been assessment and measurement. Currently, the field lacks a consensus on which measures to use with this population, which has obscured the definition of what it is to be minimally verbal

across research studies (Tager-Flusberg & Kasari, 2013; Kasari et al., 2013; Koegel et al., 2020). Researchers may rely on different instruments and use different cutoffs to define and obtain their sample of minimally verbal children.

A systematic review that examined "nonverbal" and "minimally verbal" individuals within social communication intervention studies found that the assessment measures used to evaluate and define these samples lacked consistency. Across the 31 studies and 650 unique participants in this review, only four studies assessed participants with natural language samples, and only eight studies included more standardized parent reports on language (i.e., Vineland Adaptive Behavior Scales and the MacArthur–Bates Communicative Development Inventories), indicating that not even half of these studies used objective measures to evaluate language ability (Koegel et al., 2020). Although more standardized, these parent-reported measures are subject to biases. By implementing potentially biased, variable, and often unreliable assessment methods like non-standardized descriptions or informal parent/teacher reports, researchers are likely affecting sample characteristics and possibly contributing to the variability in outcome trajectories among this subpopulation.

The same review by Koegel et al. (2020) noticed considerable age and language level variability among participants classified as minimally verbal. Minimally verbal participants in this systematic review ranged from 2 to 16 years, a wide age range. Age is an essential variable in determining minimally verbal status as some young children could be considered preverbal and will eventually acquire verbal language. The number of words spoken exhibited the same variability. Some children could repeat sounds and syllables (Gevarter & Horan, 2018), while others had up to 51-75 words (Koegel et al., 2009). This wide range of verbal ability is observed in other studies as well. Some researchers have defined their sample of minimally verbal children

as having less than 20 different word roots on numerous language samples (Yoder & Stone, 2006). Still, others may rely solely on standardized tests to split the sample into two groups: no words versus 1-9 spontaneous words (Romski et al., 2010). Although these definitions are often well-defined in individual studies, the broad range of definitions likely impacts interpreting treatment effects and child outcomes across studies.

In response to these discrepancies among research studies, the National Institutes of Health (NIH) hosted a workshop in 2010 which brought together a broad array of scientists to discuss minimally verbal individuals with ASD. Defining minimally verbal status and identifying appropriate assessments for this population were the primary outcomes (Kasari et al., 2013; Tager-Flusberg & Kasari, 2013). Thus, the current study adopts these consensus guidelines to define minimally verbal as school-aged individuals who use fewer than 20 functional words, yet measures used to determine language were not specified (National Institute on Deafness and Other Communication Disorders, 2010).

Targeted Social Communication Interventions for Minimally Verbal Children

Research on communication-based interventions for minimally verbal children over the age of 5 is sparse. Additionally, the majority of early interventions often focus more on requesting language skills (e.g., "I want") than on other functions of speech such as commenting language or question asking (Kasari et al., 2013; Tager-Flusberg & Kasari, 2013; Pickett et al., 2009). However, theoretically, young children generally learn language in social interactions where commenting language can emerge; additionally, similar findings are noted for children with autism (Shih et al., 2021). Practically, it makes sense to target developmentally upstream predictors of later social-communicative language, such as joint attention and joint engagement.

What is the Role of Joint Attention in Expressive Language Development?

Social communication impairment can look different across the autism spectrum, so it is essential to distinguish which behaviors relate to spoken language. Before spoken words, children use several intentional communication acts with their caregiver, including eye contact, pointing, and showing to coordinate another person's attention to share the experience of an object or event (Mundy et al., 1994; Tomasello & Farrar, 1986). These communicative gestures, known as joint attention, are often intertwined with caregiver interactions. Parents' ability to follow-in on child activities promotes joint engagement between child and caregiver and facilitates these joint attention skills (Kasari et al., 2006, 2008, 2010). This joint engagement between child and caregiver offers a space where caregivers can introduce language during social interactions (Adamson et al., 2004, 2009; Tomasello, 1995).

The linkage between joint attention and subsequent language skills has been well documented in children with ASD and neurotypical children, suggesting that it is a requisite skill for language development (Charman, 2003; Loveland & Landry, 1986; Mundy et al., 1990; Tomasello & Farrar, 1986). Children with ASD who exhibit more developed joint attention skills concurrently exhibit more sophisticated language skills (Dawson et al., 2004; Mundy et al., 1986). Additionally, a study by Mundy and colleagues (1990) demonstrated that joint attentional gestures measured during the preschool years predicted language development one year later.

Given the concurrent and predictive properties of social communication behaviors on spoken language outcomes, communication-based interventionists have tailored therapies to target social communication behaviors like joint attention specifically (Schriebman et al., 2015; see also Kasari et al., 2006; Yoder & Stone, 2006). However, despite the focus on prelinguistic, joint attention skills in intervention studies, outcomes measures typically overlook joint attention

assays in lieu of more concrete developmental benchmarks like words gained (Koegel et al., 2020).

Treatment Outcomes

In a longitudinal RCT by Kasari and colleagues (2008), young children ages 3 and 4 years significantly improved expressive language gains one-year post-intervention if they initially received treatments that targeted joint attention or symbolic play. Researchers also examined whether initial language ability affected expressive language outcomes. At the entry of the study, children with more spontaneously initiated words showed significantly greater expressive language growth 1-year post-intervention. There was also a significant difference among intervention groups for children with low language ability (< 5 words), where the experimental group that targeted joint attention showed significantly more growth in language than the symbolic play intervention and control group.

A long-term 5-year follow-up of these same children indicated the continued growth and generalization of these social communication skills. Spoken vocabulary at the 5-year follow-up was predicted by numerous child variables at entry (when children were 3 and 4 years of age), like initiating more joint attention, demonstrating higher play levels, and receiving the experimental interventions (Kasari et al., 2012). On average, children in the joint attention and symbolic play experimental groups scored significantly higher on language measures at the 5-year follow-up compared to the control group, but the difference between treatment groups was not significant (Kasari et al., 2012). These results support that young children with language delays can improve language outcomes if an intervention specifically targets joint attention/joint engagement.

Measuring Change in Early Intervention Programs

For communication-based research studies, in particular, individual treatment response and intervention efficacy are partially determined by measures that capture expressive language abilities (Brignell et al., 2018; Kasari et al., 2006, 2008, 2012; Yoder & Stone, 2006). However, measuring discrete spoken language skills is not the only way to characterize these children, especially since there may not be enough variability in language to distinguish children from one another (DiStefano & Kasari 2016). Quality indicators of social communication like joint attention skills and joint engagement with a social partner emerge earlier in the developmental timeline of a child and are robust predictors of future language (Adamson et al., 2004; Luyster & Lord, 2009; Mundy et al., 1990; Shumway & Wetherby, 2009; Thurm et al., 2015; Wetherby, 2006). The measurement of these joint attention skills may be sensitive enough to detect differences in children's abilities that would otherwise not be apparent in measuring language alone.

However, currently, there is a dearth of assessments that can reliably measure small and meaningful differences in social communication (Bishop et al., 2019; Kasari et al., 2013; Lord & Jones, 2012). In particular, the lack of sensitive social communication measures that can be used among minimally verbal children has been a limiting factor in autism intervention research (McConachie et al., 2015; Kasari et al., 2013; Tager-Flusberg & Kasari, 2013). Nevertheless, the accurate measurement and monitoring of these crucial but often subtle social communication behaviors may indicate whether a child is on a positive language trajectory or not.

Proximal vs. Distal Outcomes. Common outcomes used to assess meaningful change among communication-based interventions are growth in joint attention skills and expressive/receptive language (Bolte & Diehl, 2013). The developmental breadth of these outcomes is measured by proximity, which describes how similar the outcome is to the

intervention's target. Outcome measures that align with intervention targets are considered proximal to the treatment. Alternatively, outcomes that assess behaviors beyond the primary target of the intervention are considered distal to the treatment (Gersten et al., 2005). For example, in a hypothetical communication-based intervention that aims to improve joint attention skills, outcomes that accurately assess those joint attention skills would be proximal to the treatment. Improved language, a developmentally downstream effect of joint attention improvement, is an example of a potential distal outcome.

Both types of measures are essential for communication-based interventions with minimally verbal children, and each has its strengths and weaknesses. Measures that assess joint attention skills may be proximal to the treatment target and thus a more sensitive indicator of quick behavior changes; however, these measures may not necessarily reflect the intervention protocol's goal, i.e., language. Alternatively, measures that assess language ability are typically the desired result among a minimally verbal population; however, language outcomes may not be sensitive to brief intervention periods and possibly only detected in long-term studies. Therefore, in intervention research, aligning proximal and distal measures results has been challenging but necessary to understand the trajectory of communication and language development.

Social Communication Measurement

The importance of targeting social communication behaviors like joint attention in early intervention is evident. First, it is associated with language development, but more importantly, the clearly defined set of behaviors is malleable and receptive to improvement within an intensive intervention context (Fuller & Kaiser; Kasari et al., 2014, 2016; Mundy, 2016). However, measuring these behaviors can be tricky, especially among minimally verbal children.

Why is it Difficult?

Some variables, such as social partner and context, can influence joint attention skills, thus altering the perceived quality of joint engagement between a child and another person. Interpreting data from various joint attention measures is complex too. Depending on features of the psychometric measure (rating scale vs. frequency count) and source of information (observation vs. parent/teacher-report), measures may be tapping into different aspects of the same social communication construct (Wetherby, 2006).

How Do Social Communication Measures Gather Information?

In a research context, joint attention skills can be objectively observed from a natural environment or within a structured, standardized interaction. These observations can be coded and interpreted by an expert who is blind to treatment assignments. Alternatively, social communication data can be derived from surveys, questionnaires, or interviews. Each measurement method has its strengths, weaknesses, and role in overall measurement theory.

Reported Information. Caregivers and teachers have insider knowledge of child behaviors in various contexts, but these person-reports may be vulnerable to bias and placebo effects (Anagnostou et al., 2015; Bolte and Diehl, 2013). In a double-blind, placebo-controlled study, Guastella and colleagues (2015) demonstrated that caregivers' opinion of treatment assignment could sway their report on behaviors. Therefore, it has been recommended that early intervention researchers discard parent-report measures as an indication of child behavioral response as they are inherently biased, and their results are largely uninterpretable (Jones et al., 2017; Sandbank et al., 2020).

Observed Behaviors. Measuring observed behaviors may be the most objective method to assess behaviors in an intervention study context (Masi et al., 2017). However, data may look different depending on where and with whom behaviors are assessed.

Observing social communication in a natural environment may elicit more organic joint attention skills on average, but controlling for the variability in joint attention opportunities across contexts and individuals may be difficult. A structured interaction may account for context and individual variability, but it may be challenging to obtain reliable data due to the contrived environment; children may have issues with compliance, attention, or rapport with researchers which may impact data quality (Wetherby, 2006).

A child's social partner could also contribute to variability in social communication skills too. The social partner may be someone familiar, like a caregiver or sibling, or someone foreign like a new interventionist, indicating that the presentation of true social communication abilities could depend on the social partner. Interpreting measures that gather data from these different sources in an intervention study should be disclosed.

Measurement with Minimally Verbal Children

Minimally verbal children, a population where it is typically challenging to perform assessments, exacerbate the difficulty of objective social communication measurement (Tager-Flusberg & Kasari, 2013). For children with limited language ability, the basic skills required to undergo standardized testing, such as understanding the testing protocol and maintaining sustained attention, may not be within their repertoire of skills (Tager-Flusberg, 1999).

Given how difficult it is to measure social communication skills objectively, it is not surprising that there is a paucity of objective social communication measures that can be used for children with minimal language abilities (Kasari et al., 2013). The dearth of measures that quantify social communication apart from other developmental variables like language has hindered autism intervention research (Bishop et al., 2019; Lord & Jones, 2012). A possible consequence of the dearth of social communication measures is the lack of definitively

efficacious interventions for minimally verbal children (Kasari et al., 2013; Rogers & Vismara, 2008; Tager-Flusberg & Kasari, 2013). Therefore, although social communication measures that assess joint attention skills and joint engagement are good indicators of language growth, measurement is difficult, tedious, and may not be very precise for these minimally verbal children.

Communication measures for children with ASD are typically reported using total scores from standardized assessments that measure language ability (e.g., Natural Language Sample and Mullen Scales of Early Learning) and social communication/joint attention (e.g., Communication and Symbolic Behavior Scales). Using total scores under a classical test theory framework may not be appropriate for use with minimally verbal children, as this scoring method may fail to capture individual variability in outcomes (Charman, 2003; Tager-Flusberg & Kasari, 2013). As a result, two recommendations have been proposed to help measure social communication among minimally verbal children: analyze item-level data from assessments and establish spoken language benchmarks corresponding to developmental changes (Tager-Flusberg et al., 2009). Both recommendations aim to understand the variability in social communication skills among minimally verbal children.

Methodological Advances in Measurement for Minimally Verbal Children

Currently, intervention researchers are in a predicament: interventions that improve the core ASD behaviors of social communication and, subsequently, language are hindered by insufficient methods to measure these behaviors accurately and sensitively among minimally verbal children (McConachie et al., 2015). This study recognizes the recommendations put forth by Tager-Flusberg et al. (2009) and employed two methodological analytic techniques to improve our understanding of the social communication skills of minimally verbal children.

Item-Level Analysis of Social Communication Data

In communication-based autism research, outcomes for children are commonly reported with total sum scores from standardized assessments (Matson et al., 2010). However, these overall sum scores may overlook significant variability that may differentiate one child from another. In minimally verbal children, this issue is exacerbated as total scores may be very low on social communication measures (i.e., exhibiting floor effects) and may not capture individual variability (Kasari et al., 2013). Additionally, the subtlety in the function of these gestures may be lost within total scores, i.e., differentiating between joint attention and requesting social communication domains. Within some assessments like the Communication and Symbolic Behavior Scales (CSBS; Wetherby & Prizant, 1998, 2002), these domains are combined — which may pose issues in understanding children's true skills as joint attention is typically more impaired than requesting skills.

Thus, in terms of characterizing the social communication skills of minimally verbal children, an item-level analysis of individual behaviors may be more appropriate (Abbeduto et al., 2011; Charman et al., 2003). Researchers have capitalized on analyzing item-level data from standardized assessments like the Autism Treatment Evaluation Checklist (Magiati et al., 2011) and Vineland Adaptive Behavior Scales-II (Norrelgen et al., 2014) to characterize language gains in intervention studies. Within a population where it is challenging to portray variability in skills, minimally verbal children may benefit from the item-level analysis of social communication gesture data, which may provide additional and more precise information.

Item Response Theory. Item response theory (IRT) refers to a framework of mathematical models that attempts to explain the relationship between individual responses on items of an assessment scale and a latent, unobservable trait (Baker, 2001). IRT was developed as an

alternative approach to remedy the shortcomings of classical test theory (CTT) sum scores, namely that a person's ability and difficulty cannot be separately modeled, and that scores may not be measured with equal precision across the population (Doostfatemeh et al., 2015; Embretson & Reise, 2000).

In CTT, observed scores on a measure are the direct product of a participant's true score (average of observed scores over infinite repeated testing) and error (difference between observed and true scores) (Kean & Reilly, 2014). Although scores derived from CTT are easily interpretable, they may not effectively differentiate participants based on their skill levels due to inherent measurement error. Specifically, as it relates to this study, CTT's primary disadvantages include sample dependence, which may exacerbate the generalizability of findings amongst a heterogeneous population. IRT avoids these pitfalls by analyzing individual items of an assessment, which are invariant to the group's ability; therefore, determining how difficult an item is can be derived from a group with either low or high ability. Similarly, ability estimation is independent of other participants' performance on the test and items used. These invariance features of IRT allow item difficulty and ability to be measured on the same scale, which may reduce measurement error and offer a purer estimate of the construct or symptom of interest (Fries et al., 2005; Kean et al., 2018).

The premise of IRT is that it operates under the assumptions that an underlying latent trait governs or explains an individual's response on a measure and that those test items are statistically independent. Although slightly more esoteric than CTT scores, IRT and its results have proven valuable in scale development and improved precision in psychological and various health conditions measurements (Cella et al., 2007; Edelen & Reeve, 2007). In this study, using

IRT models is appropriate as children fall on a spectrum that reflects the latent trait of social communication, as measured by a specific observational assessment.

One of the goals of this study is to use IRT to improve how we measure baseline social communication skills among minimally verbal children. Analyzing item-level data may effectively capture untapped variability among minimally verbal children, thus differentiating children more sensitively based on their social communication skills. For example, under CTT, we can already determine which participants have more social communication skills. However, an experimental measure assessed with an IRT model can provide a "yardstick" to determine which participants have more social communication skills than others. Specifically, we can determine which aspects of social communication skills (i.e., joint attention and behavior regulation gestures) are most challenging to exhibit among this sample of minimally verbal children.

Rasch Poisson Count Model. IRT models typically make use of dichotomous or polytomous data, as one would see in true/false or multiple-choice question formats. Various IRT models that handle these data can effectively model the relationship between a person's endorsement of an item, which usually indicates whether that question was answered correctly, and the person's overall location on a latent construct or ability. Therefore, these models lend themselves especially well for purely psychometric purposes (e.g., shorten a questionnaire or improve test validity). Although these IRT models may be suitable for a wide range of data, they may not accommodate frequency-based behavioral data often seen in autism research and, in particular, intervention studies.

To circumvent any potential issues of recoding count data into a dichotomous or a graded format, another suitable option is to employ the oldest one-parameter IRT model developed, the

Rasch Poisson Count Model (RPCM; Rasch, 1960/1980). Although the RPCM is arguably less popular than its Rasch model counterpart, RPCMs have been used to analyze tests of attention (Baghaei & Doebler, 2019), processing speed (Doebler & Holling, 2016), oral reading errors (Rasch, 1960), memory (Jendryczko et al., 2020), and physical activity (sit-ups) (Zhu & Safrit, 1993). In each of these examples, the same task or numerous simple tasks are given to a participant, and the aggregation of hits/misses is the metric for understanding performance. For assessments that have been historically analyzed using CTT, the RPCM is a logical progression into the IRT framework (Doebler & Holling, 2016). Therefore, one key advantage to RPCMs is that they may explain results at the item and person level more profoundly than CTT alone, all while maintaining the integrity of the initial scoring procedure. Within the context of behavioral assessments among minimally verbal children with ASD, small tasks are usually sequentially introduced, and the aggregation of the frequency of observed behaviors is a way to measure overall social communication ability. From a measurement standpoint, using RPCMs may retain crucial variability that may be essential to characterize behaviors further.

RPCMs belong to the family of Rasch models, where the frequency or count of errors/successes are modeled instead of the response to a specific item as seen in the common Rasch, 1, and 2 parameter models (Doebler et al., 2014; Jansen, 1994; Rasch, 1960). Nevertheless, although it has drawn less attention, RPCMs are not dissimilar from their ubiquitous Rasch counterpart. They are both unidimensional latent trait models (i.e., measure one ability), and they both accurately separate person and item parameters for objective comparisons of persons and items. Similar to other IRT models, the number of correctly solved tasks (i.e., demonstration of behaviors in this study) is directly related to the function of a person's ability and a specific item-easiness parameter (Masters & Wright, 1984). RPCMs assume a Poisson

distribution, where the probability of a number of events (hits or misses) in a fixed period is expressed if the average number of hits or misses is known (Baghaei & Doebler, 2019). The RPCM assumes that the distribution of responses for each item follows a Poisson distribution with the rate λ, a parameter representing both its mean and variance. Under this conditional distribution, the mean equals the variance $E(X|\theta) = Var(X|\theta)$. The location and spread of the conditional distribution of responses *X* are determined by the rate, which is dependent on difficulty and latent ability. Therefore, the RPCM links an item's difficulty deterministically since the same parameter determines both location and spread (Beisemann, 2022). Violation of this equidispersion assumption is not uncommon in social science data as the variance is often greater than the mean, which may motivate the use of *post-hoc* modifications to account for the skewness (Thall & Vail, 1990).

Negative binomial model. One method to handle potential overdispersion is the negative binomial regression model (NBRM), which adds an additional random effect term to handle the extra variation in the data by allowing the mean and variance to be different (Hilbe, 2011; Hung, 2012). Poisson models, as seen in the RPCM, are a particular case of the NBRM model. When overdispersion is present, NBRMs may yield reliably larger standard errors, which reflect the additional variance in the outcome measure and help guard against Type-1 errors (Atkins et al., 2013).

Zero-inflated Negative Binomial Model. Given that joint attention skills are difficult for minimally verbal children, there is a possibility that many children are not able to demonstrate these skills within an assessment at baseline, resulting in many zeros for certain variables. Negative binomial models can typically fit highly skewed data like this, including data that contain many zeros; however, when the stack of zeros is robust across variables and the

distribution of non-zero data is not a smooth extension from the zeros, alternative models such as zero-inflated extensions may be more appropriate (Lambert, 1992). The zero-inflated model allows for frequent zero-valued observations since it is based on a zero-inflated probability distribution. Also called two-part or mixture models, ZINBMs combine a logit distribution with a negative binomial distribution (Hilbe, 2014; Lambert, 1992).

Establishing Language Benchmarks

Social communication can be exceptionally subtle for minimally verbal children, so measuring more discrete behaviors like joint attention is necessary to portray this population's skill variability. These discrete joint attention skills serve as a proxy for social communication development, and may be a more sensitive measure than language alone (Kasari et al., 2012). However, it may be unclear how to derive clinically meaningful interpretations regarding broader social communication abilities from experimental measures that only assess discrete behavioral skills. Specifically, clinical trials have not demonstrated a comprehensive method to calibrate joint attention skills from an experimental measure to more standardized measures of language development.

Classification and Regression Trees. One method suited to help establish language benchmarks within joint attention skills is classification and regression trees (CART; Breiman et al., 1984). CART is a statistical learning technique that uses recursive partitioning to create subgroups from a final set of predictor variables to specify cutoff values within those predictors (Hastie et al., 2009). This method outlines a set of easy-to-follow rules to classify observations into mutually exclusive groups based on the combinations and interactions of the explanatory variables. Compared to linear models, recursive partitioning methods like classification trees can work well when interactions among data or non-linearities cannot be identified a priori. One

strength of CART is that it is conceptually simple and easy to interpret. As a result, its rule-based decisions have made CART popular in many fields, including biomedicine, business analytics, and healthcare (Carrizosa et al., 2021).

The field of autism research has benefited from the CART method and its extensions as it can be used to form subgroups or find associations within the complex, non-normal data often present within autism research. For example, autism researchers have utilized the CART method to identify subgroups of differential treatment response (Shih et al., 2016) and to identify the most important predictors of expressive language development in minimally verbal children with autism (Bal et al., 2020). Bal et al. (2020) highlight the utility of the CART analysis to determine which variables at age three best predicted categorical language outcomes (verbal or minimally verbal) at age nineteen. Fine motor skills were the most important predictor of language ability at nineteen, where the majority of children (31/40) with delayed fine motor scores (i.e., FM-T < 20) on the Mullen Scales of Early Learning remained minimally verbal. Initiations of joint attention were the second most important predictor. Of the remaining 46 children who had better fine motor skills (i.e., FM-T ≥ 20), all 13 children who imitated joint attention developed phrase speech at nineteen (Bal et al., 2020). CART effectively minimized the heterogeneity in autism phenotypic presentation by creating subgroups of children based on cut points in important predictors.

As ASD research continues to leverage the benefits of machine learning techniques, CART has not yet been applied to illuminate benchmarks within social communication measurement among minimally verbal children with autism. This study suggests we can understand how sensitive joint attention measurements relate to broader language outcomes by applying CART across these proximal and distal social communication measures. Therefore, the

impetus for the machine learning algorithms was to close this interpretation gap between proximal and distal results so that information may be consolidated across multiple and crucial measures, thus, giving a more comprehensive characterization of abilities within this minimally verbal population.

Study Aims

Improving language is vital for minimally verbal children, given the prognosis for later optimal outcomes (Anderson et al., 2007; Nordin & Gillberg, 1998). Although interventions for this population that aim to improve social communication are growing, the number of appropriate social communication measures lags behind (Anagnostou et al., 2015; McConachie et al., 2015; Pickles et al., 2014). Currently, the field often uses experimental measures that assess discrete joint attention skills to monitor progress in an intervention. However, narrowing the developmental scope of experimental measures may obfuscate broader clinical interpretations of children's abilities. Even among these more sensitive experimental measures, subtle variability may not be sufficiently portrayed among minimally verbal children.

The field of autism intervention research is slowly addressing these needs for more accurate measures of social communication for minimally verbal children, but it is a time-intensive process (Fletcher-Watson & McConachie, 2015). This study proposes a more immediate alternative: analyzing an existing measure of joint attention with newer statistical techniques that may add additional meaning to scores and better characterize the skills of minimally verbal children. As researchers, if we can understand social communication skills before the intervention, we can better understand the progress made and determine if changes in the intervention protocol need to be made.

This study contains two parts. First, it examined individual, spontaneous social communication gestures to operationalize a hierarchy of gesture difficulty with an IRT framework. Next, machine learning algorithms calibrated salient and clinically-meaningful benchmarks as they relate to other standardized assessments to improve the meaning behind differing frequencies of behaviors. This study is among the first to use these novel quantitative methodologies to improve the characterization of social communication skills among minimally verbal children with autism.

Aim 1: Operationalize a hierarchy for spontaneous joint attention and behavioral regulation gestures among minimally verbal children with ASD with novel item response theory frameworks.

1a- Establish the best-fitting IRT structure for the social communication gestures by evaluating the Akaike Information Criterion (AIC; Akaike, 1974), Bayesian Information Criterion (BIC; Schwarz, 1978), and log-likelihood.

1b- Evaluate item parameters of the IRT model, including the easiness of spontaneous gesture initiations, which may determine the organization of social communication behaviors.

1c- Determine the correlation between ESCS raw response scores and IRT ability scores to corroborate the results of the final IRT model.

Aim 2: Create cut points in an experimental measure of social communication that relate to scores within standardized social communication assessments.

2a- Construct separate CART models to calibrate scores from the Early Social Communication Scales (ESCS; Mundy et al., 2003) to the Autism Diagnostic Observation Scales Social Affect domain and joint attention factor scores (Gotham et al.,

2007; Lord et al., 2012; Oosterling et al., 2010), Vineland Adaptive Behavior Scales, 2nd edition expressive and receptive language domains (VABS-II; Sparrow et al., 2005), and a Natural Language Sample (NLS; Kasari et al., 2014).

2b- Confirm CART findings with a machine learning ensemble method known as a random forest.

Methods

Participants

Participants from four different randomized, controlled trial (RCT) projects comprised the sample of the current analysis. All four projects included a naturalistic developmental behavioral intervention (NDBI) known as JASPER and focused on minimally verbal children with ASD (Kasari et al., 2006, 2008, 2012, 2014). The projects were referred to as Projects 1, 2, 3, and 4. All projects had inclusion criteria that included a confirmed ASD diagnosis by a licensed clinician via the Autism Diagnostic Observation Schedule (Lord et al., 2012). The exclusion criteria included major medical conditions other than autism, such as genetic disorders (Fragile X or Down syndrome and sensory disabilities (blindness or deafness). A University Institutional Review Board approved all original studies, and all parents provided written consent.

Project 1 included 61 individuals aged 5 to 8 years across three sites. All children had fewer than 20 spontaneous, different words used on a natural language sample and had a receptive language age of at least 24 months (Kasari, et al., 2014).

Project 2 included approximately 194 minimally verbal children with autism aged 5 to 8 years. Individuals were recruited across four sites. All children displayed less than 20 spontaneous, unique, and functional words during screening assessments and had a nonverbal

cognitive age equivalent of at least 18 months on the Mullen Scales of Early Learning (manuscript under review; MSEL, Mullen, 1997).

Project 3 included 42 minimally verbal children aged 5 to 11 years. Minimally verbal in this study was defined as having fewer than 30 functional words obtained from a natural language sample, parent report, and standardized tests.

Project 4 included approximately 160 children aged 33 to 54 months. Children had fewer than 30 spontaneous communicative words, as determined by a natural language sample, and a cognitive age equivalence of at least 12 months as determined by the visual reception or receptive language scales on the MSEL or Reynell Developmental Language Scales (manuscript in preparation; RDLS, Reynell, 1977). Project 4 recruited participants through public schools across three different research sites.

Study Designs

All projects implemented different study designs and timelines. Projects 1 and 2 were Sequential Multiple Assignment Randomized Trial (SMART) designs, meaning treatment protocols were altered depending on children's progress via mid-treatment randomization. For projects 1 and 2, there were 24 and 16 weeks of active behavioral intervention, respectively. Project 3 was a combination treatment for augmenting language in minimally verbal children with ASD. All children received the JASPER intervention and were randomized to either aripiprazole (flexibly dosed from 2-10 mg per day) or a placebo for 12 weeks.

Project 4 was a traditional randomized controlled trial comparing the JASPER intervention to another early intensive behavioral intervention on child outcomes across six months of intervention.

Due to the varied studied designs, only baseline data were used to address both aims. Future directions may incorporate the longitudinal data and examine how classification subgroups and joint attention gesture difficulty/discriminability change over time.

Measures

Autism Diagnostic Observation Scales (ADOS; Lord et al., 2012)

The ADOS is a well-validated, semi-structured assessment primarily used to help diagnose ASD (Lord et al., 2000). Within the assessment, the clinician first chooses the appropriate module based on the individual's language level. For this population, Module 1 was given to all participants, which is designed for nonverbal individuals or those who use single-word communication. An ADOS contains a standardized set of materials and prompts that are used to assess a child's social communication, repetitive, and social behaviors. These operationally defined behaviors are converted to scores and are used to determine if an autism classification is warranted based on specific cutoff scores. Subdomain raw scores for Social Affect (SA) and Restricted, Repetitive Behaviors (RRB) can be used to calculate an autism severity score.

Given the scope of this study and its relation to social communication, this study examined Social Affect (SA) scores from the ADOS in detail. Within SA domain scores, there is evidence that a joint attention factor exists, indicating a more distinct cluster of items that exclusively measure joint attention as a construct (Gotham et al., 2007; Oosterling et al., 2010). These items included gesturing, showing, initiating joint attention, and unusual eye contact. One item differed depending if the child had “few to no words” or “some words,” as determined by the first item on the assessment that measured the overall level of non-echoed spoken language.

If the child had no words, the joint attention factor included the response to the joint attention item. Conversely, the pointing item was included if the child was deemed to have some words.

Since this study included a range of minimally verbal children, both the pointing and response to joint attention items were included, along with the gesturing, showing, initiation of joint attention, and unusual eye contact items. This joint attention factor and the SA domain were scored using the same diagnostic algorithm, with higher scores indicating greater social communication difficulty (Gotham et al., 2007).

For analytic purposes, SA domain and joint attention factor scores were split at the median to create new dichotomous variables. The new binary variables were used as classification variables in the Aim 2 analyses.

Early Social Communication Scales (ESCS; Mundy et al., 2003)

The ESCS is a 15-20 minute, semi-structured measure of children's early social communication abilities. It is a common assessment used to evaluate young children with developmental delays, including those who are minimally or preverbal (Mundy et al., 2003; Anagnostou et al., 2015; Wetherby, 2006). During the administration of the ESCS, the child and a blinded staff member sit across from each other with a set of standardized toys that are in view but out of reach of the child. The toy set includes wind-ups, a car, a ball, glasses, and a book. A trained assessor (>80% fidelity) presents the toys consecutively, thus allowing the child opportunities to exhibit social communication skills.

The ESCS has demonstrated good reliability and validity in developmentally delayed children (Mundy et al., 1987, 1988, 1990, 1995, 2007). Additionally, high interrater reliability has been well documented across studies, indicating its coding protocol is relatively easy to master (Anagnostou et al., 2015). Social communication skills measured by the ESCS have been

associated with subsequent language outcomes in ASD populations, thus supporting it as a valid measurement protocol (Kasari et al., 2006; Mundy et al., 1990; Remington et al., 2007; Yoder and Lieberman, 2010).

Natural Language Sample (NLS; Kasari et al., 2014)

The NLS is a 20-minute standardized, naturalistic assessment in which an adult and child play with a set of specified toys. The NLS provides a standard time, toy set, and interaction style that can be used to measure a child's spontaneous verbal abilities (Tager-Flusberg et al., 2009). In the NLS, the adult remains responsive to children's verbal and nonverbal communication but does not prompt the child to talk. The NLS is a sensitive measure of language production over repeated measures among ASD samples in intervention contexts (Miller & Iglesias, 2010; Tager-Flusberg et al., 2009).

Blinded research staff administered the NLS. Before administering the NLS, the research staff was trained to at least 90% fidelity criterion on the NLS procedures. The NLS interaction was videotaped and later coded.

Vineland Adaptive Behavior Scales, 2nd edition (VABS-II; Sparrow et al., 2005)

The VABS-II is a semi-structured interview with a parent or caregiver that measures adaptive behavior in children and adults from birth to 90 years of age. A research professional conducts the interview, assessing four major domains of adaptive behavior: communication, daily living skills, socialization, and motor skills. Individual domain scores, as well as an overall sum score, are yielded. Raw scores can be converted to standard and age equivalent scores for each domain.

The VABS-II provides scores that demonstrate stability over repeated applications. It has strong inter-rater reliability (0.78-0.80), test-retest reliability ($r = .95$ to $r = .99$), and validity. In

the previous version of the VABS, scores are a sensitive unit of measurement (see Dawson et al., 2010; Williams et al., 2006). However, it has been recommended that the study design be at least six months to use the VABS-II data as an outcome measure.

Raw scores for the expressive and receptive language domains were split at the median and used as the classification variables in the machine learning analyses for Aim 2.

Coding and Reliability

Early Social Communication Scales

The ESCS coding procedure separates social communication behaviors into two categories: joint attention and behavior regulation skills. These subtypes of skills can be further categorized into initiations or responses to bids of behaviors, effectively creating four distinct domains of behaviors. For this study, ESCS data were not only analyzed at the domain level but also at the item level, i.e., the behaviors that constitute those domains.

Across all projects, blinded coders maintained at least 80% reliability as measured by intraclass correlation coefficients. Reliability was measured at the individual gesture level to ensure less frequent behaviors were accurately detected and coded.

Frequency of Individual Joint Attention and Behavior Regulation Gestures. The ESCS measures the number of joint attentional and behavioral regulation gestures. These gestures are often pooled together to create total IJA and IBR frequencies. For Aim 1, individual gesture frequencies were calculated within JA and BR domains. For example, the ESCS measures joint attentional points in the form of points alone, points with eye contact, points with language, and points with eye contact and language. All combinations of points were added together to achieve the total sum score for the pointing gesture. This step was repeated for the five joint attentional gestures/skills: eye contact (alternate gazes and coordinated joint looks),

language (by itself and paired with eye contact), points, shows, and gives. The five behavior regulation gestures/skills consisted of eye contact, language, points, gives, and reaches. Figure 1 outlines the ESCS code sheet to see more clearly all gestures measured. Individual gesture frequencies were examined in Aim 1 with an IRT framework and again in Aim 2 to determine their relationship to other assessments.

Initiations of Joint Attention (IJA; Mundy et al., 2003). IJA frequencies are the number of times children use nonverbal behaviors (e.g., pointing, showing, and looking) and verbal communication spontaneously to coordinate another person's attention to share an event or object.

Initiations of Behavioral Regulation (IBR; Mundy et al., 2003). IBR behaviors are a set of child gestures and verbal communication aimed at recruiting the help of another person to fulfill a request or need, e.g., using points to obtain a specific object. Like IJA, all IBR frequencies are combined to create a sum score.

Response to Joint Attention (RJA; Mundy et al., 2003). RJA is the number of children's successful responses to an adult's bid for joint attention. RJA is coded as a percentage.

Response to Behavior Regulation (RJA; Mundy et al., 2003). RBR is the number of times a child successfully responded to an adult's bid for behavioral requests. RBR is coded as a percentage.

Natural Language Sample

NLS videos were transcribed by blind research staff using the Systematic Analysis of Language Transcripts conventions (SALT; Miller & Iglesias, 2010). Once transcribed, the nature of children's language was coded by its type and function. Types of communication include: spontaneous, imitated, elicited, or prompted. Functions of language include requesting,

commenting, or other. Coders and transcribers were trained to reliability above 80% as measured by intraclass correlation coefficients.

Spontaneous Communicative Utterances (SCU; Kasari et al., 2014). SCU refers to the number of requests, comments, or protests spontaneously initiated by the child. Blinded raters verified transcripts and codes to ensure correctness. Spontaneous utterances of all types were included in the calculation of SCU.

SCU, a continuous variable, was dichotomized at the median to create a binary outcome variable. The binary variable effectively classified individuals with low and high SCU, which was used as an outcome variable in the CART analysis for Aim 2.

Number of Different Word Roots (NDWR; Kasari et al., 2014). NDWR measures the number of unique, spontaneous words a child uses in the NLS. This variable is similar to SCU but is not inflated by additional utterances of spontaneous words. Similar to other classification variables used in Aim 2, NDWR was dichotomized at its median to divide children into those with fewer and more word roots.

Analytic Plan

Aim 1: Create a Hierarchy of Social Communication Gesture Difficulty

1a. Establish the best-fitting IRT model. As mentioned earlier, the ESCS presents a series of play-based opportunities for a child to demonstrate social communication skills. Data are considered count data as the number of behaviors are non-negative integers that conceivably have no limit to the frequency of each skill. However, if researchers want to investigate item-level data, often typical among psychological tests, this type of data format would not accommodate typical IRT models, which require dichotomous or polytomous data (Embretson & Reise, 2000; Hambleton et al., 1991). Therefore, the data structure of the ESCS not only lends

itself to being a prime candidate for RPCM, but a good model fit also corroborates initial CTT scoring methods due to the sufficiency of the total score statistic within the RPCM.

An RPCM model and its extensions were fit to the ten initiation skills (items) as measured by the ESCS. Per the ESCS scoring protocol, these items can be aggregated into higher levels of analysis (i.e., domain scores of the ESCS). Despite separate functions of these skills (joint attention vs. behavior regulation), their intent of initiation was deemed acceptable justification to include all gestures in a single unidimensional model. Under this definition and by evaluating individual initiating skills of the ESCS in this manner, it was assumed that these specific gestures represent the latent trait of social communication initiations well, thus lending itself able to detect incremental increases in skills among minimally verbal children.

A prerequisite for RPCM analyses is that data must be in its long format. The data was restructured so that items were a variable column that consisted of the individual behavior skills and a frequency variable was the number of times that skill was demonstrated for each child. These raw frequencies or counts of individual behaviors were the unit of analysis and modeled in the RPCM.

First, a "person-only" model was created, which assumed equal difficulty across items but included a person parameter (random intercept model). Next, a simple RPCM that allowed item difficulty to vary was constructed as a comparison model. In addition to item-wise intercepts, otherwise known as the easiness parameters, the RPCM included a random intercept on the person-level, with mean 0 and unknown variance. Next, a likelihood ratio test (LRT) with a Chi-squared test statistic was computed to compare the fit of the two models.

The fit of the RPCM model was assessed by observing how well the model correctly predicted total scores, as total scores are considered a sufficient statistic for estimating model

parameters (Baghaei & Doebler, 2019). Model fit was also evaluated with numerous fit indices such as AIC, BIC, and -2log-likelihood. Visual checks of model fit were employed by graphing predicted values for each child plotted against their standardized Pearson residuals. Roughly symmetrical Pearson residuals with minimal outliers indicate acceptable fit (Baghaei & Doebler, 2018). Additionally, a covariate-adjusted frequency plot was used for a graphical model check (Holling et al., 2013).

Next, a dispersion index was calculated for the simple RPCM model. The RPCM is classified as a log-linear model considered a type of generalized linear mixed model; therefore, the ability to check the presence of under or overdispersion was also implemented (Demidenko, 2013; Doebler & Holling, 2016; McCullagh & Nelder, 1989).

The Poisson distribution within the RPCM requires equidispersion, that the variance in the outcome is identical to what the model assumes. Under this assumption, the dispersion index equals one. In overdispersion, dispersion indices are greater than one and indicate that the variance is greater than expected (Bliss & Fisher, 1953). This phenomenon often occurs in empirical data.

Given the population of interest, the skills measured by the ESCS may be especially difficult for many minimally verbal children, resulting in too many zero values compared to the number expected in a classical count probability distribution, otherwise known as zero-inflation (Heilbron, 1994). Overdispersion and zero-inflation are related as an excess number of zeros can contribute to overdispersion, which would require model adjustment to ensure that the model fits the data well. Ignoring overdispersion and zero inflation in model construction would typically lead to the overestimation of standard errors and may return biased parameter estimates (Lambert, 1992; MacKenzie et al., 2002)

Therefore, this study constructed successive models and evaluated the model fit with each iteration. These models included a negative binomial regression model (NBRM) and a zero-inflated negative binomial model (ZINBM), both of which attempted to account for the overdispersion in the data. NBRMs are beneficial for data that otherwise follow a Poisson distribution but contain extra skewness, and its zero-inflated counterpart was able to account for variability in the data by directly accounting for the excess zeros.

The RPCM was estimated using the lme4 package (Bates et al., 2017). Overdispersion was assessed with the DHARMa (Hartig, 2022) and performance (Lüdecke et al., 2021) packages. Negative binomial models were handled with lme4, but the zero-inflated models were estimated using the GLMMadaptive package (Rizopoulos, 2022). All the analyses mentioned above were conducted in R (R Core Team, 2022)

1b. Evaluate item parameters of RPCM. Once the best fitting model was identified, item parameters were evaluated. The fit of individual gestures was evaluated by Pearson residual boxplots for item-specific predicted scores. Item easiness parameters and their standard errors operationalized gesture difficulty, and item fit was checked by χ^2-tests as suggested by Baghaei and Doebler (2018).

1c. Relate latent gesture ability scores to ESCS domain scores. Correlations between theta scores derived from the final IRT model, which measured latent social communication ability, and raw domain scores of the ESCS (IJA, IBR, RJA, and RBR) were conducted. Correlations between theta scores and the IJA and IBR domains were of substantive interest to corroborate model construction and fit of the unidimensional RPCM from Aim 1a. Although there is evidence that initiation of gestures and response to bids of social communication are related, there is also evidence that they may represent different underlying constructs and have

distinct developmental trajectories (Mundy et al., 2007, Mundy & Volkmar, 2013). This study sought to investigate the relationship between these constructs within this sample.

Examining correlations between theta scores and response abilities was necessary since separate IRT models could not be created for response skills. Response data were presented as proportions and unsuitable for an item response theory framework. Additionally, this study could not separate the few individual items from response domain scores. Therefore, it was considered a decent compromise to perform a correlation analysis between theta scores and RJA/RBR to uphold the psychometric cohesion of the ESCS.

Aim 2: Create Cut-Points in ESCS Scores

2a: Calibrate ESCS to three standardized social communication assessments. Aim 2a was addressed using the CART machine learning technique. CART implements recursive partitioning, a statistical technique for uncovering relationships between variables as they pertain to a binary classification outcome. Compared to other ubiquitous linear and generalized linear models, CART can discover interactions and unknown relationships amongst the variables, which may go unnoticed with other analytic techniques.

CART is well suited for this Aim 2 as we do not know how differential ESCS scores, either from individual gesture or domain scoring, are related to other standardized assessments that measure similar constructs. The main benefit of decision trees, which are a graphical illustration of the results from the machine learning algorithm, is that they are easily interpretable and follow logical decision-making processes. However, it is essential to note that CART models are for exploratory discovery rather than hypothesis testing and inference.

The CART models generated empirically derived cut points in continuous ESCS scores as they relate to six outcomes across three different standardized social communication

assessments. The outcome measures were ADOS SA and JA factor scores, SCU and NDWR derived from NLS, and expressive and receptive scores from the Vineland. These outcomes were dichotomized at their medians and served as the binary classifiers in the classification models. Classification trees were chosen over regression trees primarily to elucidate the clinical interpretation of ESCS scores. Regression trees address different questions that are outside the scope of this study.

There were two sets of independent variables used for each outcome measure. One set included individual ESCS skill frequencies from Aim 1, and the other set included the domain scores from the ESCS (e.g., IJA, IBR, RJA, and RBR). It was imperative that both sets of explanatory variables were included in order to represent the complete scoring system of the assessment. The use of independent gestures aligned with Aim 1, and was an essential step in understanding how gesture skills were situated among scores of standardized tests. Modeling domain scores reflect how the ESCS is currently scored, and examining these variables with CART presented opportunities for its widespread application.

Per standard machine learning guidelines, within every CART analysis, the dataset was first split randomly into training (75%) and testing (25%) datasets that preserved the ratio of the dichotomous outcome variable. In the training set, the classification criteria were known and were used for supervised classification to create a plausible model. The testing set represented real-world conditions where the classification was unknown and was used to test the performance of the models generated in the training sample. Although predictive ability was evaluated internally with each model, this study took a conservative approach and performed validation on a test set instead of using repeated cross-validation on the entire data set.

Combining the two sets of independent variables and the six possible outcome measures resulted in 12 separate CART analyses. Each classification tree was fit and analyzed separately. Trees were created by starting with a root node that consists of all individual data, and at each step, the algorithm determined the optimal split that maximizes the decline in the Gini coefficient. Splits created cutoff values within each explanatory predictor and continued until homogeneous subgroups were created or when certain criteria were met, specifically that a split would not be attempted if it did not improve the model's fit by a pre-defined value (Atkinson & Therneau, 2000).

A 10-fold repeated cross-validation with five repeats was used to assess the stability of the tree in the training set. This method effectively created 50 different held-out sets to assess efficacy. Before each repetition, the sample was shuffled, resulting in different sample splits. Typically, 5 or 10 are chosen as the number of folds within cross-validation as these values have been empirically shown to yield test error rate estimates that are not affected by high bias or variance (James et al., 2013). However, there is no formal rule for the number of folds. Since the difference in size between the training and resample subsets gets smaller as K gets larger, this study opted for K=10 to mitigate bias (Kuhn & Johnson, 2013). Although computationally more expensive, repeated K-fold cross-validation was chosen over standard cross-validation as it provided a more stable performance estimate. Additionally, ten separate complexity parameters were evaluated within each tree, which ultimately measured the cost of adding another variable to the model.

Pruning of the branches was based on Breiman's "one standard error" rule, which suggested that a simpler model within one standard error of the empirically optimal model would safeguard against overfitting and be the better model choice (Breiman et al., 1984). The

algorithm chose the largest complexity parameter with one cross-validation error within one standard error of the minimum cross-validation error, which effectively calculated the additional accuracy a split must add to warrant additional complexity.

All classification trees' predictive accuracy was assessed using internal (repeated cross-validation on the training set) and external validation (using the siphoned test set). Predictive probabilities are presented in each terminal node of the decision trees, and performance on the test set was evaluated with a confusion matrix.

The trees were constructed using the R package caret (Kuhn, 2008) and then visualized using rpart.plot package (Milborrow, 2018)

2b: Confirmation of CART Analyses with Random Forests. The CART model chose the most important variable that predicted the outcome measure based on the Gini index at each node. In order to corroborate variable importance, an ensemble method known as a random forest was used. A random forest fitted a more accurate model by averaging many trees together, thus reducing the variance and potential overfitting within a single tree.

Typically, if there are many explanatory variables, a random forest can extract the most important variables, which can then be used in a CART analysis. This study only used a limited number of independent variables (four or eight depending on the model); therefore, a random forest model was instead used as a corroboration technique to confirm results in the CART analyses.

Although computationally expensive, each random forest model averaged 1,000 unpruned trees together to maximize accuracy. Due to the Strong Law of Large Numbers, random forests always converge; therefore, an excessive number of trees would not overfit nor be an issue (Breiman, 2001). A separate algorithm from the caret package was used to tune the

crucial parameter of the random forest that determined the optimal number of variables randomly sampled at each split within each tree, otherwise known as mtry. Each forest used a mtry that maximized its accuracy.

The random forest models performed an internal validation check as an alternative to the split sampling method of repeated cross-validation used in the classification tree analyses. Two-thirds of the data was used as training data, and the remaining third was used to calculate the out-of-bag error, which was an indicator of accuracy. Using the out-of-bag error estimate eliminated the need for a testing set and gave unbiased estimates as compared to cross-validation, which exhibits data leakage as some data is used in the training model in some way — a factor that may affect variance and overfitting (Breiman, 2001)

The original random forest algorithm from Leo Breiman was used in this study (Breiman, 2001) and was implemented in R with the randomForest package (RColorBrewer & Shaw, 2018).

Results

Descriptive Statistics

453 participants were included in the present analyses. Demographic information on child and parent characteristics was consolidated from each project and presented in Table 1. The majority of children in this study are male (82.8%), which aligns with current differential diagnostic estimates by gender, and is similar to other autism studies. The average age of children was 64.1 months (*SD* = 19.9; 5.34 years). The race composition of the sample is relatively heterogeneous, with nearly 40% of children identifying as something other than white. Most parents had completed college: 78.9 and 77% for mothers and fathers, respectively.

Descriptive statistics were generated for each measure, including individual and domain-level ESCS scores and each of the six outcome measures from the three standardized

assessments, and are presented in Tables 2-4 and depicted in Figures 2 and 3. Inspection of the individual skill statistics from the ESCS signaled that data were right-skewed, indicating that, on average, children were not exhibiting these pre-linguistic skills often. The frequency distribution across skills further hinted that extensions to the RPCM would likely be needed to account for the skewness.

Responses to bids of joint attention and behavior regulation at the domain level included a wider range of scores due to their proportion data structure. Initiations of joint attention maintained right skewness, and initiations of behavior regulation exhibited more variability, which is consistent with developmental trajectories of this population; BR skills generally develop in accord with typical development while JA skills do not (Paparella et al., 2011)

Rasch Poisson Counts Model

The RPCM with items as predictors fitted the data significantly better than the person-only model, which assumed equal difficulty for all the items but included a person parameter [χ^2 (9) = 8150, $p < 0.001$]. The standard deviation of the ability parameter for the RPCM was estimated to be 0.67. Fit statistics of the RPCM are outlined in Table 5. The fit of the RPCM confirmed that the latent social communication variable is quantitative and that items, as well as the latent variable, were successfully measured on an interval scale with a common unit of measurement, a crucial aspect of proper model fit (Wright, 1977).

Figure 4 shows the Pearson residual plot for the RPCM, which graphed the predicted values for each person on the x-axis and the Pearson residuals on the y-axis. The residuals are the difference between the observed scores and the scores predicted by the Poisson model. Good model fit is indicated by Pearson residuals with a mean of 0 and a standard deviation of 1.0. For the RPCM, the residual variance increased for lower predicted scores, displaying potential

overdispersion in the data. The counts were seemingly too large for small values of the predicted values, indicating more variance than the model predicted. Potential overdispersion is common and is not necessarily detrimental to model construction; however, it may lead to more liberal inferences since the Poisson distribution did not predict enough variance at these low counts (McCullagh & Nelder, 1983).

Figure 5 shows Pearson residual boxplots for item-specific predicted scores. These results generally give the same information as Figure 4, and confirm that data may be overdispersed.

The blue dashed line in Figure 6 shows the expected frequency of item scores as predicted by the RPCM (Holling et al., 2013). The RPCM line approaches the observed scores more closely than the person-only model indicating a better model fit.

Table 6 displays the item easiness parameters of the RPCM on the counts level. On average, JA skills are more difficult than BR skills, with joint attention shows and gives as the most difficult skills to demonstrate.

Checking for Overdispersion and Zero-Inflation

Before further interpreting these item estimates, the dispersion parameter of the RPCM was checked. Overdispersion is frequently observed in applications of Poisson regression, and given the population and latent ability, it was hypothesized that this would likely occur. However, it is important to note that overdispersion could not be confirmed until after the RPCM model was fitted. The performance package indicated the dispersion index or the ratio of the variance to the mean of the fitted model was $\varphi = 2.97$ [$\chi^2 = 13424.84$, $p < 0.001$]. This value indicated a variance value nearly three times the mean. Thus, data were overdispersed relative to the Poisson distribution, confirming suspicions from the Pearson residual plot.

Simulation-approach. For a more thorough examination of overdispersion and zero inflation, a simulation-based approach was used with the DHARMa package, which transformed residuals to a standardized scale between 0 and 1 (Hartig, 2017). To calculate the scaled residuals, this study ran 250 simulations, which was deemed a reasonable compromise between computation time and precision. Simulations created synthetic data from the fitted RPCM model, and the cumulative distribution of simulated values was calculated for each observed value, thus rendering a standardized value corresponding to the observed value. The algorithm from the DHARMa packaged returned an object containing the simulations and scaled residuals, which was used for all plots and subsequent tests.

Figure 7 is a visual interpretation of the plot of the scaled residuals. The left panel of Figure 8 depicts a QQ-plot, which was used to detect deviations from the expected distribution, and a plot of the residuals against the predicted value. The QQ-plot indicated that more residuals were in the tail of the distribution than would be expected under the fitted model. By default, DHARMa tested for distributional accuracy (KS test), dispersion, and outliers. The Kolmogorov-Smirnov test for overall uniformity of the residuals, overdispersion tests, and outlier tests were all significant, indicating data were overdispersed and potentially zero-inflated. The significant outlier test was interpreted with caution as outliers are dependent on the number of simulations and thus not defined in terms of particular quantiles. The right panel of Figure 7 returned a plot of the residuals against the predicted value. Values were rank transformed to aid in visually identifying patterns in data. Simulation outliers are highlighted as red stars and are defined as data points outside the range of simulated values. Like the QQ-plot, outliers are interpreted with caution as we do not know how much these outliers deviate from the model expectation.

Formal goodness-of-fit tests were conducted on the simulated residuals to support the visual inspection of the residuals. First, a dispersion test was conducted, and details and results are presented in Figure 8. In Figure 8, the red line represents the variance of observed raw residuals, which was displayed against a histogram depicting the variance of the simulated residuals. The dispersion ratio was 1.45 and was significant with $p = 0.024$, thus confirming overdispersion.

Next, a separate zero-inflation test was implemented, which compared the distribution of expected zeros in the data against the observed zeros. The test was significant, $p < .001$, with a ratio of observed zeros against the expected zeros to be 1.46, indicating the model was underfitting zeros and that there was zero inflation. Figure 9 is a visual representation of this test.

Post Hoc Analyses for Overdispersion and Zero-Inflation

Unfortunately, the dispersion of the Poisson model often underestimates the observed dispersion, otherwise known as overdispersion. This phenomenon is common among Poisson data and is either caused by population heterogeneity, an excess number of zeroes, or both within the data (Hilbe, 2011). When there is sufficient dispersion in the data, the single Poisson parameter is often insufficient to describe the sample population and therefore leads to the violation of the core equidispersion assumption of the Poisson distribution (Bohning et al., 1999). The RPCM violated this assumption, ultimately obfuscating statistical inferences from the item parameters (Lindsey, 1993; Thall & Vail, 1990). However, despite the overdispersion, the RPCM and its item parameters may still be a useful approximation of the true model. Nevertheless, to have greater flexibility in the relationship between the mean and variance, extensions of the RPCM were implemented.

The first model considered was the negative binomial regression model, which directly addressed overdispersion. Next, another model extension was used, the zero-inflated negative binomial counts model, which considered the excess zeros in the data. It was imperative to fit and compare both models since it is difficult to differentiate between overdispersion and zero-inflation reliably a priori.

Negative Binomial Regression Model

The first post hoc model embedded the RPCM within an overdispersion framework via a negative binomial regression model (NBRM; Hung, 2012), which effectively relaxed the assumption of equidispersion. In this model, the parameters were estimated more accurately due to an additional random effect that allowed the variance to differ from the mean, thus handling the extra variation in the data (Engel, 1984; Hung, 2012; Lawless, 1987). Despite the significant zero-inflation in the RPCM, it was decided that adding a zero-inflation term was not warranted for this next model iteration. It was hypothesized that accounting for the overdispersion within a negative binomial distribution may effectively model the excess zeros.

Model Fit. Per an ANOVA test, the NBRM with items as fixed effects and random intercepts at the person level fitted the data significantly better than the RPCM [$\chi^2(1) = 4351.4$, $p < 0.001$]. This improvement in model fit was expected when overdispersion was found in a Poisson model. The fit statistics are outlined in Table 5 and demonstrate that AIC and BIC are significantly lower than those of the RPCM. Item parameters and their standard errors are outlined in Table 6.

Model Inspection. Similar to the RPCM, dispersion and zero-inflation were checked for the NBRM. An initial test with the performance package indicated that overdispersion was not detected, indicating the residual distribution fitted much better to a negative binomial distribution

[$\varphi = 0.953$, $\chi^2 = 4304.3$, $p = 0.10$]. Moreover, a zero-inflation test indicated that the ratio of observed and predicted zeros (0.96) was within the tolerance range but had opportunities for improvement.

Simulation Tests. For thoroughness, additional tests were conducted with simulations via the DHARMa package. The initial QQ plot in Figure 10 indicates that dispersion may still be an issue. Testing for dispersion within the DHARMa package indicated significant underdispersion in Figures 10 and 11. It seems the residual distribution did not fit very well with a negative binomial distribution, potentially due to outliers observed in the right panel of Figure 10. In order to account for the outliers and likely excess zeroes, DHARMa may have adjusted its parameters, which may not have been needed in other parts of the data, resulting in underdispersion.

Figure 12 shows the results of the significant zero-inflation test, which further justified that another iteration of the model would need to account for these excess zeros.

Zero-Inflated Negative Binomial Regression Model

Justification for Zero-Inflated Model. Given the significant zero inflation test within the NBRM, the next logical progression of model construction was the zero-inflated negative binomial model (ZINBM), whose distribution has additional flexibility to assign the probability of zero counts beyond that of the NBRM (Greene, 1994). Additionally, the suspected presence of structural zeros in the data, which are theoretically expected among this population, is another key justification for using a zero-inflated model. Children with structural zeros do not exhibit these gestures at this early baseline stage of the intervention process because they may be too difficult for them. This process is different from those who are classified as random zeroes, otherwise known as children who could potentially exhibit gestures but do not. Random zeroes may occur due to sampling variability or external factors like poor rapport with the tester within

the behavioral assessment. In a ZINBM, both processes are accounted for, which effectively addresses overdispersion that is not only caused by zero inflation.

The ZINBM, also called a mixture distribution model, generated two separate models based on different distributions and allows for the possibility of subpopulations exhibiting different latent trait distributions (Broek, 1995; Farewell & Sprott, 1988; Lambert, 1992; Mullahy, 1997). Children who do not exhibit any skills (degenerate distribution at zero) were modeled with a logistic regression which returned zero-part coefficients that determined how unlikely it would be to exhibit a zero count for that skill. Those children who displayed some skills, including random zeros, were modeled with a negative binomial model, which returned easiness parameters as seen in the RPCM. In the model specification, a random effect term of the zero-inflated part was also used to estimate parameters.

Although statistical modeling is partly concerned with representing the true structure of phenomena, models must also help extract meaningful information from the data regarding the population of interest (Konishi & Kitigawa, 2008). These IRT model iterations not only helped determine the best fitting model, but they also aided in accomplishing the research aims in approximating the explainable information regarding minimally verbal children and their abilities within this set of empirical data. Therefore, the use of the zero-inflated model as the final model, which modeled zeros differently than count values (Washington et al., 2003), was justified.

Model Fit. A Vuong test was used to compare model fit between the non-nested NBRM and a ZINBM (Vuong, 1989), and the results confirmed that the ZINBM fitted the data significantly better than the NBRM [$p < 0.001$]. Lowered AIC and BIC of the ZINBM in Table 5 indicated improvement above all other models.

Despite its ability to model count data, the Poisson distribution in the RPCM did not address the issue of overdispersion. Moreover, the negative binomial regression was less capable than the zero-inflated negative binomial model of addressing the problem of excess zeros in the data. Therefore, the zero-inflated negative binomial model fitted the observed number of joint attention and behavior regulation skills better than any of the other models tested.

Item-specific residual boxplots for each item are depicted in Figure 13. Figure 13 shows Pearson residual boxplots for item-specific predicted scores and aid in the interpretation model fit. Mean residuals were roughly zero for every item, and as expected, most settled between -2 and 2.

Model Inspection with Simulation Tests. Inspection of simulated residuals via the DHARMa package indicated that the ZINBM successfully accounted for the overdispersion and zero-inflatedness. Although the KS test indicated a significant deviation from the assumed distribution, the model's fit was still deemed acceptable. The QQ-plot in Figure 14 is nearly perfectly linear, which suggests that the overall distribution is satisfactory. The plot of the residuals against the predicted value in Figure 14 looks approximately normal with potentially higher density towards the middle with very few outliers, who are likely responsible for the significant KS test. Separate tests of dispersion and zero-inflation are depicted in Figures 15 and 16, respectively; both were not significant and indicated that the model fitted well.

Interpretation of Item Parameters of ZINBM

The fixed effect and zero-part coefficient estimates and their standard errors are outlined in Table 6. Separate interpretations occurred for each part of the ZINBM.

Fixed Effects

The fixed effects of the ZINBM are interpreted similarly to the RPCM item estimates. A joint attention show was the most challenging gesture, indicating that a child required a more refined social communication ability to exhibit these gestures. Conversely, a behavior regulation reach was the easiest skill and required relatively less latent social communication ability to demonstrate this gesture. On average, joint attention gestures were generally the most difficult gestures to exhibit, followed by all the behavior regulation gestures. Behavior regulation eye contact was the exception and was considered the third most difficult gesture. BR eye contact may have been modeled as more difficult than it truly is as these gestures are often not observed alone and are typically paired with another gesture. The complete ranking of item easiness parameters is outlined in Table 7.

A version of the Rasch person-item map is displayed in Figure 17. This item-order map displays the location of the item parameters along the latent social communication dimension. Items are ordered by average location parameter, and minimum and maximum values of the lines represent the lowest and highest value of the location parameter associated with each item. These maps are generally helpful in comparing the range and position of the item measure distribution. Items span the full scale, which indicates that the ESCS successfully and meaningfully measures skills and gestures that constitute the range of social communication ability. From a psychometric standpoint, this is an important characteristic that bolsters its internal validity.

Theta distributions were also included and are displayed in Figure 18. All theta scores of each child, otherwise known as their latent abilities or random effects, were plotted by frequency. Higher ability scores indicated greater social communication ability. The depicts most children falling within the average ability range.

Zero-Part Coefficients

The ZINBM explicitly modeled the zero counts present in the data. Zero-part coefficients gave the log odds of *not* exhibiting the gesture, which was another way of identifying difficulty across gestures. Therefore, if coefficients were positive for the zero-inflated part, then the log odds of not having any of those skills were relatively higher than other skills. Parameters indicated that the log odds of not exhibiting any behavior regulation reaches (count of zero) were the lowest compared to all other skills, thus corroborating results from the count portion of the model. Generally, BR skills were less likely to have a zero-count compared to JA skills. However, JA eye contact, also recognized as the reference item, was considered the third easiest skill, i.e., among the top three skills most likely not to exhibit a zero-count. For these children who generally are not exhibiting gestures, facilitating joint eye contact may be a good entry skill to facilitate more complex joint attention skills. All zero-part coefficients are displayed in the bottom panel of Table 6.

Association Between Theta Values and Domain Scores

RJA and RBR scores were calculated as percentages representing the percentage of successful responses to bids of joint attention or social communication. These domain scores also represent a different social communication construct compared to IJA and IBR frequencies. For these reasons, prior IRT analyses could not model RJA and RBR scores within a unidimensional model. To reconcile these two constructs of social communication, theta scores from the ZINBM, which represent initiation abilities, were compared to RJA and RBR scores as a corroboration technique.

Theta abilities of initiation abilities were significantly correlated to both RJA and RBR frequencies, with a higher Pearson's correlation coefficient with RJA. These results indicate that although there are separate constructs, ability levels derived from the ZINBM are related to a

child's ability to respond to an adult's bids for attention. Table 8 shows the correlation matrix for theta, RBR, and RJA scores.

Classification Trees

Descriptives and Data Preparation

Two sets of classification trees were created. One set included the frequency of individual ESCS items as the predictors, and another set included the ESCS domain scores as the predictors. This study decided that including both sets of trees was imperative to further uphold the overarching study aim to characterize minimally verbal children. Specifically, using individual ESCS scores aided in connecting findings from Aim 1 regarding item-level analyses. Additionally, using domain scores maintained the integrity of the ESCS scoring protocol, as extant research typically reports results in this manner.

All trees included the same classifiers: ADOS SA and JA factor scores, NLS SCU and NDWR, and Vineland Expressive and Receptive scores. Table 4 shows the descriptives for the classifiers. SCU and NDWR from the Natural Language sample were the most right skewed, as demonstrated in Figure 3. These outcomes singularly measured verbal language ability, and it is therefore unsurprising that the frequency of spontaneous communicative utterances (M = 11.1, SD = 15.8) and the number of word roots (M = 7.28, SD = 9.97) were especially low for this minimally verbal population. The distributions of scores from the other outcomes are also presented in Figure 3, demonstrating more normal distributions. Evaluation of data confirmed that classification trees, as opposed to regression trees, would be most valuable in interpreting these variables.

Before CART analyses could be implemented, the classifier variables were dichotomized. Although creating a binary variable from continuous data may inherently decrease the

variability, it was decided that at this initial and exploratory stage in the machine learning process, classifications would give more meaning to cut points in ESCS scores as opposed to a regression output. All outcome variables were split at their median, which provided theoretically meaningful and easily interpretable splits in the variables.

Correlations between individual ESCS gestures were examined prior to the introduction of the machine learning algorithm. Results are depicted in the Figure 19 correlogram, where the color gradient from purple to red represents the size of the Pearson correlation coefficient. Red represents the strongest positive association. Joint attention and behavior regulation language skills were strongly correlated and were excluded as independent variables to optimize the algorithm. Theoretically, relating cut points in behavior regulation and joint attention language to other standardized assessments measuring language did not seem to contribute unique and meaningful information.

Classification Tree Results

Figure 20 shows all classification trees with independent ESCS skills as predictors, and Figure 21 contains all trees with ESCS domains. All models were used to identify ESCS count thresholds corresponding to binary classifications across the following dependent variables: NDWR, SCU, expressive and receptive language from the Vineland, and SA and JA factor scores of the ADOS. Each node in the trees shows the predicted class (higher or lower scores on outcome variable), the predicted probability of belonging to the group with higher scores (green terminal node), and the percentage of observations in the node.

ESCS Individual Gesture Trees. Across most trees that used individual gestures, behavior regulation points were identified as the only and, therefore, most important splitting variable with an optimal cut-off of 1 initiation. For all outcomes apart from SA scores from the

ADOS, one behavior regulation point split the training data into subgroups with the greatest class purity. One BR point was able to distinguish children with more spontaneous communicative utterances, greater number of word roots, more robust expressive and receptive language, and greater joint attention skills from their counterparts.

The tree for SA scores was the most complex, with BR points being the first predictor that maximized the decline in impurity via the Gini index followed by JA eye contact and BR EC. There was a 79% probability that a child had less severe SA scores if they had three or more BR points. If a child had less than three BR points but four or more JA eye contact initiations, there was 61% probability of belonging to the less severe SA group. Although not identified as the most important predictor in the tree, numerous joint attention eye contact initiations were a crucial skill that predicted social affect scores of the ADOS. Alternatively, less than three behavior regulation points, four JA eye contact initiations, and four initiations of behavior regulation eye contact classified the child as having more severe SA scores.

Prediction on Test Data. The prediction metrics of each model were determined on the test data set and are presented in Table 9. All information was derived from the confusion matrix, a performance measurement for machine learning classification. In a confusion matrix, the actual target values were compared with those predicted by each machine learning model. The accuracy statistic indicates the percentage of correctly predicted observations. More detailed metrics like precision and sensitivity were calculated within overall accuracy.

Precision, also known as the positive predictive value, is a metric that calculated the number of correct predictions made out of the total number of positive predictions. Sensitivity is the true positive rate, which was calculated as the number of correct positive predictions made out of all positive predictions that could have been made. The main difference between

these two metrics is that sensitivity considers missed positive predictions (Fernández et al., 2018). The F-Measure (F1) combines both sensitivity and recall into one metric and can be used similarly to accuracy to summarize total model performance. All metrics range from 0.0 to 1.0, where higher scores indicate better predictive ability.

Overall accuracy was not high across the CART models that used individual gestures as independent variables, as shown in Table 9. However, sensitivity was especially high for spontaneous communicative utterances and word roots, and precision was especially high for ADOS SA and JA factor scores. These metrics and outcomes may be more meaningful given the population and research aim to characterize minimally verbal children. Sensitivity may be more appropriate if minimizing false negatives is the goal, which may often be the case in intervention research. In intervention research, it may be more meaningful to protect against classifying a child with more robust language ability when they do not (false negative) as oppose to the overall accuracy of the classification tree.

Similarly, in the ADOS trees, the scoring is reversed; the less severe score group (more robust joint attention and social affect) is the positive class. Therefore, the precision estimate may be more meaningful as it minimizes false positives and guards against predicting that a child has more robust social communication skills when they do not.

ESCS Domain Trees. Across all trees apart from the ADOS SA tree, responding to joint attention was the most important predictor. RJA was the single most important variable that predicted class membership and effectively decreased the Gini index within the constraints of the tree's splitting rules for trees that included Vineland's receptive and expressive language and LS number of word roots. Responding to joint attention around 35% of the time (average across trees with RJA as the lone predictor) was the crucial cut point to differentiate children regarding

their specific language skills as reported by parents (receptive and expressive) and on an objective language sample (NDWR).

The SCU tree is more complicated than the NDWR tree, likely due to its greater variability in scores. Again, responding to joint attention 37% of the time was the first division in the tree, and not complying with this criterion led to fewer SCU and constituted 39% of the sample. Among the rest of the children who responded to joint attention 37% of the time or more, those who initiated behavior regulation gestures 15 times or more were classified as having more spontaneous communicative utterances with a 69% predicted probability– the highest probability across all four terminal nodes in the tree. If there were less than 15 initiations of behavior regulation, children could still be classified as having more SCUs if they had five or more initiations of joint attention.

The joint attention factor tree was the most complex, which may be because the ESCS and the JA factor measure the same construct of joint attention. Responding to bids of joint attention 30 percent of the time was the first cut point, and failing to do so classified children with more severe joint attention factor scores with a predicted probability of 73%. A combination of an RJA of 30 or more with 11 or more initiations of behavior regulation resulted in a less severe JA factor score with a predicted probability of 75%, indicating that 75 percent were correctly classified and 25 percent were misclassified. For those who responded to joint attention more than 30% of the time but had less than 11 initiations of behavior regulation, an RBR score of 57 or greater was able to salvage a less severe JA factor classification; however, this only constituted 3% of the sample.

Prediction on Test Data. All prediction metrics are presented in the bottom panel of Table 9. The accuracy of the testing data set that examined the other 25% of the sample was

satisfactory across trees. Prediction metrics were the highest for the ADOS trees, with overall prediction accuracies of .70 and .72 for the social affect and joint attention.

Random Forests

The random forest ensemble method was used for every classification model. In these random forests, 1,000 unpruned classification trees were generated from the entire dataset by bootstrap aggregating (bagging; Breiman, 1996). By bagging, the random forest algorithm arbitrarily learned the dataset and combined learning results across numerous trees, thus effectively solving potential overfitting within a single tree.

During the classification tree construction phase, a separate tuning algorithm determined the optimal number of variables as a split criterion to perform prediction, ultimately minimizing out-of-bag (OOB) error. OOB error is a metric that internally validates the random forest model. The OOB score was computed as the average error calculated from the samples not used in their bootstrap sample (Hastie et al., 2009). The OOB score was preferred to guard against data leakage in training the model, which is observed in cross-validation and may affect the predictive ability of the model.

Across all trees, the optimal number of variables to sample at each split was 2 (mtry = 2), which is similar to the default recommendation of $\sqrt{p}$, where p is the number of independent variables in the model (4 or 8).

Results from random forests for all individual gesture trees are presented in Figure 22 and Figure 23. Each panel depicts metrics of variable impact in the form of the mean decrease in the Gini coefficient and mean decrease in accuracy. The mean decrease in the Gini coefficient measured how much each variable contributed to the purity or homogeneity of the nodes and corresponding final classification leaves (Hastie et al., 2009; Sandri & Zuccolotto, 2008).

Variables with a greater mean Gini coefficient are considered more important. Mean decrease in accuracy measured how much the overall accuracy would decrease if that variable were not included in the model. The metrics mentioned above were used to corroborate the relative importance of the explanatory variables depicted in each classification tree.

Across all random forests that used individual ESCS gestures as predictors, behavior regulation points were the most important predictor, and their removal from the model would greatly reduce the accuracy of the model.

Apart from the ADOS SA random forest, all the remaining random forests that contained ESCS domain scores as predictors confirmed that responding to joint attention was the most important predictor. The ADOS SA random forest's most important variable was initiations of behavior regulation, which again matched results from its corresponding classification tree. In the JA factor and SCU trees with domains as predictors, RJA and IBR seemed equally as important. In examining their mean decrease in accuracy metric, both IBR and RJA variables exhibit strong separation from the other variables, implying that permuting only one would not obstruct the model.

All forests, however, were not especially strong leaners, as exhibited by their OOB errors presented under each panel of variable importance metrics in Figures 22 and 23. Despite relatively high OOB error rates, the variable importance in each tree was further corroborated in their respective random forests, thus providing additional confidence in decision tree interpretation.

Discussion

The proliferation of communication-based autism interventions and their overarching aim to ameliorate the challenges of limited verbal ability and difficulty with social communication

have been an important and burgeoning area of research. However, numerous barriers have hindered more rapid progress, especially for minimally verbal children. These barriers generally encompass the paucity of appropriate measurement tools that assess social communication behaviors accurately. Given these barriers to accurate measurement among minimally verbal children, current assessments may miss children who truly make progress and identify others who actually do not within an intervention protocol (Grzadzinski et al., 2020). This measurement gap and barriers for minimally verbal children served as the impetus for the current study, and specific solutions derived from this paper are discussed below.

The present study offers several primary findings regarding the presentation of early social communication skills. First, it was determined that an item response theory framework could be used to score and analyze the ESCS assessment to extract additional meaning, especially regarding its creation of an objective hierarchy of social communication gesture difficulty.

Second, machine learning algorithms can connect information from separate assessments measuring related or similar constructs to consolidate information and return a comprehensive evaluation of skills.

Finally, by fitting appropriate models, this study contributes inherent methodological value. It was demonstrated that these statistical methods used in tandem might provide more insight than using one method alone.

Together, this study offers an immediate and practical solution to the lack of appropriate social communication measures that can be used among minimally verbal children.

Utilizing an IRT Framework for Sensitive Measurement of Social Communication

Overall, IRT analyses added to the measurement and intervention literature by providing supplementary information regarding social communication gesture presentation, which have already been identified as salient predictors of later language (Mundy, 1990). More detailed analyses of these developmentally upstream gestures portrayed additional variability within the minimally verbal population. Study results not only aid in understanding the difficulty of gesture presentation, but the continued use of these analyses may also aid in retrospectively differentiating between preverbal and those who remained minimally verbal.

Evidence of Subpopulations within the Minimally Verbal Status

As opposed to the CTT approach that scored ESCS based on the sum of observed behaviors, IRT models implemented in this study effectively modeled the relationship between a child's underlying latent social communication ability and their presentation of various joint attention and behavior regulation gestures. Therefore, IRT scored children's social communication behaviors directly on that latent scale.

IRT has been used extensively in the educational testing context, where it is assumed that the underlying latent trait follows a normal distribution. More recently, IRT has been applied to psychiatry to explain the relationship between symptomatology and the diagnosis of a given psychiatric disorder (Aggen et al., 2005; Chan et al., 2004). Similar to psychiatric symptom batteries, assessments that measure social communication among minimally verbal children with autism often exhibit floor effects, with many children not demonstrating gestures. A zero-inflated negative binomial regression model fitted the best out of all competing models, suggesting it can be a suitable model for researchers studying count data within this population.

By definition, the best-fitting ZINBM confirms what clinicians and researchers have suspected all along, that there is untapped variability within the minimally verbal status and subpopulations exist, at least within this sample and definition of minimally verbal status.

Comprehensive Difficulty Hierarchy of Social Communication Gestures

For intervention research, identifying ability at the start of intervention may influence the tailoring of treatment, i.e., implementing a more structured or naturalistic approach. Additionally, entry characteristics like joint attention and social responsiveness are strong predictors in treatment response from behavioral intervention (Kasari et al., 2012; Sallows & Graupner, 2005). In this study, one would assume that strict inclusion criteria would result in a relatively homogeneous sample at this pre-intervention stage, and for the most part, it does; however, the ZINBM suggests subtle differences in the sample may influence subsequent intervention strategies.

BR Gestures and JA Eye Gaze as Initial Targets for Children with Limited Skills. The zero-part coefficients of the model indicated that the five easiest gestures were BR reaches, BR gives, JA eye contact, BR language, and BR eye contact. These skills mostly comprised of behavior regulation skills and were technically the least likely to exhibit a zero count. Clinically, these skills could be interpreted as the first gestures that will eventually be demonstrated for children who do not currently present with skills. Therefore, for clinicians and researchers targeting social communication gestures to improve verbal language, information from this statistical model may offer initial targets for many children, especially for those with no existing joint attention skills.

JA Gestures are Especially Difficult. In the count part of the ZINBM, the first four easiest skills were behavior regulation skills. Surprisingly, BR eye contact was the third most difficult

gesture, and this was suspected due to its likely pairing with other gestures like reaches, which would code its frequency with that other gesture. On average, joint attention gestures were more difficult than behavior regulation gestures. Joint attention shows and gives were the most difficult and exhibit vast ranges of perceived difficulty, as seen in the item order map in Figure 17.

In a typical IRT model, JA shows and gives would likely be removed from the model as they are not adding much information due to low frequency and high difficulty. However, in this study, it was intentional to include all gestures to maintain the integrity of the ESCS and provide a compressive appraisal of gesture difficulty.

Portraying "Levels" of JA and BR Gestures on the Same Scale

The item parameters from the count part of the ZINBM gave an objective hierarchy of gesture difficulty and are depicted in Table 7 and Figure 17. These results corroborate observed developmental trends from extant literature, namely that young children are potentially more adept at using social communication gestures to request objects than they are able to use their joint attention gesture counterparts to orient attention to an object or event (Mundy, 1995). However, this study adds to the social communication literature by objectively comparing and definitively ranking social communication gestures with an IRT framework, thus potentially giving more detailed information regarding developmental trends among minimally verbal children.

Previous research has classified social communication gestures as high or low-level behaviors, where conventional gestures such as pointing and showing would be considered high-level gestures, and low-level gestures were typically the number of eye contacts and alternates (see Mundy et al., 1994; Schietecatte et al., 2012). It has been of interest to differentiate these

gestures as they may be associated with distinct behaviors. Higher-level gestures may involve other aspects of development, such as social motivation, and may be predictive of other skills, such as theory of mind (Baron-Cohen, 1989). Lower-level gestures do not require the same degree of social communication ability and, therefore, are less able to differentiate children with autism from their typically developing peers (Mundy et al., 1986; Charman, 1998). This study built upon that foundation of differentiating gesture "levels" but gave a more precise and detailed analysis by placing gesture difficulty along a united continuum via IRT, which directly operationalized the level of every gesture.

Results indicate that the lower and higher-level behaviors could instead be re-defined as more difficult and easier behaviors for this minimally verbal population. JA shows and gives could be one classification and represent the "very difficult" set of gestures. BR eye contact and JA points cluster together on the item order map in Figure 17 and offer separation from other items; therefore, these items could be considered "difficult" gestures. "Medium difficulty" gestures could include JA language, JA eye contact, and BR gives. "Easy" gestures would include BR points and BR language, and the "easiest" gesture would be a BR reach. Defining these levels more clearly may provide information regarding the quality of social communication gestures in intervention studies, an aspect of autism research that has largely been ignored (Lawton & Kasari, 2012). Future research could investigate replicating the results of the item order map in other samples of minimally verbal children.

RJA and RBR are Significantly Associated with Initiation Ability

Previous research suggests significant correlations exist between IJA and RJA skills of children with ASD (e.g., Bono et al. 2004; Dawson et al. 2004; Siller & Sigman 2008). Although theta scores represent the overall ability to initiate social communication, it was important to

determine whether these correlations with RJA and RBR existed in this sample. Theta values were significantly correlated with both domain values. However, a stronger correlation of .44 existed with RJA, which is consistent with prior research.

Although the direction of this association cannot be extracted from these analyses, the strong connection between RJA and initiation ability may be an important avenue to consider for future intervention research.

Redefining Scoring of Social Communication Gestures

The hierarchy of gesture difficulty derived from the IRT analysis may help redefine what researchers consider improvement is within these initiating behaviors. Since gestures exhibit various difficulty levels, differentially weighing each additional count across gestures may make sense. At pre-intervention or even in the early stages of the intervention timeline, analyzing ESCS data with an IRT framework may provide adjunct information to determine which children are presenting differently from one another, which may influence individual modes of treatment. Additionally, the results of the final ZINB IRT model suggest that an IRT scoring method may help guard against scores that are inflated by easier gestures, thus, inaccurately portraying skills of the child.

Justification for Increased ESCS Use

Although the ESCS is touted as a useful outcome measure (Anagnostou et al., 2015), early intervention research has demonstrated mixed results regarding its sensitivity to change over time (see Ingersoll, 2012; Kasari et al., 2006; Lawton & Kasari, 2012; Murza et al., 2016). Additionally, Grzadzinski et al. (2020) suggest that more research is needed to determine the capability of the ESCS to capture changes in early intervention trials. This uncertainty may explain why the ESCS is not often used in intervention trials despite its direct measurement of

joint attention, which is often the target construct. This study strongly endorses that the gestures measured by the ESCS are crucial in examining the skills of minimally verbal children and sought to improve how it is analyzed.

Although only modeled on baseline data, the successful use of the IRT scoring procedure may be a straightforward solution to this measurement dilemma and offer a method that captures the subtleties and sensitive change in intensive but short intervention protocols (Krstovska-Guerrero & Jones, 2016; Nordahl-Hansen et al., 2016).

Machine Learning Algorithms to Connect Information Across Assessments

The CART models determined which cut points in gesture frequencies were especially influential in predicting membership to more robust language outcomes at baseline. On average, responding to joint attention nearly one-third of the time was enough to classify children into high language score groups from standardized assessments. Among the trees that included individual gestures, one behavior regulation point was a prominent cut point that was able to classify children. Overall, at baseline, only a few skills are needed to classify children, which directly helped reorganize the clinical meaning behind these frequencies. Although CART is becoming increasingly popular in the field, using these algorithms to gain more insight into the scaling of social communication gesture frequencies had not been done.

Cut-points in ESCS Frequencies that Relate To Standardized Measures

In the construction of the trees, all were pruned according to Breiman's "one standard error" rule, which was a method to increase accuracy by removing sections of the tree that provided little power to classify ESCS frequencies into the outcome variables. However, as a result, many of the trees were pruned to stumps, which included only one node immediately connected to the leaf/classifier. Consequently, in most of these trees, the prediction was able to

occur based on only one value of a single variable. Despite these stumps, it was determined that prioritizing accuracy was more important than creating potentially more complex and robust trees.

Individual Behavior Regulation Gestures as a Differentiating Feature. In the trees that used individual ESCS gestures as predictors, the most important input feature was behavior regulation points, with a crucial cut point occurring at only one demonstration of this gesture amongst all the decision stumps. For the number of word roots, spontaneous communicative utterances, caregiver's perception of expressive and receptive language, and joint attention factor scores from the ADOS, one behavior regulation point alone was able to classify children into respective outcome groups. Behavior regulation points were considered the third easiest gesture via IRT, suggesting that the machine learning algorithm determined that enough variability existed across the sample to discriminate participants at baseline and classify them into various language outcomes.

Although results from the decision trees slightly veer from the initial aim of attributing additional meaning to ESCS individual gestures frequencies, they provide insights into methods to further characterize this population.

Implications for Intervention. In this study, all children were recruited from various studies that utilized a naturalistic developmental behavioral intervention known as JASPER (Joint Attention Symbolic Play and Engagement and Regulation; Kasari et al., 2008, 2010). JASPER has been shown to improve spoken language through play routines that boosted engagement and social communication (Kasari et al., 2008) and was deemed an ideal targeted intervention for these minimally verbal children in this study. The differentiating characteristic of JASPER is that it explicitly targets joint attention when other interventions do not. However,

these results indicate that behavior regulation gestures and behavior regulation language may be more important targets than initially thought. The examination of the BR frequencies at baseline may better characterize the skills of minimally verbal children, and could be important initial targets that facilitate joint attention.

This study cannot comment on changes in skills over time; however, its initial results indicate that looking towards the increase in behavior regulation gestures over short periods of intervention may be a more sensitive indicator of improvement. Almirall and colleagues (2016) found that within an adaptive intervention protocol that implemented JASPER, initiations of behavior regulation gestures improved in all groups at the 12-week, stage 1 timepoint. Therefore, behavior regulation gestures may be thought of as developing in concert with joint attention gestures in this population. The information from decision trees indicates that behavior regulation gestures may be a more sensitive metric to help inform optimal treatment routes at pre-intervention or very early stages of interventions compared to joint attention gestures or overall language ability.

RJA is Predictive of Language Ability. Among the vast majority of classification trees with ESCS domain scores as predictors, RJA was the first splitting variable and was, therefore, the most important variable in classifying outcomes. RJA's importance in classification trees corroborates results from numerous studies where a significant relationship between RJA and expressive language was found (Kasari et al., 2008; Murray et al., 2008; Pickard and Ingersoll, 2015; Schietecatte et al., 2012; Sigman et al., 1999; Yoder et al., 2015). For the outcomes that directly measure language, either objectively via the number of different word roots or subjectively via parent report, RJA was not only determined as an important predictor of verbal language but the *only* splitting variable, resulting in tree stumps.

Results from these classification stumps support findings from Siller and Sigman (2008), where the longitudinal change in the rate of language growth among 28 children with autism during early and middle childhood was significantly predicted by children's responsiveness to bids of joint attention. Similarly, in a sample of 29 4-year-olds who received 24 hours of various communication-based interventions per week for one year, children with well-developed RJA at the beginning of the study developed significantly more significant language gains than children with less-developed RJA (Bono et al., 2004).

Reframing Scoring Expectations in Minimally Verbal Children

The examination of classification trees that included domain scores as predictors were of particular interest as performance on the ESCS is typically presented in this manner. These initial results begin to defuse preconceived notions of the performance within social communication gesture presentation and reframe what it means to demonstrate a skill well in this population pre-treatment. Classification tree results indicated that only relatively modest scores were necessary among this sample to sort children into more robust language classifications. These specific cut points detailed in the trees can be considered inflection points, where more skills past these points are not necessary for classification purposes.

Per the classification stumps, responding to bids of joint attention at least one-third of the time was enough to classify children into more robust language ability subgroups. Therefore, children who missed half of the opportunities to respond to bids of attention, which may erroneously be considered a poor outcome, would be classified well into a more robust language outcome at baseline.

In the trees that include splitting variables beyond RJA, initiation of behavior regulation gestures was the second most important set of skills, and their respective cut points ranged from

11-15 frequencies. Since domain scores include frequencies from all gestures within that respective construct, these ranges may feel modest, especially to parents or other researchers unfamiliar with the population. However, CART analyses reveal that these frequency delineations are much more important than they may appear. For example, the first tree in Figure 21 predicted more/less robust spontaneous communicative utterances. The third most important variable behind RJA and IBR was initiating joint attention gestures, with a cut point of 5.

In general, these cut-points tell us a few things. First, children are not initiating many joint attention gestures, which is expected at baseline. Next, the influential cut points in CART analyses reveal that 5 IJA may be more significant than initially thought. Even though more initiations of joint attention are inherently beneficial for more positive language and social development (Sigman et al., 1999, 2005), this study begins to unearth a target frequency that is especially predictive of classifying children and reveal an anchor into what may be considered robust scores.

Contribution to Joint Attention Theory. A growing body of literature has suggested that early social attention and later social cognitive skills exist on a continuous axis (Baron-Cohen, 1989; Mundy, 2018; Mundy & Sigman, 1989). This study adds a crucial and largely unexplored portion of that continuous axis by beginning to understand social communicative gesture presentation among a minimally verbal sample more thoroughly. Although this study cannot comment on the longitudinal presentation of gestures, initial CART and IRT results may help redefine what realistic targets are within the ESCS and, subsequently, a better idea of what progress may look like in an intervention study.

Final Considerations

The autism intervention literature has demonstrated that a single intervention is unlikely to be effective for all children of a given sample (Kasari & Smith, 2013). Although more challenging to execute, an individualized approach may more accurately address the heterogeneous nature of autism. Newer approaches that propose a sequence of interventions based on individual behavioral responses may be most beneficial for individual outcomes (Almirall et al., 2014; Collins et al., 2004, 2014). These tailored interventions that consider a child's baseline characteristics and treatment response history may maximize optimal outcomes in a given domain (Costello & Maughan, 2015; Georgiades & Kasari, 2018). However, there is a lack of guidance in determining treatment routes, forcing experts to rely on their clinical judgment to determine if a treatment should be augmented (Steidtmann et al., 2013).

This study contributes a concrete method to help determine treatment routes and subsequent gains. The IRT results explained that the difficulty of initiation gestures can vary, which provided additional insights into defining the quality of gestures. The CART results attached additional clinical significance to gesture frequencies, thus, improving the understanding of gesture quantity. Together, these detailed facets of social communication can be used alongside clinical judgment to help illuminate what may work best for minimally verbal children with ASD.

Limitations and Future Directions

This study is a crucial step in better characterizing minimally verbal children with autism via advanced quantitative methods. However, like with every study, this one is not without its limitations. This study is a secondary data analysis; therefore, the measures selected for this study were limited to those common across original projects that comprised the current sample. Although all assessments in this study measure social communication, it has been demonstrated

that specific assessments can affect sample characteristics among this minimally verbal population (Bal et al., 2016). Therefore, analyses should be interpreted with caution as these are only a handful of assessments that have been used with the minimally verbal population.

The measurement limitation posed by this secondary analysis raises larger issues in the field: not only is there a dearth of reliable and valid assessments for minimally verbal children, but the field also lacks consensus regarding which existing assessments to use among this population (Kasari et al., 2013; Fletcher-Watson & McConachie, 2017). These issues exist in the broader context of intervention research and its variable participants and are not isolated to the minimally verbal population. Recent reviews suggest that up to 289 unique measurement tools were used in intervention studies, with 62% used in only one study and none used in more than 7% of the studies (Bolte & Diehl, 2013). Ideally, the field will soon establish more sensitive and agreed-upon measures across RCTs, especially those involving minimally verbal children.

There are several limitations in interpreting results. First, IRT models included joint attention and behavior regulation gestures under the same construct of social communication initiations. Theoretically, this decision is justified, but future research may explore a bifactor model if the sample size is large enough. Logical expansions to this initial IRT model pose for exciting future research. For example, future studies may want to consider various sources of differential item function (DIF) like cognitive ability, which would determine different probabilities of gesture presentation depending on group membership. Theta values, or a child's social communication ability, derived from IRT models may also be used as adjunct independent variables in researchers' analysis of intervention effects.

While mentioned in the results, it is worth reiterating the concern regarding the stability of the classification trees. Despite relatively weak learners, these CART models are primarily

tools for exploratory discovery rather than hypothesis testing, and therefore results and prediction accuracies are deemed acceptable given the heterogeneous minimally verbal population. Additionally, influential splits in the CART models were corroborated by the importance metrics derived from the random forest models, which helped bolster the confidence in interpreting the findings of the classification trees. Influential cut points found in these models are unfortunately isolated to this sample and cannot be generalized to other samples of minimally verbal children.

Overall, the current study was considered exploratory and primarily aimed to determine if these statistical methods extracted additional and meaningful information from the ESCS regarding social communication presentation. For that reason, only baseline data were examined cross-sectionally. The next logical step is to expand IRT and CART analyses to accommodate the longitudinal nature of these data, specifically, to determine how individual gesture difficulty changes over time, if influential cut points remain stable across timepoints, or if certain characteristics throughout an intervention are better able to predict treatment response.

Conclusion

The main strength of this paper is the thorough examination of joint attention as measured by the ESCS measurement tool. In terms of behaviors, joint attention may be the smallest unit of measurement that predicts language outcomes. However, in minimally verbal samples, it still may not be sensitive enough to differentiate children; therefore, children may be deemed similar when differences likely exist. This study acknowledges that investigating the heterogeneity within minimally verbal children may be especially urgent as the stakes for efficacious interventions are high.

The utility of the paper is two-fold: (1) to improve the characterization of minimally verbal children by thoroughly examining their social communication behaviors and (2) to explain the benefit of the novel quantitative methodologies used to examine those behaviors. Regarding its methodological contributions, this paper determined that IRT models and machine learning algorithms can be repurposed to examine behavioral data to offer additional value in analyzing social communication gesture presentation.

Overall, this study demonstrated that repurposing advanced statistical techniques can extract additional variability from existing assessments, thus, providing a meticulous evaluation of the heterogeneity of social communication presentation within minimally verbal children. With the improved characterization of minimally verbal children, we can move towards the next phase of tailored interventions and understanding their treatment effects among subpopulations of the minimally verbal status.

Tables and Figures

Table 1. *Child and Parent Demographics*

Child/Parent Characteristics N (%)	N = 457
Chronological Age (Months): Mean (SD)	64.1 (19.9)
Gender	
Female	75 (17.2%)
Race/Ethnicity	
African American/Black	59 (13.6%)
American Indian/Alaska Native	1 (0.2%)
Asian/Pacific Islander	59 (13.6%)
Caucasian	216 (49.7%)
Native Hawaiian/Pacific Islander	2 (0.5%)
Other/Not Disclosed	51 (11.7%)
More than one race	47 (10.8%)
Ethnicity	
Hispanic or Latino	104 (22.7%)
Not Hispanic or Latino	335 (73.1%)
Other/Not Disclosed	19 (4.1%)
Mother's Education	
Less than 7th grade	7 (1.6%)
Junior High	11 (2.4%)
Some High School	13 (2.9%)
High School graduate	49 (10.9%)
Special Training after High School	21 (4.7%)
Some College	98 (21.8%)
College graduate	156 (34.7%)
Graduate/Professional Training	95 (21.2%)
Father's Education	
Less than 7th grade	8 (1.9%)
Junior High	10 (2.3%)
Some High School	22 (5.1%)
High School graduate	72 16.7%)
Special Training after High School	23 (5.3%)
Some College	75 (17.4%)
College graduate	121 (28.1%)
Graduate/Professional Training	99 (23%)

Table 2. *Descriptives of ESCS Individual Gesture Frequencies*

	JAec	**JAlang**	**JApoints**	**JAgives**	**JAshows**	**BRec**	**BRlang**	**BRpoints**	**BRgives**	**BRreaches**
Mean	3.13	1.81	0.669	0.0618	0.0221	1.29	3.87	2.03	3.29	6.83
Median	2	0	0	0	0	0	2	0	2	6
Standard deviation	4.34	3.60	2.14	0.353	0.147	2.26	4.69	4.24	3.43	5.03
Minimum	0	0	0	0	0	0	0	0	0	0
Maximum	32	27	24	4	1	14	33	26	20	25

Table 3. *Descriptives of ESCS Domain Scores*

	IJA_TOTAL	**IBR_TOTAL**	**RJA_TOTAL**	**RBR_TOTAL**
Mean	5.70	16.6	47.4	38.0
Median	3	15	46.0	35.0
Standard deviation	6.82	10.3	29.5	27.0
Minimum	0	0	0.00	0.00
Maximum	51	71	100	100

Table 4. *Descriptives of Outcomes used in Classification Trees*

	LS_NDWR	LS-SCU	ADOS-JA-Factor	ADOS-SA	Vineland-Receptive	Vineland-Expressive
Mean	7.28	11.1	8.92	14.7	22.2	21.8
Median	3.00	4.50	9.00	15.0	21.0	20.0
Standard deviation	9.97	15.8	2.27	3.59	10.7	11.0
Minimum	0	0	0	0	4	2
Maximum	62	102	12	24	73	72

Table 5. *Fit Statistics of IRT Models*

Model	AIC	BIC	Log-likelihood
Person-only RPCM	27301.1	27313.9	-13648.5
RPCM	19169	19239.6	-9573.5
Negative Binomial	14819.6	14896.7	-7397.8
Zero-Inflated Negative Binomial	14328.7	14427.48	-7140.351

Table 6. *Item Parameters of IRT Models*

Fixed Effects	**RPCM**		**NBRM**		**ZINB**	
	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error
JAec	0.94	0.04	0.91	0.07	1.19	0.06
JAlang	0.39	0.05	0.23	0.08	1.17	0.08
JApoints	-0.61	0.07	-0.82	0.09	0.68	0.12
JAgives	-2.99	0.19	-3.01	0.21	-0.81	0.52
JAshows	-4.02	0.32	-4.05	0.34	-2.66	0.72
BRec	0.05	0.05	0.01	0.08	0.44	0.09
BRlang	1.15	0.04	1.08	0.07	1.41	0.06
BRpoints	0.50	0.05	0.31	0.07	1.41	0.08
BRgives	0.99	0.04	1.10	0.07	1.22	0.06
BRreaches	1.72	0.04	1.88	0.07	1.92	0.05

Zero-Part Coefficients		
	Estimate	Std. Error
(Intercept)	-2.33	0.35
JAlang	2.45	0.36
JApoints	3.73	0.41
JAgives	5.09	0.72
JAshows	3.83	1.34
BRec	0.96	0.37
BRlang	0.05	0.34
BRpoints	2.81	0.37
BRgives	-1.04	0.45
BRreaches	-2.58	0.61

Table 7. *Item Easiness Rankings of ZINB*

Item	Difficulty	Rank
ItemBRreaches	1.92	10
ItemBRlang	1.41	9
ItemBRpoints	1.41	8
ItemBRgives	1.22	7
ItemJAec	1.19	6
ItemJAlang	1.17	5
ItemJApoints	0.68	4
ItemBRec	0.44	3
ItemJAgives	-0.81	2
ItemJAshows	-2.66	1

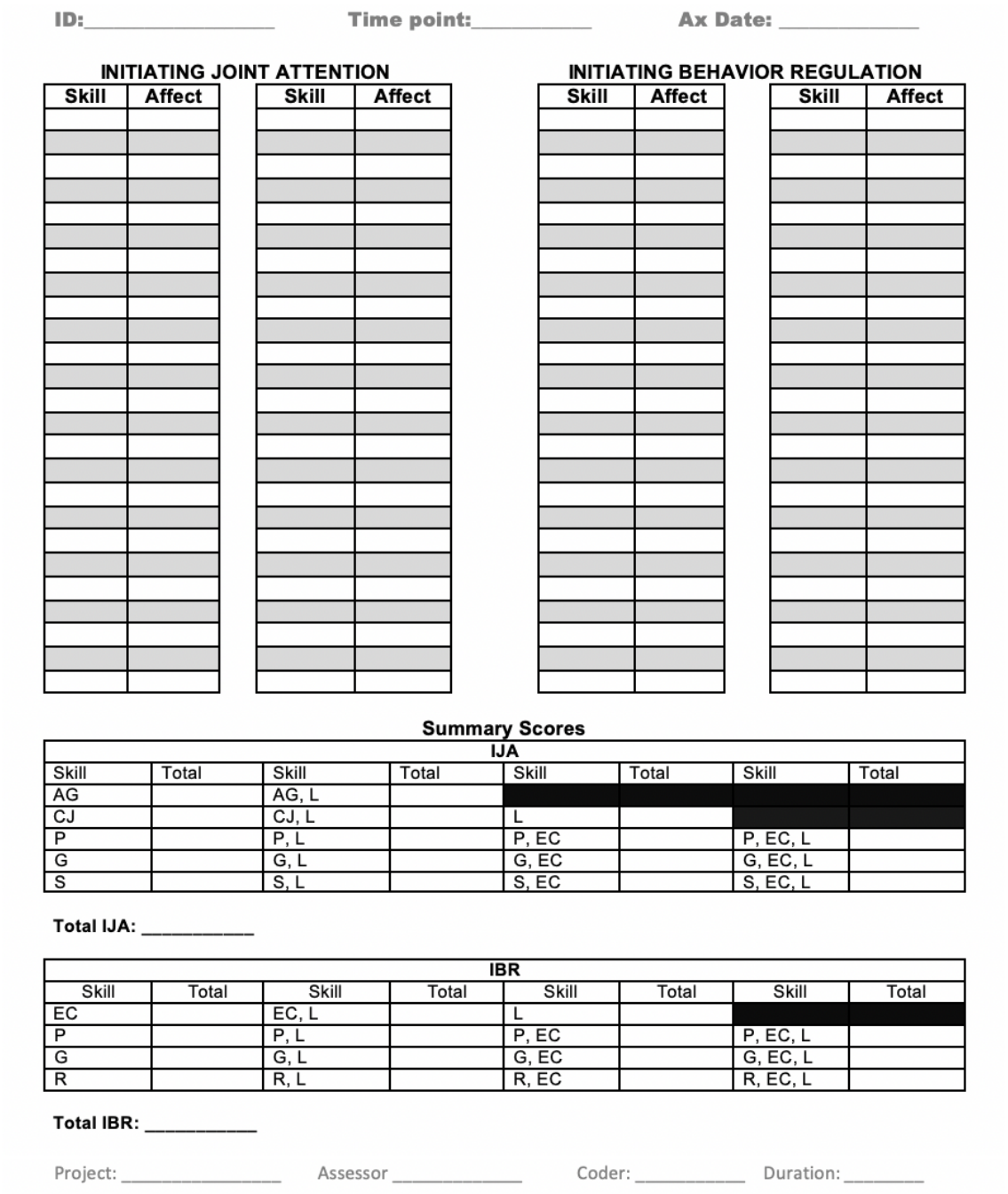
Table 8. *Theta Values and Domain Scores Correlation Matrix*

		RJA_TOTAL	RBR_TOTAL	Theta Scores
RJA_TOTAL	Pearson's r	—		
	p-value	—		
RBR_TOTAL	Pearson's r	0.34	—	
	p-value	< .001	—	
Theta Scores	Pearson's r	0.44	0.30	—
	p-value	< .001	< .001	—

Table 9. *Prediction Statistics on Test Data*

CART + Initiation Gestures	**Sensitivity**	**Precision**	**Accuracy**	**F1**
Vineland Receptive	0.84	0.64	0.67	0.73
Vineland Expressive	0.68	0.61	0.60	0.64
LS SCU	0.85	0.58	0.62	0.69
LS NDWR	0.84	0.69	0.71	0.76
ADOS SA	0.63	0.76	0.70	0.69
ADOS JA Factor	0.47	0.80	0.64	0.59
CART + Domains				
Vineland Receptive	0.33	0.66	0.56	0.44
Vineland Expressive	0.54	0.73	0.66	0.62
LS SCU	0.75	0.65	0.67	0.70
LS NDWR	0.56	0.71	0.65	0.63
ADOS SA	0.73	0.70	0.70	0.71
ADOS JA Factor	0.80	0.73	0.72	0.76

Figure 1. *ESCS code sheet*



ID:_____ Time point:_____ Ax Date: _____

INITIATING JOINT ATTENTION

Skill	Affect	Skill	Affect

INITIATING BEHAVIOR REGULATION

Skill	Affect	Skill	Affect

Summary Scores

IJA

Skill	Total	Skill	Total	Skill	Total	Skill	Total
AG		AG, L					
CJ		CJ, L		L			
P		P, L		P, EC		P, EC, L	
G		G, L		G, EC		G, EC, L	
S		S, L		S, EC		S, EC, L	

Total IJA: _____

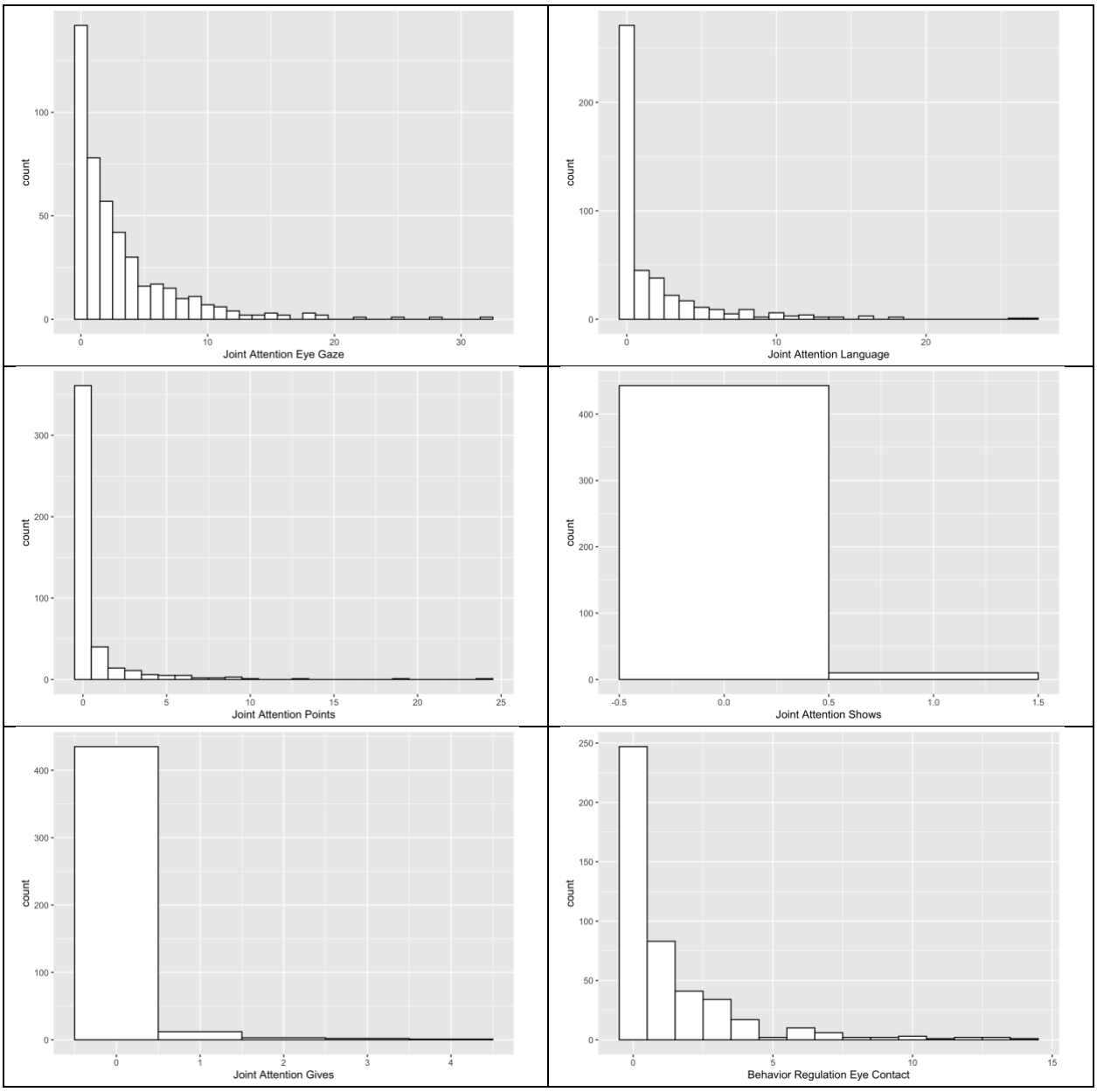
IBR

Skill	Total	Skill	Total	Skill	Total	Skill	Total
EC		EC, L		L			
P		P, L		P, EC		P, EC, L	
G		G, L		G, EC		G, EC, L	
R		R, L		R, EC		R, EC, L	

Total IBR: _____

Project: _____ Assessor _____ Coder: _____ Duration: _____

Figure 2. *ESCS frequency distributions by gesture*



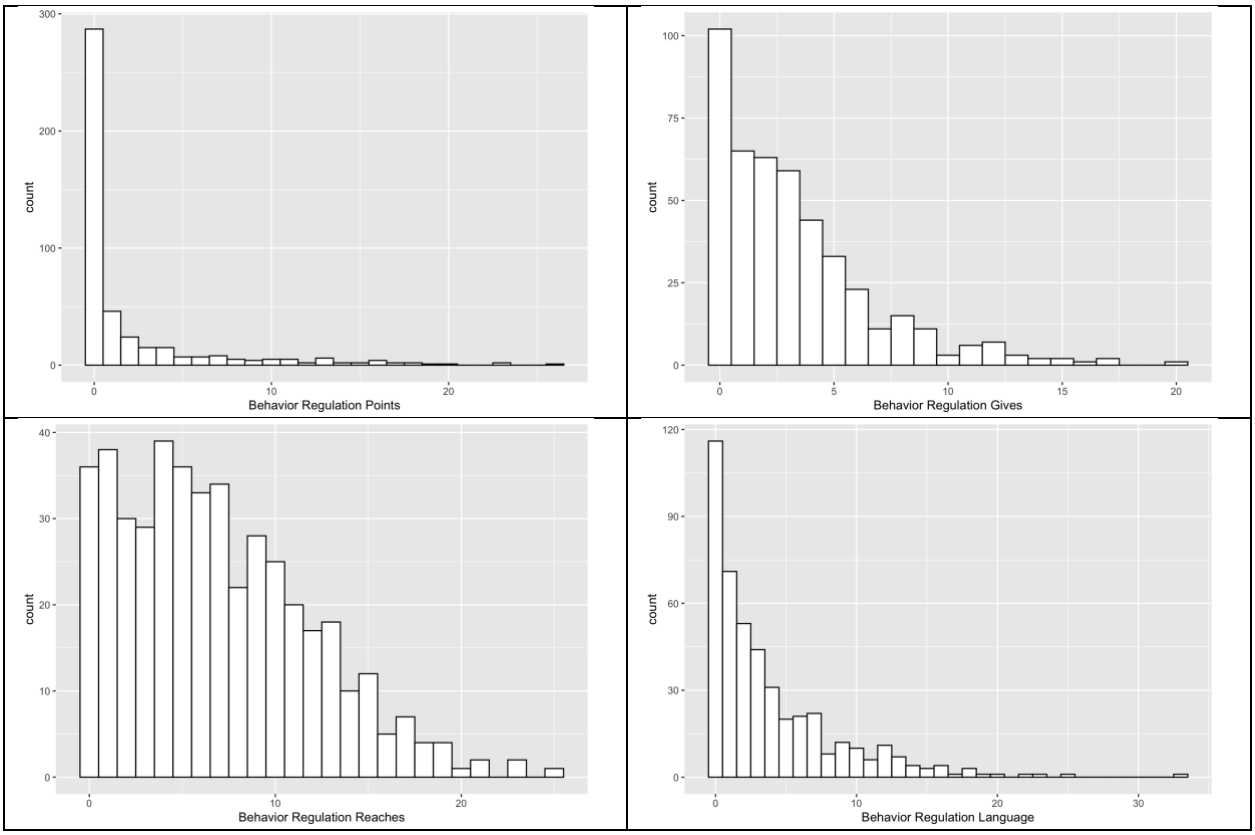

count
Behavior Regulation Points
count
Behavior Regulation Gives
count
Behavior Regulation Reaches
count
Behavior Regulation Language

Figure 3. *Dependent outcome frequency distributions*

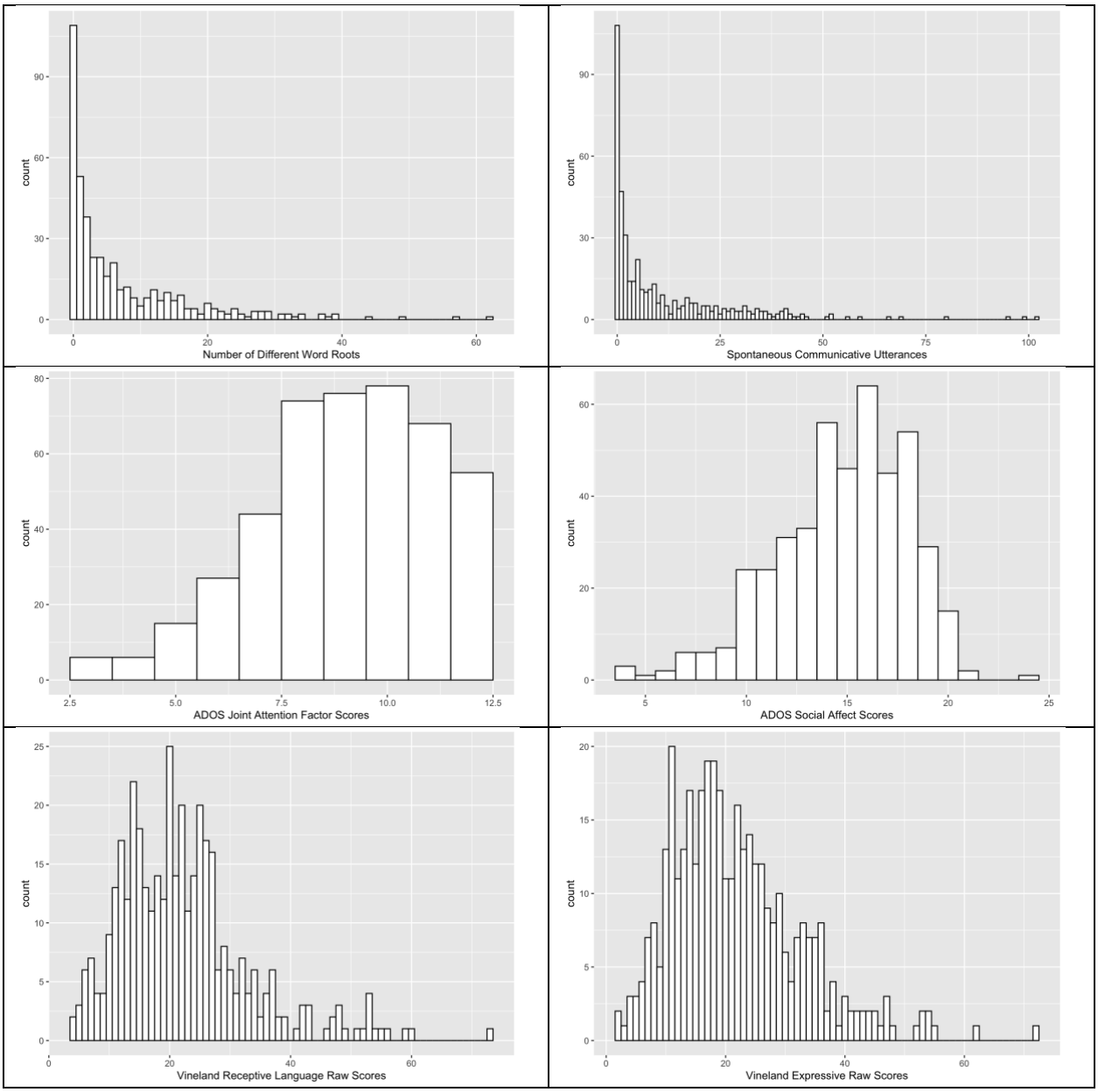


Figure 4. *Predicted values plotted against standardized Pearson residuals*

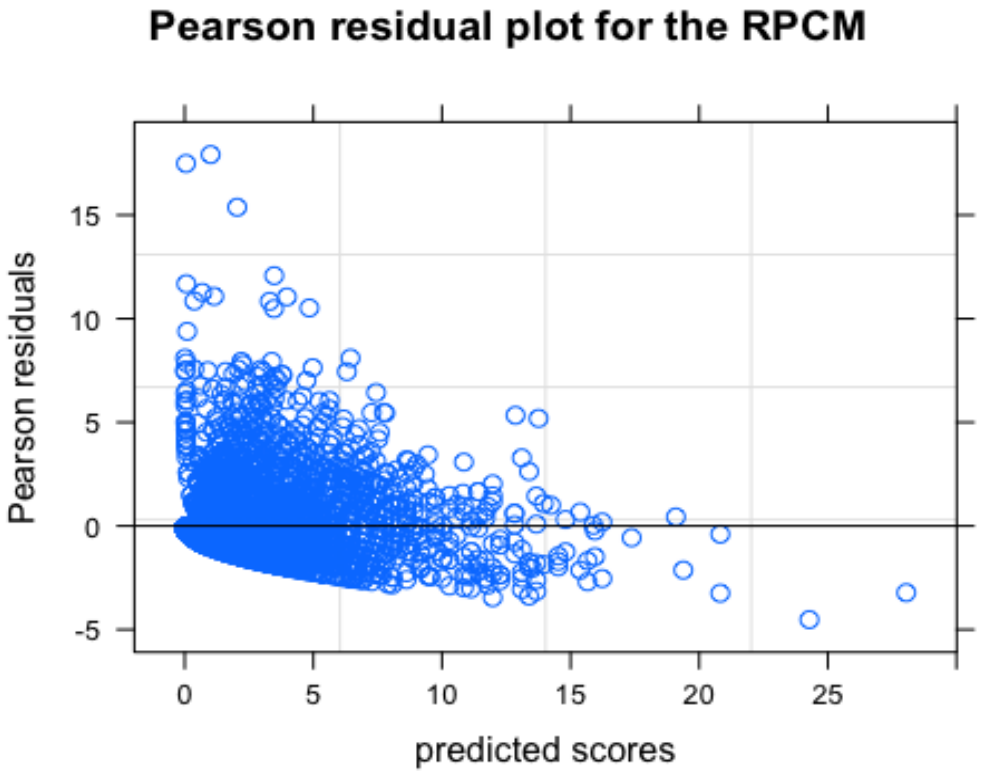


Figure 5. *Pearson residual boxplots for item-specific predicted scores of the RPCM*

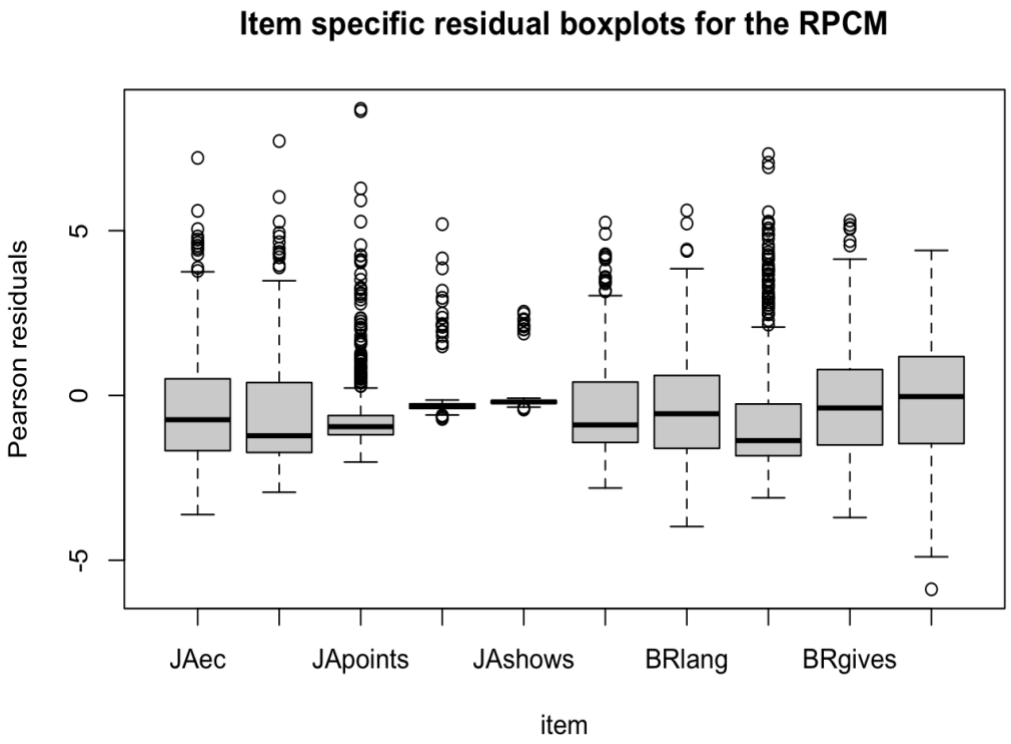


Figure 6. *Covariate adjusted frequency plots for IRT models*

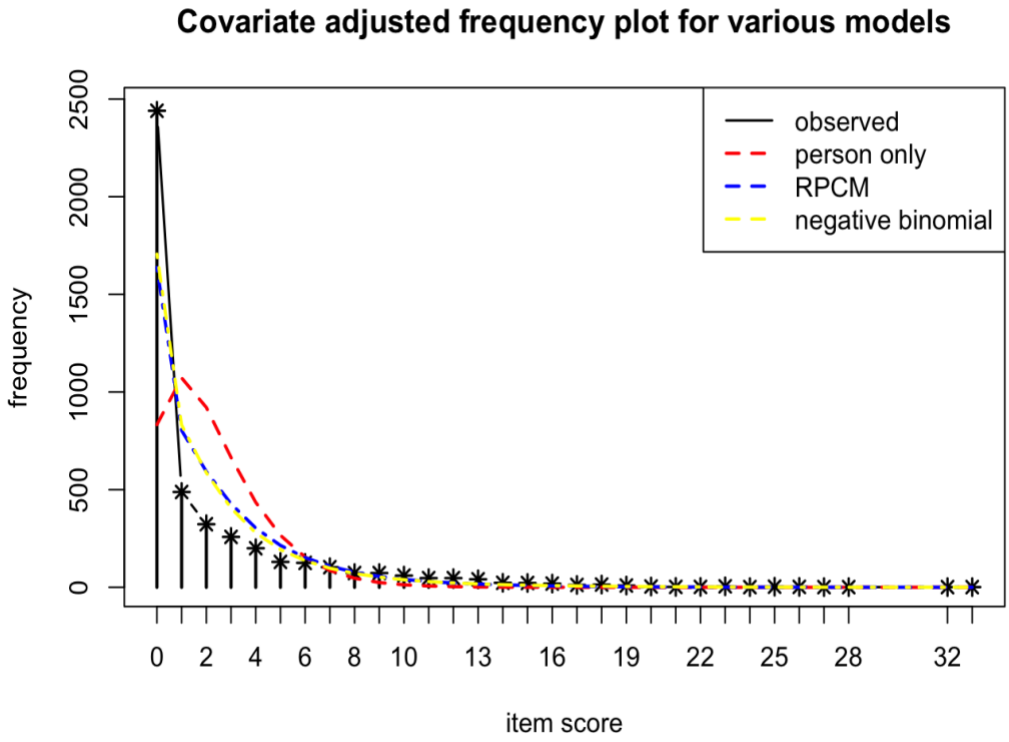


Figure 7. *Plots of the simulated scaled residuals of the RPCM*

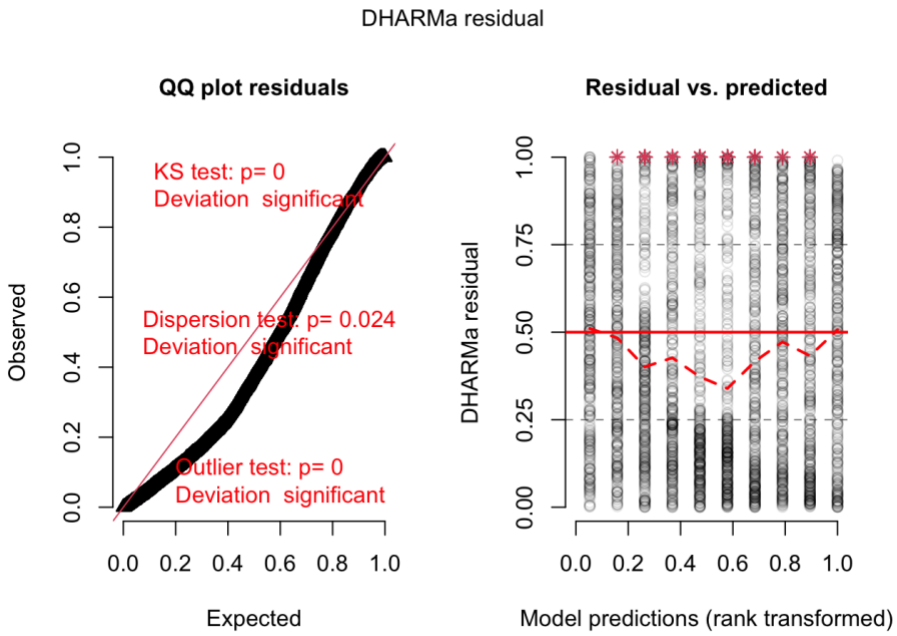


Figure 8. *Simulated dispersion test for the RPCM*

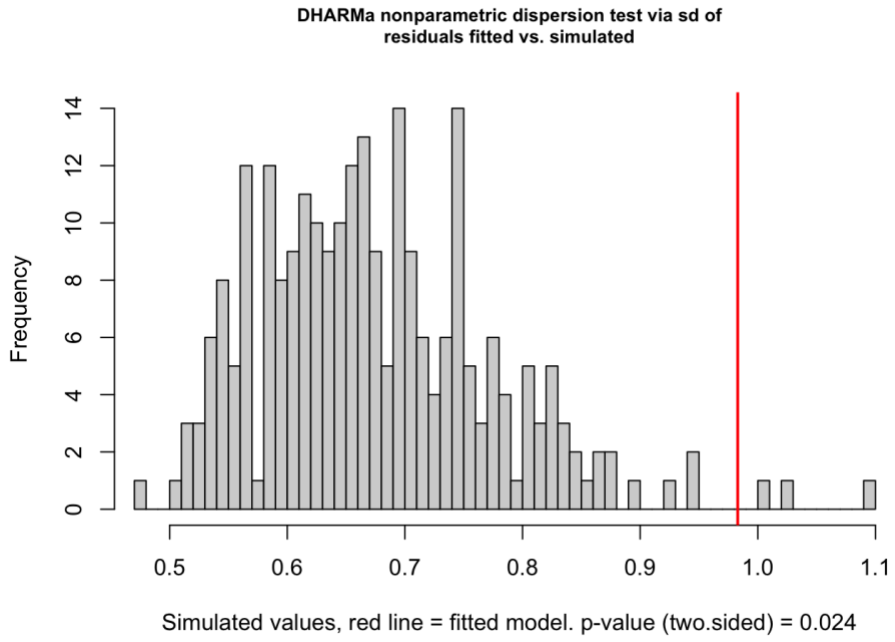


Figure 9. *Simulated zero-inflation test for the RPCM*

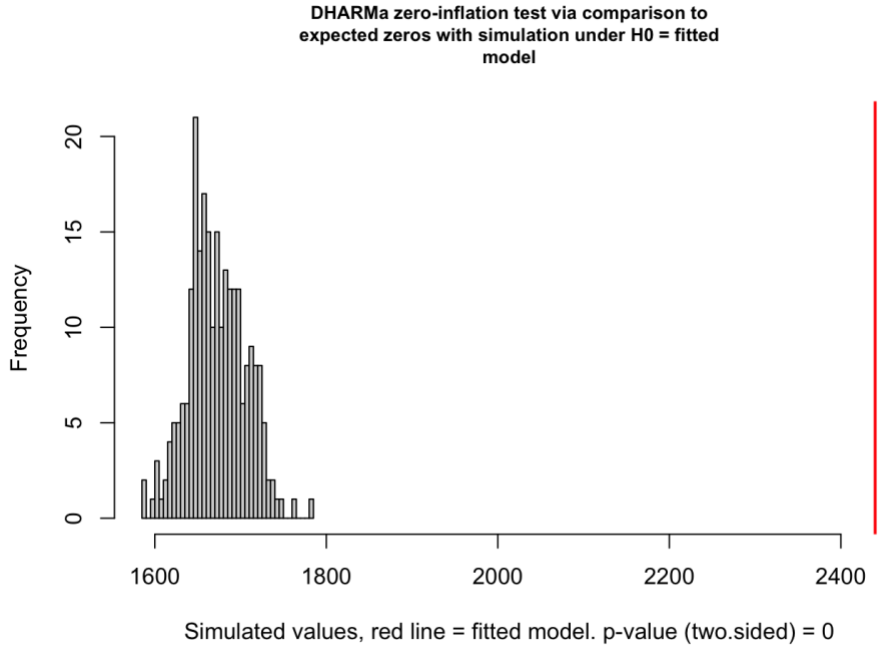


Figure 10. *Plots of the simulated scaled residuals of the NBRM*

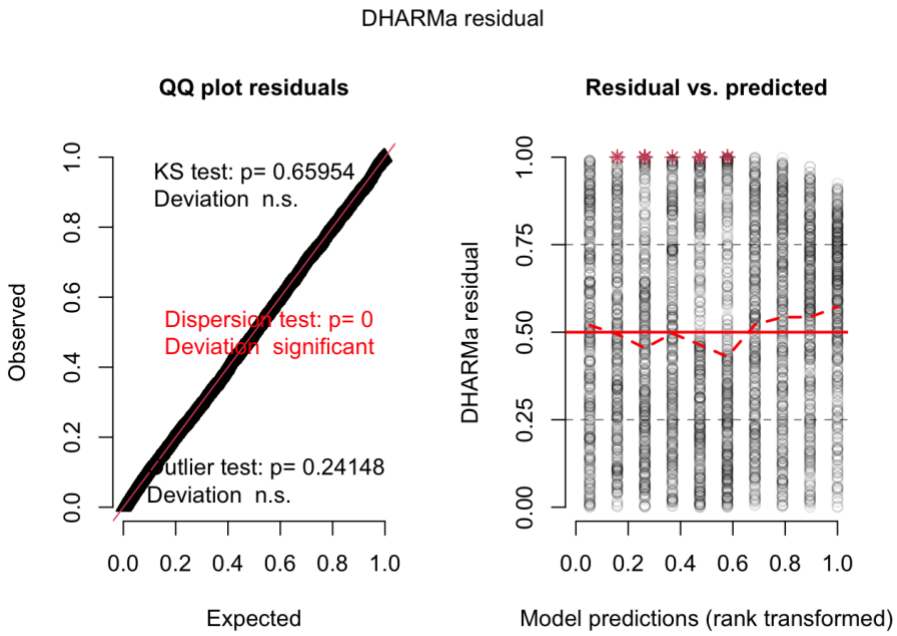


Figure 11. *Simulated dispersion test for the NBRM*

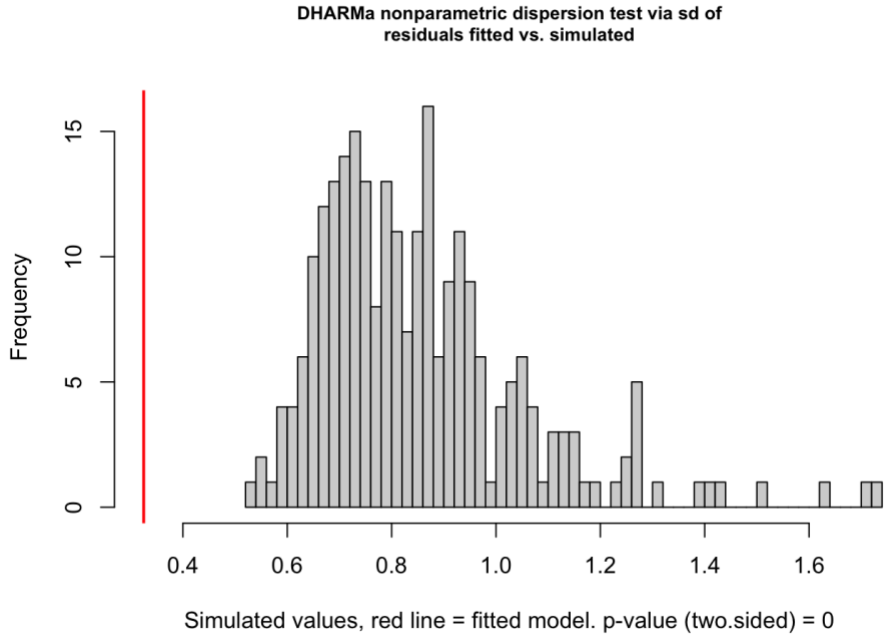


Figure 12. *Simulated zero-inflation test for the NBRM*

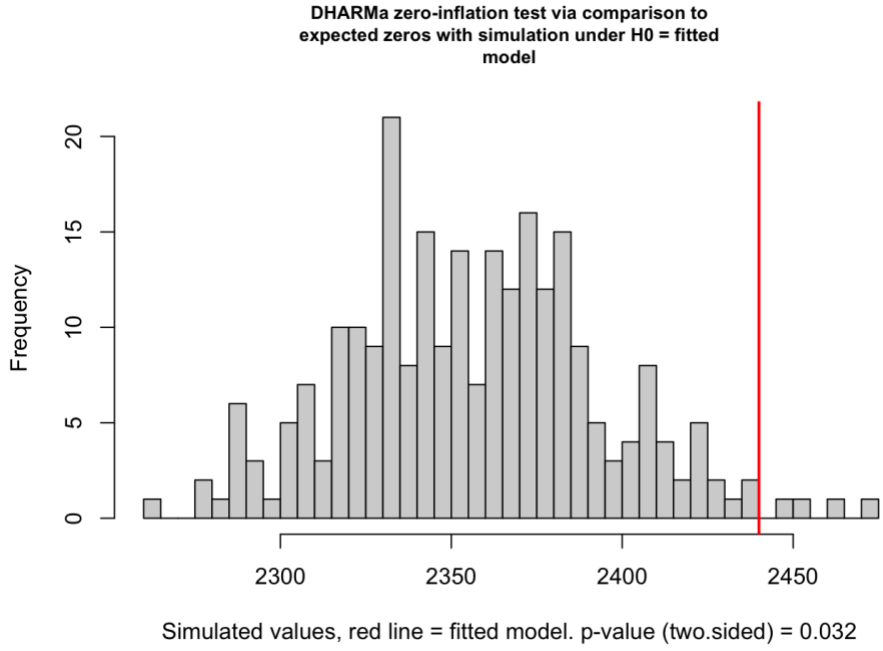


Figure 13. *Pearson residual boxplots for item-specific predicted scores of the ZINBM*

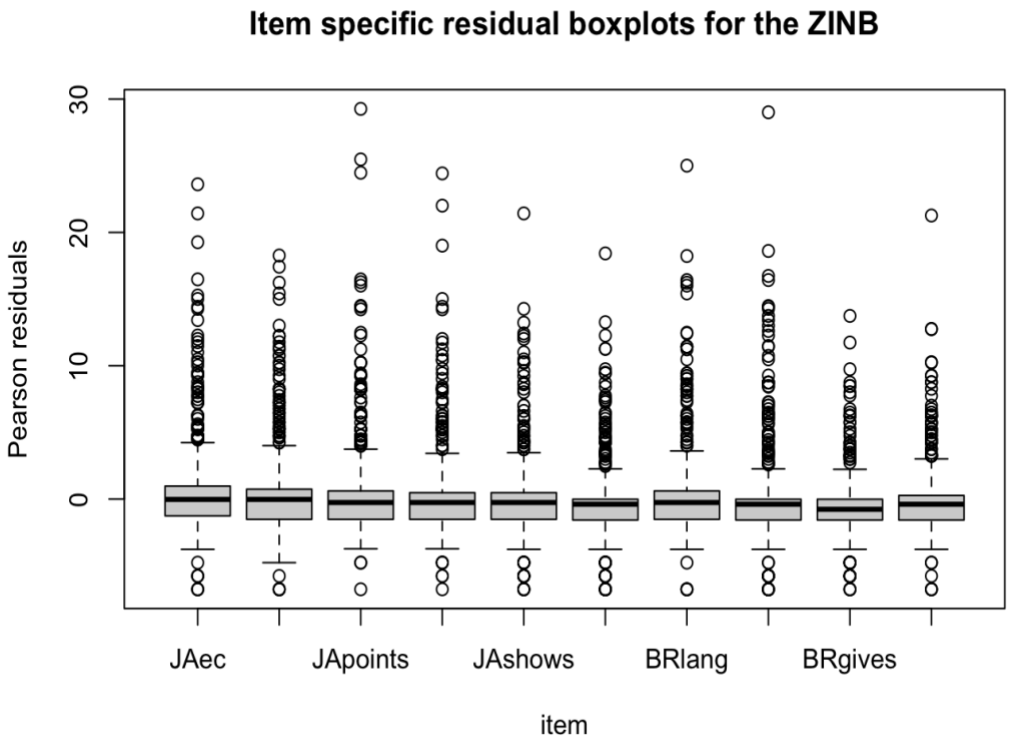


Figure 14. *Plots of the simulated scaled residuals of the ZINBM*

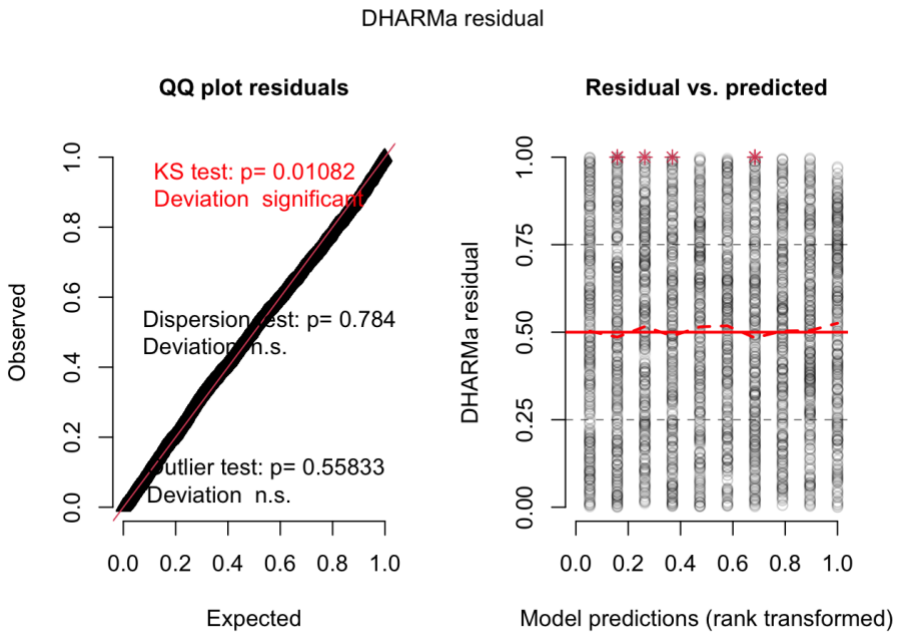


Figure 15. *Simulated dispersion test for the ZINB*

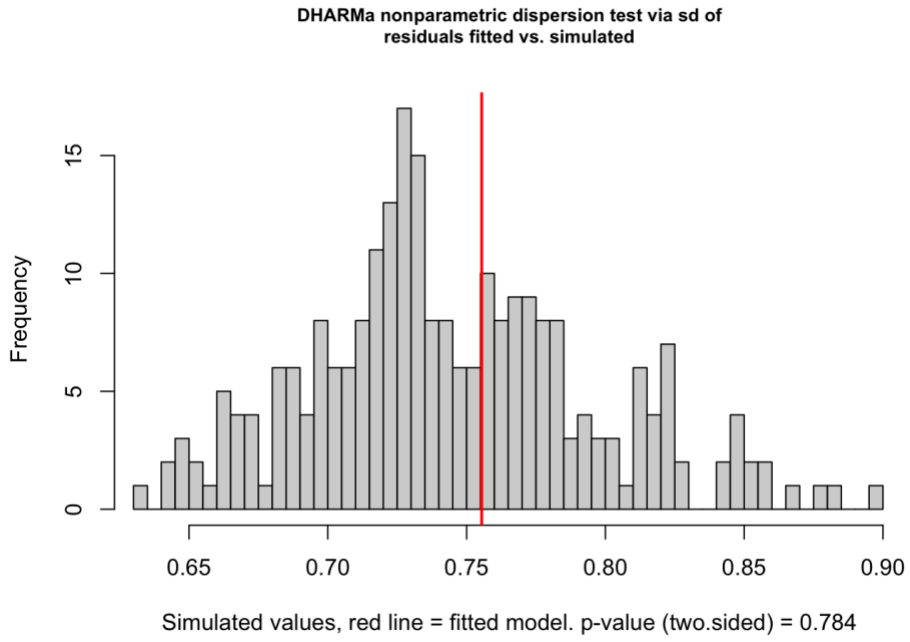


Figure 16. *Simulated zero-inflation test for the ZINB*

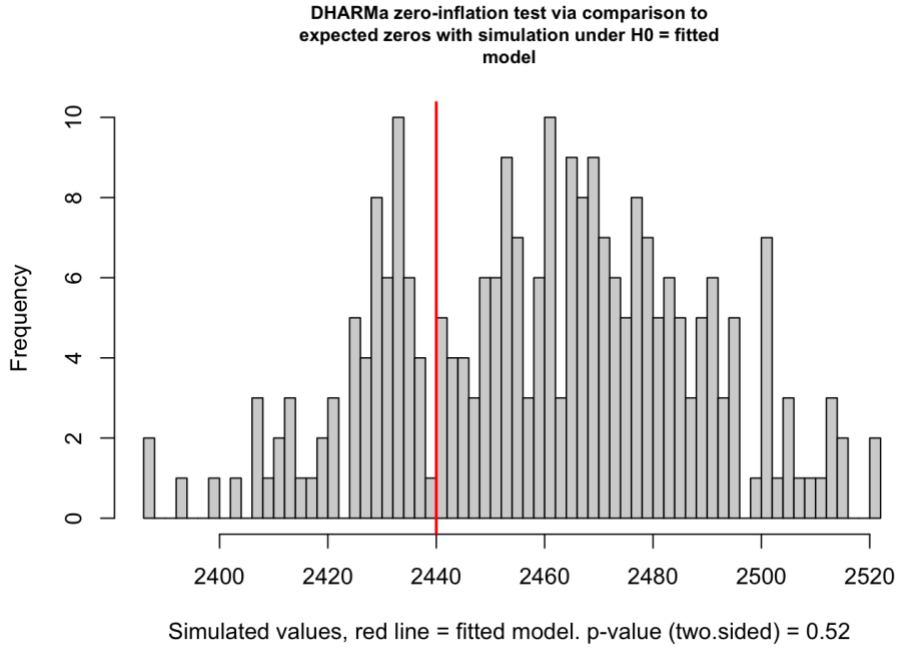


Figure 17. *Item order map of ZINB*

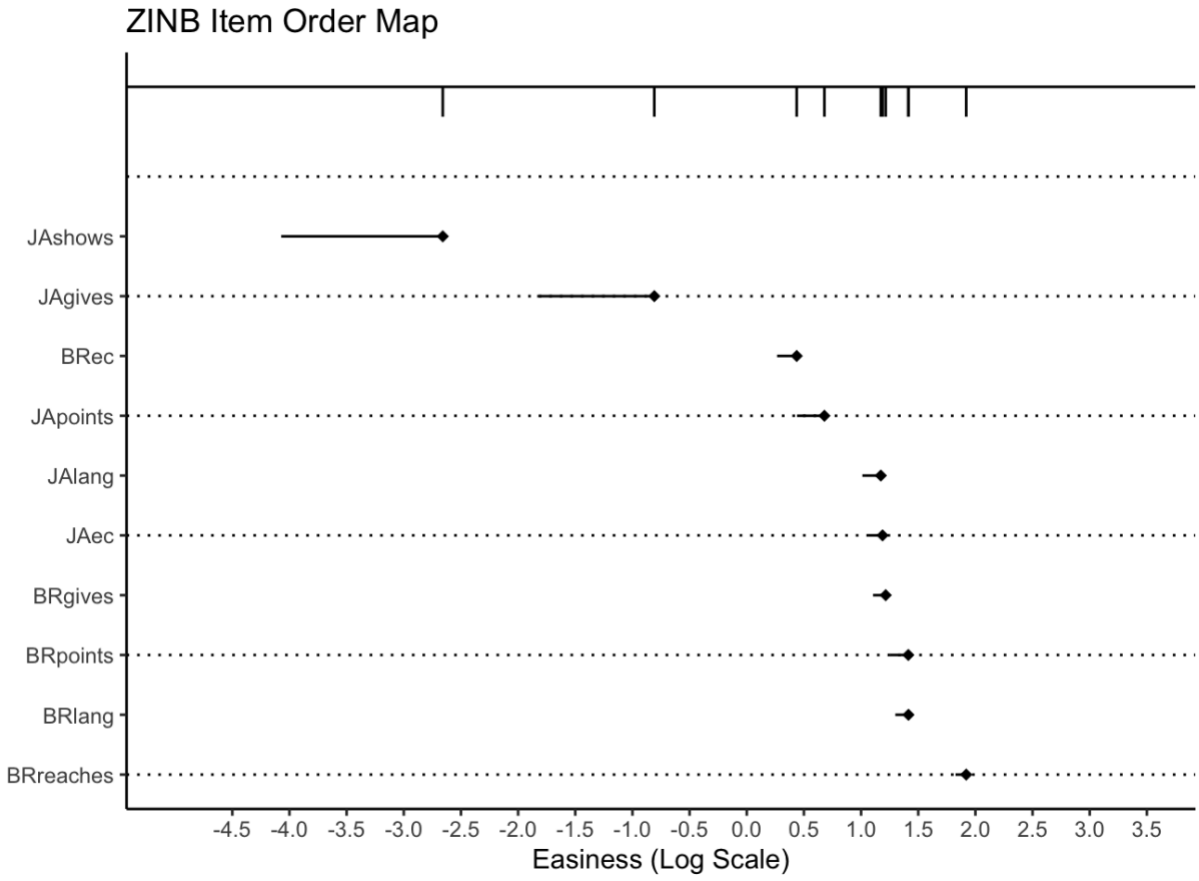


Figure 18. *Distribution of children's social communication ability (theta scores)*

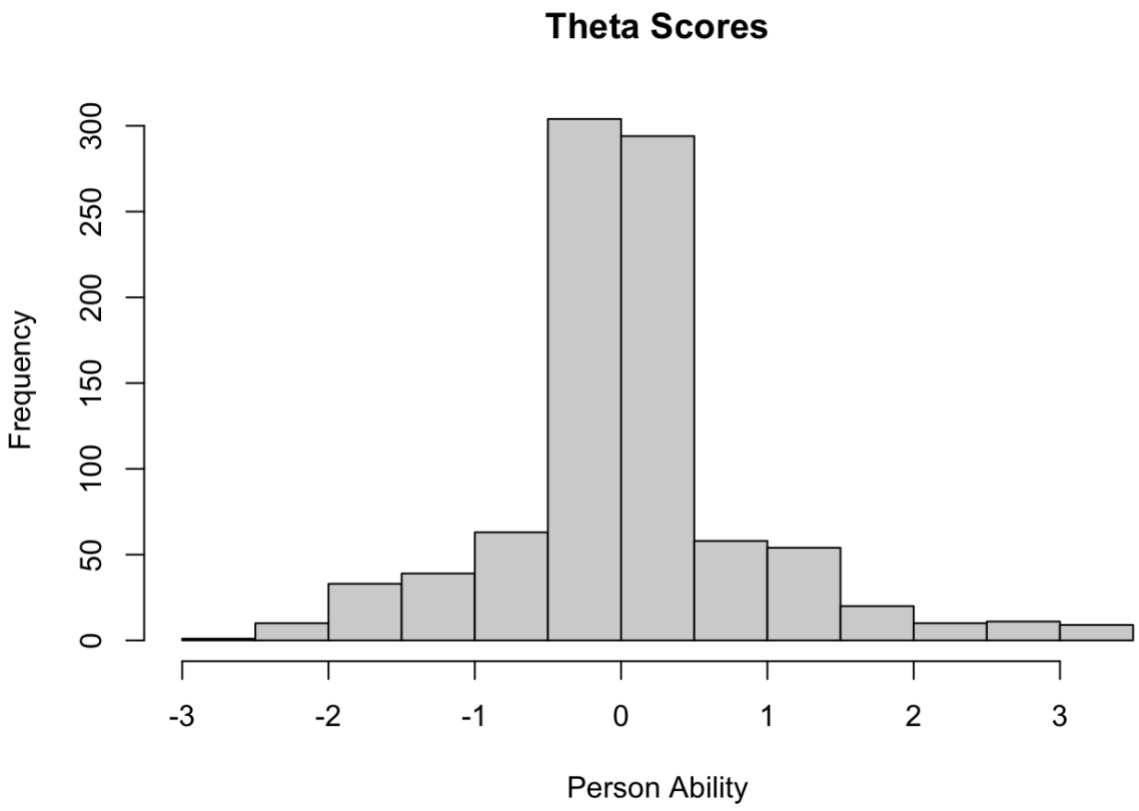
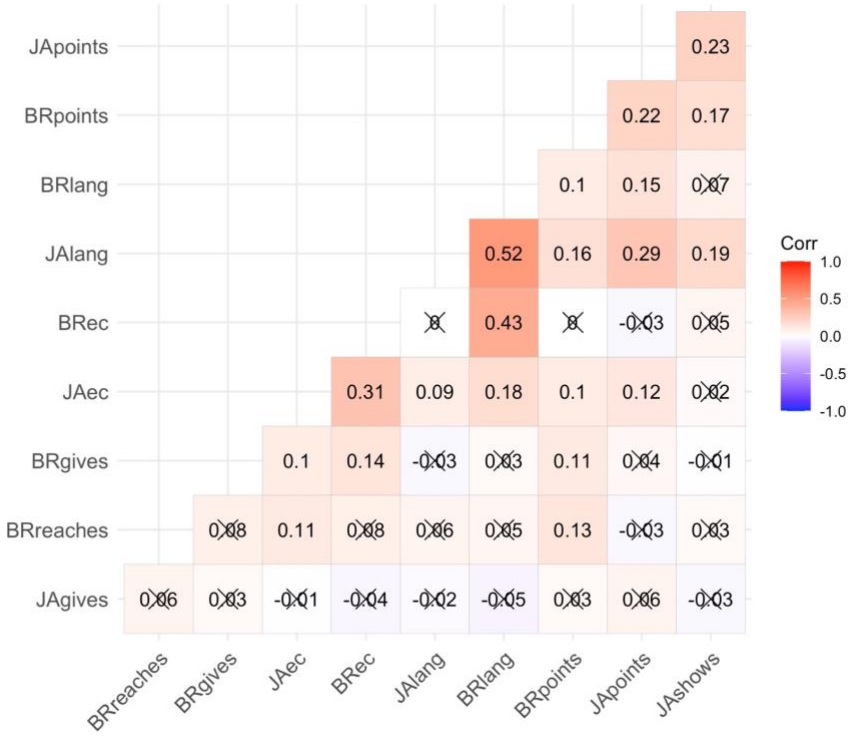
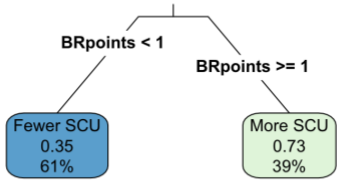


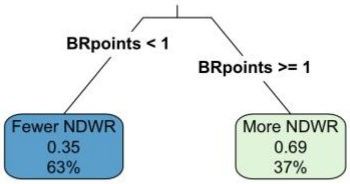
Figure 19. *Correlogram with hierarchical clustering for individual ESCS gestures*

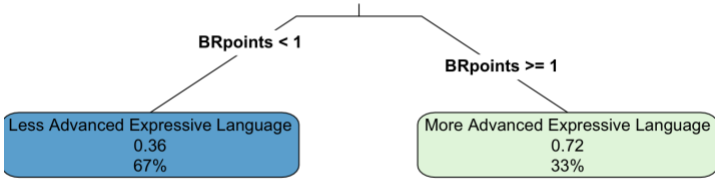


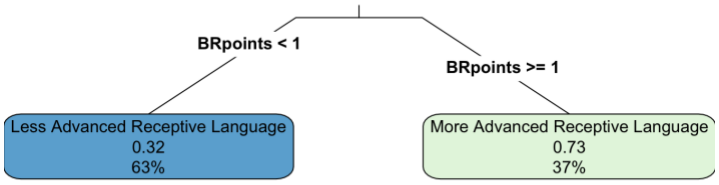
Note: Color gradient from purple to orange represents the size of the Pearson correlation coefficient. Purple is the strongest negative association; orange is the strongest positive association. All correlations are significant at $p = .05$ level except those with an ×.

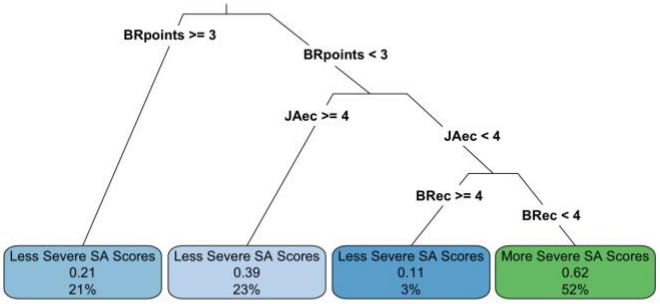
Figure 20. *Classification trees with ESCS individual gestures*








BRpoints < 1
BRpoints >= 1
Less Advanced Receptive Language
0.32
63%
More Advanced Receptive Language
0.73
37%


BRpoints >= 3
BRpoints < 3
JAec >= 4
JAec < 4
BRec >= 4
BRec < 4
Less Severe SA Scores
0.21
21%
Less Severe SA Scores
0.39
23%
Less Severe SA Scores
0.11
3%
More Severe SA Scores
0.62
52%

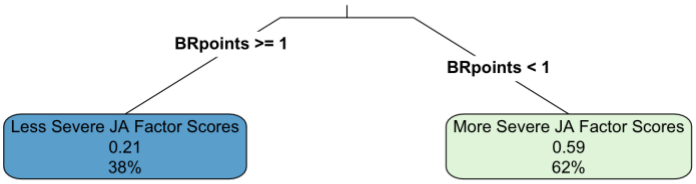
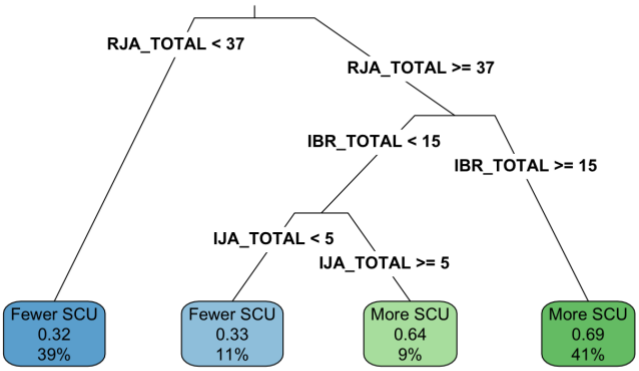
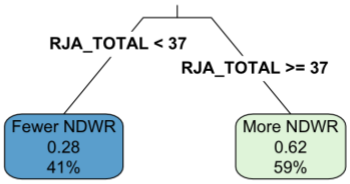
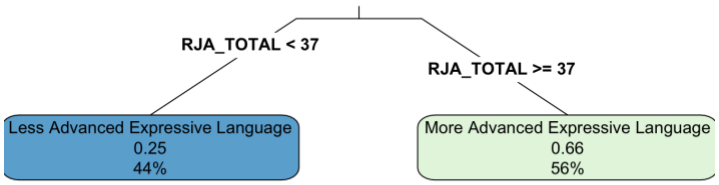
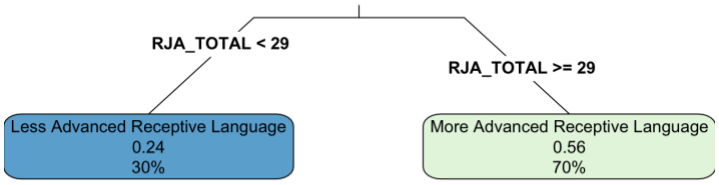

BRpoints >= 1
BRpoints < 1
Less Severe JA Factor Scores
0.21
38%
More Severe JA Factor Scores
0.59
62%

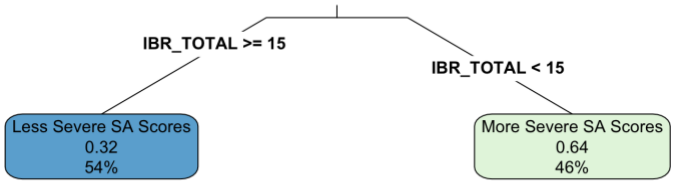
Figure 21. *Classification trees with ESCS domains*








RJA_TOTAL < 29
RJA_TOTAL >= 29
Less Advanced Receptive Language
0.24
30%
More Advanced Receptive Language
0.56
70%


IBR_TOTAL >= 15
IBR_TOTAL < 15
Less Severe SA Scores
0.32
54%
More Severe SA Scores
0.64
46%

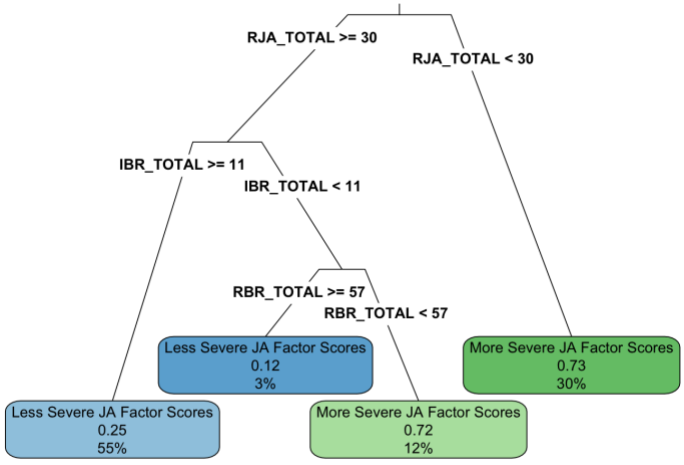

RJA_TOTAL >= 30
RJA_TOTAL < 30
IBR_TOTAL >= 11
IBR_TOTAL < 11
RBR_TOTAL >= 57
RBR_TOTAL < 57
Less Severe JA Factor Scores
0.12
3%
More Severe JA Factor Scores
0.73
30%
Less Severe JA Factor Scores
0.25
55%
More Severe JA Factor Scores
0.72
12%

Figure 22. *Random forest results with ESCS Individiual gesture frequencies*

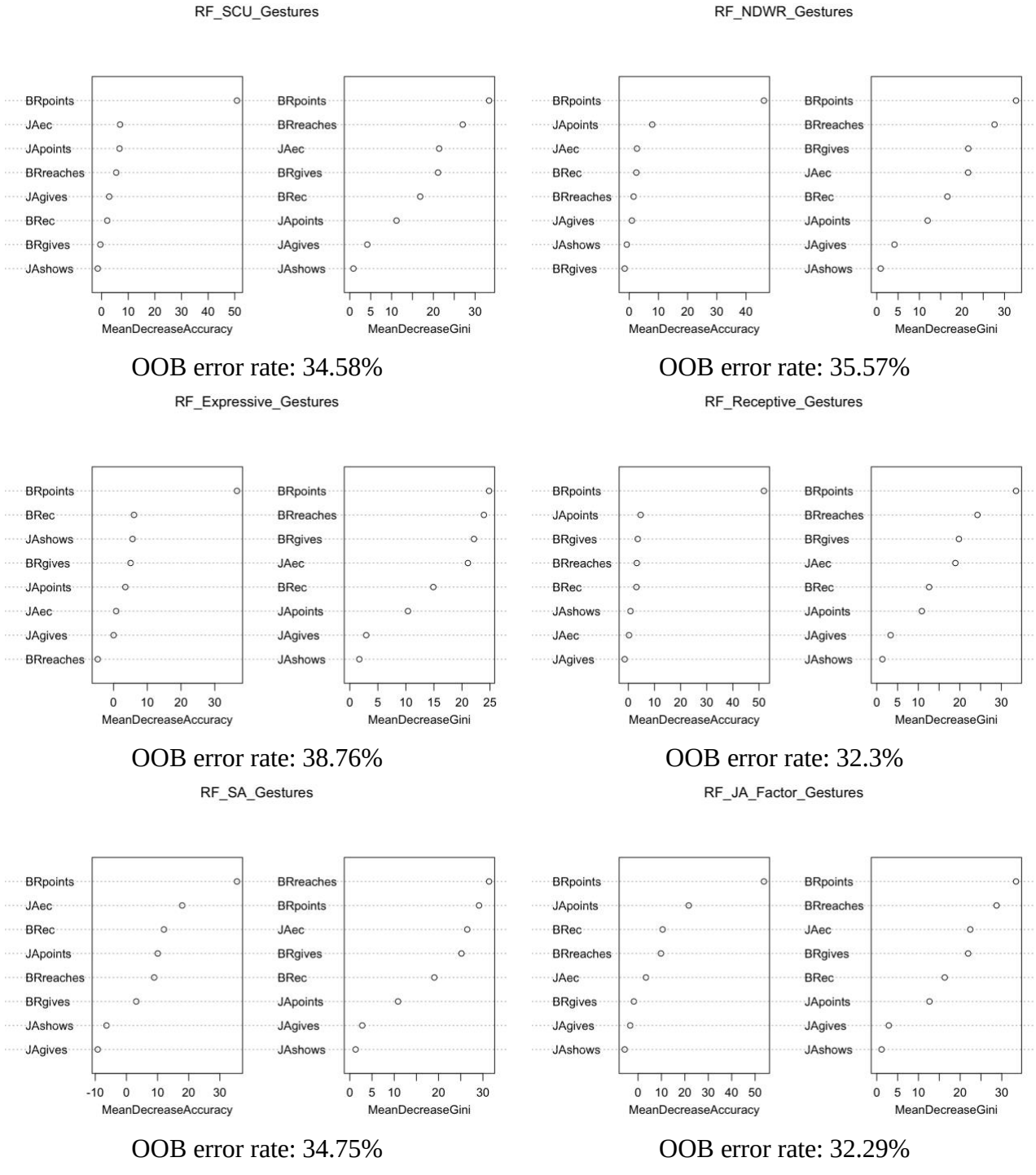
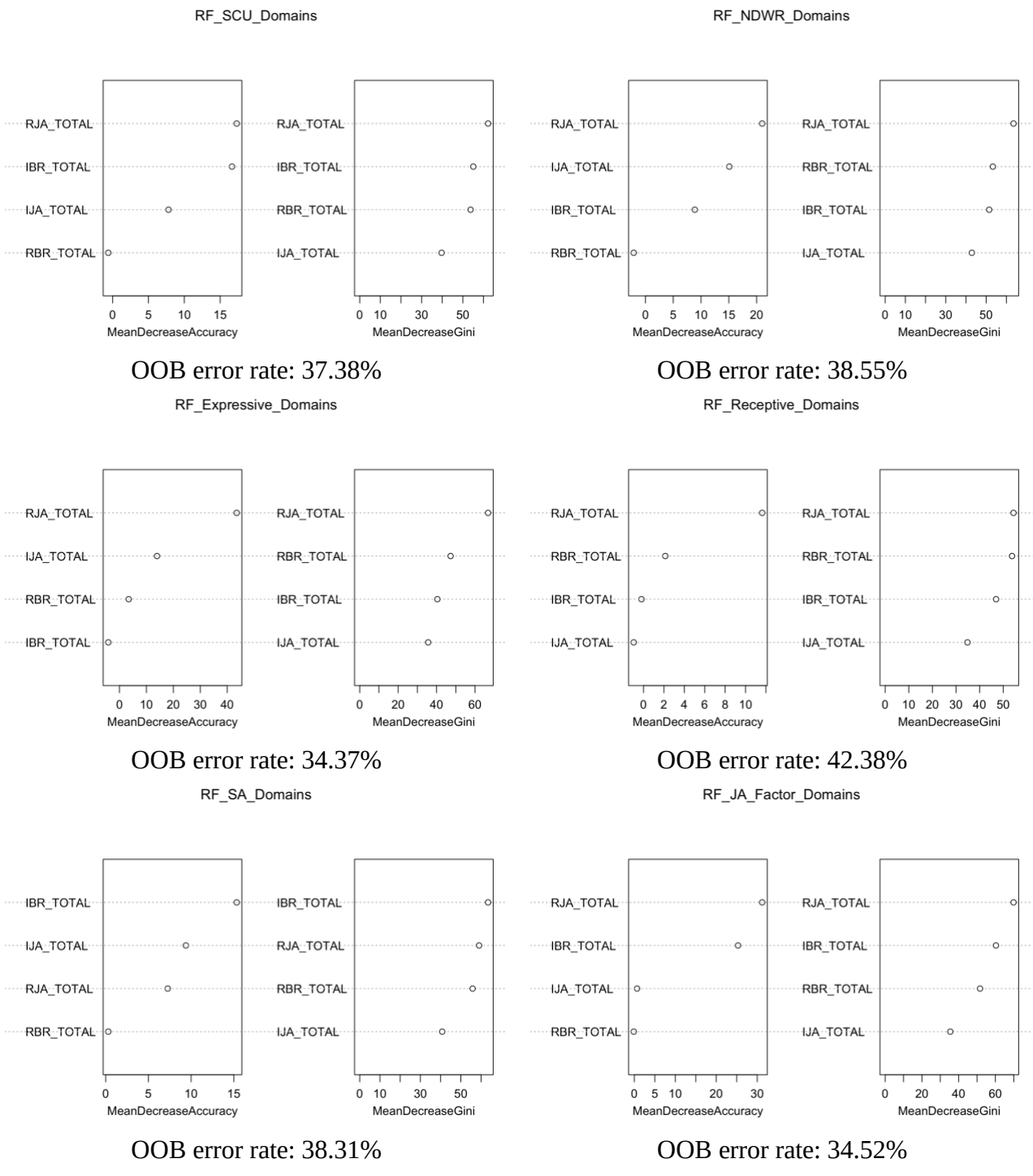


Figure 23. *Random forest results with ESCS domain scores*



References

Abbeduto L., Kover S. T. & McDuffie A. (2011) Studying the language development of children with intellectual disabilities. In: *Research Methods in Child Language* (ed.E. Hoff). Wiley-Blackwell, Hoboken.

Adamson, L. B., Bakeman, R., & Deckner, D. F. (2004). The development of symbol-infused joint engagement. *Child development, 75*(4), 1171–1187. https://doi.org/10.1111/j.1467-8624.2004.00732.x

Aggen, S. H., Neale, M. C., & Kendler, K. S. (2005). DSM criteria for major depression: evaluating symptom patterns using latent-trait item response models. *Psychological medicine, 35*(4), 475-487.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716-723.

Almirall, D., DiStefano, C., Chang, Y. C., Shire, S., Kaiser, A., Lu, X., Nahum-Shani, I., Landa, R., Mathy, P., & Kasari, C. (2016). Longitudinal Effects of Adaptive Interventions With a Speech-Generating Device in Minimally Verbal Children With ASD. *Journal of clinical child and adolescent psychology : the official journal for the Society of Clinical Child and Adolescent Psychology, American Psychological Association, Division 53*, 45(4), 442–456. https://doi.org/10.1080/15374416.2016.1138407

Almirall, D., Nahum-Shani, I., Sherwood, N. E., & Murphy, S. A. (2014). Introduction to SMART designs for the development of adaptive interventions: With application to weight loss research. *Translational Behavioral Medicine, 4*(3), 260–274.

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (5th ed.). Washington: American Psychiatric Association.

Anagnostou, E., Jones, N., Huerta, M., Halladay, A. K., Wang, P., Scahill, L., Horrigan, J. P., Kasari, C., Lord, C., Choi, D., Sullivan, K., & Dawson, G. (2015). Measuring social communication behaviors as a treatment endpoint in individuals with autism spectrum disorder. *Autism*, *19*(5), 622–636. https://doi.org/10.1177/1362361314542955

Anderson, D. K., Lord, C., Risi, S., DiLavore, P. S., Shulman, C., Thurm, A., Welch, K., & Pickles, A. (2007). Patterns of growth in verbal abilities among children with autism spectrum disorder. *Journal of Consulting and Clinical Psychology*, *75*(4), 594–604. https://doi.org/10.1037/0022-006X.75.4.594

Atkins, D. C., Baldwin, S. A., Zheng, C., Gallop, R. J., & Neighbors, C. (2013). A tutorial on count regression and zero-altered count models for longitudinal substance use data. *Psychology of Addictive Behaviors : Journal of the Society of Psychologists in Addictive Behaviors*, *27*(1), 166–177. https://doi.org/10.1037/a0029508

Atkinson, E. J., & Therneau, T. M. (2000). An introduction to recursive partitioning using the RPART routines. *Rochester: Mayo Foundation*.

Baghaei, P., & Doebler, P. (2019). Introduction to the Rasch Poisson Counts Model: An R Tutorial. *Psychological Reports*, *122*(5), 1967–1994. https://doi.org/10.1177/0033294118797577

Bailey, A., Phillips, W., & Rutter, M. (1996). Autism: Towards an Integration of Clinical, Genetic, Neuropsychological, and Neurobiological Perspectives. *Journal of Child Psychology and Psychiatry*, *37*(1), 89–126. https://doi.org/10.1111/j.1469-7610.1996.tb01381.x

Baio, J., Wiggins, L., Christensen, D. L., Maenner, M. J., Daniels, J., Warren, Z., Kurzius-Spencer, M., Zahorodny, W., Rosenberg, C., White, T., Durkin, M.S., Imm, P.,

Nikolaou., L., Yeargin-Allsopp, M., Lee, L., Harrington, R., Lopez, M., Fitzgerald, R.T., Hewitt, A., . . . Dowling, N. F. (2018). Prevalence of autism spectrum disorder among children aged 8 years—Autism and developmental disabilities monitoring network, 11 sites, United States, 2014. *Morbidity and Mortality Weekly Report, 67,* 1–23.

Baker, F. (2001). *The basics of item response theory*. College Park, MD: ERIC.

Bal, V. H., Fok, M., Lord, C., Smith, I. M., Mirenda, P., Szatmari, P., Vaillancourt, T., Volden, J., Waddell, C., Zwaigenbaum, L., Bennett, T., Duku, E., Elsabbagh, M., Georgiades, S., Ungar, W. J., & Zaidman-Zait, A. (2020). Predictors of longer-term development of expressive language in two independent longitudinal cohorts of language-delayed preschoolers with Autism Spectrum Disorder. *Journal of Child Psychology and Psychiatry*, *61*(7), 826–835. https://doi.org/10.1111/jcpp.13117

Bal, V. H., Katz, T., Bishop, S. L., & Krasileva, K. (2016). Understanding definitions of minimally verbal across instruments: evidence for subgroups within minimally verbal children and adolescents with autism spectrum disorder. *Journal of child psychology and psychiatry, and allied disciplines*, *57*(12), 1424–1433. https://doi.org/10.1111/jcpp.12609

Baron-Cohen, S. (1989). Perceptual role taking and protodeclarative pointing in autism. *British Journal of developmental psychology*, *7*(2), 113-127.

Bates, D., Maechler, M., Bolker, B, Walker, S., Christensen, R., Singmann, H., et al. (2017). *lme4: Linear Mixed-Effects Models using 'Eigen' and S4.* R package, version 1.1-14. https://cran.r-project.org/web/packages/lme4/index.html.

Bishop, S., Farmer, C., Kaat, A., Georgiades, S., Kanne, S., & Thurm, A. (2019). The Need for a Developmentally Based Measure of Social Communication Skills. *Journal of the*

American Academy of Child & Adolescent Psychiatry, *58*(6), 555–560. https://doi.org/10.1016/j.jaac.2018.12.010

Bliss, C. I., & Fisher, R. A. (1953). Fitting the negative binomial distribution to biological data. *Biometrics*, *9*(2), 176-200.

Bohning, D., Dietz, E., Schlattmann, P., Mendonca, L., & Kirchner, U. (1999). The Zero-Inflated Poisson Model and the Decayed, Missing and Filled Teeth Index in Dental Epidemiology. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *162*(2), 195–209. http://www.jstor.org/stable/2680578

Bolte, E. E., & Diehl, J. J. (2013). Measurement Tools and Target Symptoms/Skills Used to Assess Treatment Response for Individuals with Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders*, *43*(11), 2491–2501. https://doi.org/10.1007/s10803-013-1798-7

Bono, M. A., Daley, T., & Sigman, M. (2004). Relations among joint attention, amount of intervention and language gain in autism. *Journal of autism and developmental disorders*, *34*(5), 495-505.

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123-140.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Brignell, A., Morgan, A. T., Woolfenden, S., Klopper, F., May, T., Sarkozy, V., & Williams, K. (2018). A systematic review and meta-analysis of the prognosis of language outcomes for individuals with autism spectrum disorder. *Autism & Developmental Language Impairments*, *3*, 239694151876761. https://doi.org/10.1177/2396941518767610

van den Broek J. (1995). A score test for zero inflation in a Poisson distribution. *Biometrics, 51*(2), 738-743.

Carrizosa, E., Molero-Río, C., & Romero Morales, D. (2021). Mathematical optimization in classification and regression trees. *TOP*, *29*(1), 5–33. https://doi.org/10.1007/s11750-021-00594-1

Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., Ader, D., Fries, J. F., Bruce, B., & Rose, M. (2007). The Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care*, *45*(5 Suppl 1), S3–S11. https://doi.org/10.1097/01.mlr.0000258615.42478.55

Chan, K. S., Orlando, M., Ghosh-Dastidar, B., Duan, N., & Sherbourne, C. D. (2004). The interview mode effect on the Center for Epidemiological Studies Depression (CES-D) scale: an item response theory analysis. *Medical care*, 281-289.

Charman, T. (1998). Specifying the nature and course of the joint attention impairment in autism in the preschool years: Implications for diagnosis and intervention. *Autism*, *2*(1), 61-79.

Charman, T. (2003). Why is joint attention a pivotal skill in autism? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *358*(1430), 315–324. https://doi.org/10.1098/rstb.2002.1199

Collins, L. M., Murphy, S. A., & Bierman, K. L. (2004). A conceptual framework for adaptive preventive interventions. *Prevention Science, 5*(3), 185–196.

Collins, L. M., Nahum-Shani, I., & Almirall, D. (2014). Optimization of behavioral dynamic treatment regimens based on the Sequential, Multiple Assignment, Randomized Trial (SMART). *Clinical Trials, 11*(4).

Costello, E. J., & Maughan, B. (2015). Annual Research Review: Optimal outcomes of child and adolescent mental illness. *Journal of Child Psychology and Psychiatry*, *56*(3), 324–341.

Cunningham, A. B. (2012). Measuring change in social interaction skills of young children with autism. *Journal of Autism and Developmental Disorders*, *42*(4), 593-605.

Dawson, G., Rogers, S., Munson, J., Smith, M., Winter, J., Greenson, J., Donaldson, A., & Varley, J. (2010). Randomized, Controlled Trial of an Intervention for Toddlers With Autism: The Early Start Denver Model. *Pediatrics*, *125*(1), e17–e23. https://doi.org/10.1542/peds.2009-0958

Dawson, G., Toth, K., Abbott, R., Osterling, J., Munson, J., Estes, A., & Liaw, J. (2004). Early Social Attention Impairments in Autism: Social Orienting, Joint Attention, and Attention to Distress. *Developmental Psychology*, *40*(2), 271–283. http://dx.doi.org/10.1037/0012-1649.40.2.271

Demidenko, E. (2013). *Mixed models: theory and applications*. Hoboken, NJ: John Wiley & Sons.

DeMyer, M. K., Barton, S., DeMyer, W. E., Norton, J. A., Allen, J., & Steele, R. (1973). Prognosis in autism: A follow-up study. *Journal of Autism and Childhood Schizophrenia*, *3*(3), 199–246. https://doi.org/10.1007/BF01538281

DiStefano, C., & Kasari, C. (2016). The Window to Language is Still Open: Distinguishing Between Preverbal and Minimally Verbal Children With ASD. *Perspectives of the ASHA Special Interest Groups*, *1*(1), 4–11. https://doi.org/10.1044/persp1.SIG1.4

Doebler, A., Doebler, P., & Holling, H. (2014). A Latent Ability Model for Count Data and Application to Processing Speed. Applied Psychological Measurement, 38(8), 587–598. https://doi.org/10.1177/0146621614543513

Doebler, A., & Holling, H. (2016). A processing speed test based on rule-based item generation: An analysis with the Rasch Poisson Counts model. *Learning and Individual Differences*, *52*, 121–128. https://doi.org/10.1016/j.lindif.2015.01.013

Dominick, K. C., Davis, N. O., Lainhart, J., Tager-Flusberg, H., & Folstein, S. (2007). Atypical behaviors in children with autism and children with a history of language impairment. *Research in Developmental Disabilities*, *28*(2), 145–162.

Doostfatemeh, M., Taghi Ayatollah, S. M., & Jafari, P. (2016). Power and Sample Size Calculations in Clinical Trials with Patient-Reported Outcomes under Equal and Unequal Group Sizes Based on Graded Response Model: A Simulation Study. *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research*, *19*(5), 639–647. https://doi.org/10.1016/j.jval.2016.03.1857

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research*, *16*(S1), 5–18. https://doi.org/10.1007/s11136-007-9198-0

Ellis Weismer, S., & Kover, S. T. (2015). Preschool language variation, growth, and predictors in children on the autism spectrum. *Journal of Child Psychology and Psychiatry*, *56*(12), 1327–1337. https://doi.org/10.1111/jcpp.12406

Engel, J. (1984). Models for response data showing extra-Poisson variation. *Statistica Neerlandica*, *38*(3), 159-167.

Embretson, S. E., & Reise, S. P. (2000*). Item Response Theory for Psychologists*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Farewell, V. T., & Sprott, D. A. (1988). The use of a mixture model in the analysis of count data. *Biometrics*, *44*(4), 1191–1194.

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). *Learning from imbalanced data sets* (Vol. 10, pp. 978-3). Berlin: Springer.

Folstein, S., & Rutter, M. (1977). Infantile Autism: A Genetic Study of 21 Twin Pairs. *Journal of Child Psychology and Psychiatry*, *18*(4), 297–321. https://doi.org/10.1111/j.1469-7610.1977.tb00443.x

Fombonne, E. (2009). Epidemiology of Pervasive Developmental Disorders. *Pediatric Research*, *65*(6), 591–598.

Franchini, M., Duku, E., Armstrong, V., Brian, J., Bryson, S. E., Garon, N., Roberts, W., Roncadin, C., Zwaigenbaum, L., & Smith, I. M. (2018). Variability in Verbal and Nonverbal Communication in Infants at Risk for Autism Spectrum Disorder: Predictors and Outcomes. *Journal of Autism and Developmental Disorders*, *48*(10), 3417–3431. https://doi.org/10.1007/s10803-018-3607-9

Fries, J., Bruce, B., & Cella, D. (2005). The promise of PROMIS: Using item response theory to improve assessment of patient-reported outcomes. *Clinical and Experimental Rheumatology*, *23*, S53-7.

Fuller, E. A., & Kaiser, A. P. (2019). The Effects of Early Intervention on Social Communication Outcomes for Children with Autism Spectrum Disorder: A Meta-analysis. *Journal of Autism and Developmental Disorders*. https://doi.org/10.1007/s10803-019-03927-z

Georgiades, S., & Kasari, C. (2018). Reframing Optimal Outcomes in Autism. *JAMA Pediatrics*, *172*(8), 716–717. https://doi.org/10.1001/jamapediatrics.2018.1016

Gersten, R., Fuchs, L. S., Compton, D., Coyne, M., Greenwood, C., & Innocenti, M. S. (2005). Quality indicators for group experimental and quasi-experimental research in special education. *Exceptional children*, *71*(2), 149-164.

Gevarter, C., O'Reilly, M. F., Kuhn, M., Mills, K., Ferguson, R., Watkins, L., Sigafoos, J., Lang, R., Rojeski, L., & Lancioni, G. E. (2016). Increasing the vocalizations of individuals with autism during intervention with a speech-generating device. *Journal of applied behavior analysis*, *49*(1), 17–33. https://doi.org/10.1002/jaba.270

Gotham, K., Risi, S., Pickles, A., & Lord, C. (2007). The Autism Diagnostic Observation Schedule: Revised Algorithms for Improved Diagnostic Validity. *Journal of Autism and Developmental Disorders*, *37*(4), 613–627. https://doi.org/10.1007/s10803-006-0280-1

Green, J., Charman, T., McConachie, H., Aldred, C., Slonims, V., Howlin, P., Le Couteur, A., Leadbitter, K., Hudry, K., Byford, S., Barrett, B., Temple, K., Macdonald, W., Pickles, A., & PACT Consortium (2010). Parent-mediated communication-focused treatment in children with autism (PACT): a randomised controlled trial. *Lancet (London, England)*, *375*(9732), 2152–2160. https://doi.org/10.1016/S0140-6736(10)60587-9

Greene, W. H. (1994). Accounting for excess zeros and sample selection in Poisson and negative binomial regression models.

Grzadzinski, R., Janvier, D., & Kim, S. H. (2020). Recent Developments in Treatment Outcome Measures for Young Children With Autism Spectrum Disorder (ASD). *Seminars in pediatric neurology*, *34*, 100806. https://doi.org/10.1016/j.spen.2020.100806

Guastella, A. J., Gray, K. M., Rinehart, N. J., Alvares, G. A., Tonge, B. J., Hickie, I. B., Keating, C. M., Cacciotti-Saija, C., & Einfeld, S. L. (2015). The effects of a course of intranasal oxytocin on social behaviors in youth diagnosed with autism spectrum disorders: A

randomized controlled trial. *Journal of Child Psychology and Psychiatry*, *56*(4), 444–452. https://doi.org/10.1111/jcpp.12305

Hallmayer, J., Cleveland, S., Torres, A., Phillips, J., Cohen, B., Torigoe, T., Miller, J., Fedele, A., Collins, J., Smith, K., Lotspeich, L., Croen, L. A., Ozonoff, S., Lajonchere, C., Grether, J. K., & Risch, N. (2011). Genetic Heritability and Shared Environmental Factors Among Twin Pairs With Autism. *Archives of General Psychiatry*, *68*(11), 1095–1102. https://doi.org/10.1001/archgenpsychiatry.2011.76

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Vol. 2). Sage.

Hartig, F. (2022). *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models*. R package version 0.4.5. https://CRAN.R-project.org/package=DHARMa

Hartley, S. L., Sikora, D. M., & McCoy, R. (2008). Prevalence and risk factors of maladaptive behaviour in young children with autistic disorder. *Journal of Intellectual Disability Research*, *52*(10), 819–829.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning – Data Mining, Inference, and Prediction*. New York: Springer. doi:10.1007/978-0-387-84858-7_4

Heilbron, D. C. (1994). Zero-altered and other regression models for count data with added zeros. *Biometrical Journal*, *36*(5), 531-547.

Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge University Press.

Hilbe, J. M. (2014). *Modeling count data*. Cambridge University Press.

Holling, H., Böhning, W., Böhning, D., and Formann, A. K. (2013). The covariate adjusted frequency plot. *Stat. Methods Med. Res*. 25, 902–916. doi: 10.1177/0962280212473386

Howlin, P., Gordon, R. K., Pasco, G., Wade, A., & Charman, T. (2007). The effectiveness of Picture Exchange Communication System (PECS) training for teachers of children with autism: A pragmatic, group randomised controlled trial. *Journal of Child Psychology and Psychiatry*, *48*(5), 473–481. https://doi.org/10.1111/j.1469-7610.2006.01707.x

Howlin, P., Mawhood, L., & Rutter, M. (2000). Autism and Developmental Receptive Language Disorder—a Follow-up Comparison in Early Adult Life. II: Social, Behavioural, and Psychiatric Outcomes. *Journal of Child Psychology and Psychiatry*, *41*(5), 561–578. https://doi.org/10.1111/1469-7610.00643

Hung, L.-F. (2012). A Negative Binomial Regression Model for Accuracy Tests. Applied Psychological Measurement, 36(2), 88–103. https://doi.org/10.1177/0146621611429548

Ingersoll B. (2012). Brief report: effect of a focused imitation intervention on social functioning in children with autism. *Journal of autism and developmental disorders*, *42*(8), 1768–1773. https://doi.org/10.1007/s10803-011-1423-6

Interagency Autism Coordinating Committee (IACC). (2011). IACC Strategic plan for autism spectrum disorder research. https://iacc.hhs.giv/strategicplan/2011/index.shtml.

Interagency Autism Coordinating Committee (IACC). (2017). *IACC strategic plan for autism spectrum disorder research: 2016-2017 update*. US Department of Health and Human Services Interagency Autism Coordinating Committee. National Institutes of Health.

Jack, A., & Pelphrey, K. A. (2017). Annual Research Review: Understudied populations within the autism spectrum – current trends and future directions in neuroimaging research.

Journal of Child Psychology and Psychiatry, *58*(4), 411–435. https://doi.org/10.1111/jcpp.12687

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.

Jansen, M. G. H. (1994). Parameters of the latent distribution in Rasch's Poisson counts model. In G. Fischer & D. Laming (Eds.), *Contributions to mathematical psychology, psychometrics, and methodology* (pp. 319-326). New York, NY: Springer.

Jendryczko, D., Berkemeyer, L., & Holling, H. (2020). Introducing a Computerized Figural Memory Test Based on Automatic Item Generation: An Analysis With the Rasch Poisson Counts Model. *Frontiers in Psychology*, *11*, 945. https://doi.org/10.3389/fpsyg.2020.00945

Johnson, C. P., Myers, S. M., & and the Council on Children With Disabilities. (2007). Identification and Evaluation of Children With Autism Spectrum Disorders. *PEDIATRICS*, *120*(5), 1183–1215. https://doi.org/10.1542/peds.2007-2361

Jones, R. M., Carberry, C., Hamo, A., & Lord, C. (2017). Placebo-like response in absence of treatment in children with Autism. *Autism research : official journal of the International Society for Autism Research*, *10*(9), 1567–1572. https://doi.org/10.1002/aur.1798

Kanner, L. (1943). Autistic disturbances of affective contact.

Kasari, C., Brady, N., Lord, C., & Tager-Flusberg, H. (2013). Assessing the Minimally Verbal School-Aged Child With Autism Spectrum Disorder: Assessing minimally verbal ASD. *Autism Research*, *6*(6), 479–493. https://doi.org/10.1002/aur.1334

Kasari, C., Freeman, S., & Paparella, T. (2006). Joint attention and symbolic play in young children with autism: A randomized controlled intervention study. *Journal of Child*

Psychology and Psychiatry, *47*(6), 611–620. https://doi.org/10.1111/j.1469-7610.2005.01567.x

Kasari, C., Gulsrud, A., Freeman, S., Paparella, T., & Hellemann, G. (2012). Longitudinal Follow Up of Children with Autism Receiving Targeted Interventions on Joint Attention and Play RH = Targeted Interventions on Joint Attention and Play. *Journal of the American Academy of Child and Adolescent Psychiatry*, *51*(5), 487–495. https://doi.org/10.1016/j.jaac.2012.02.019

Kasari, C., Gulsrud, A., Paparella, T., Hellemann, G., & Berry, K. (2015). Randomized comparative efficacy study of parent-mediated interventions for toddlers with autism. *Journal of consulting and clinical psychology*, *83*(3), 554.

Kasari, C., Gulsrud, A. C., Wong, C., Kwon, S., & Locke, J. (2010). Randomized controlled caregiver mediated joint engagement intervention for toddlers with autism. *Journal of autism and developmental disorders*, *40*(9), 1045-1056.

Kasari, C., Kaiser, A., Goods, K., Nietfeld, J., Mathy, P., Landa, R., Murphy, S., & Almirall, D. (2014). Communication Interventions for Minimally Verbal Children With Autism: Sequential Multiple Assignment Randomized Trial. *Journal of the American Academy of Child and Adolescent Psychiatry*, *53*(6), 635–646. https://doi.org/10.1016/j.jaac.2014.01.019

Kasari, C., Paparella, T., Freeman, S., & Jahromi, L. B. (2008). Language outcome in autism: Randomized comparison of joint attention and play interventions. *Journal of Consulting and Clinical Psychology*, *76*(1), 125–137. https://doi.org/10.1037/0022-006X.76.1.125

Kasari, C., & Smith, T. (2013). Interventions in schools for children with autism spectrum disorder: Methods and recommendations. *Autism*, *17*(3), 254–267.

Kasari, C., Sturm, A., & Shih, W. (2018). SMARTer Approach to Personalizing Intervention for Children With Autism Spectrum Disorder. *Journal of Speech, Language, and Hearing Research*. https://doi.org/10.1044/2018_JSLHR-L-RSAUT-18-0029

Kean, J., Brodke, D. S., Biber, J., & Gross, P. (2018). An introduction to Item Response Theory and Rasch Analysis of the Eating Assessment Tool (EAT-10). *Brain Impairment : A Multidisciplinary Journal of the Australian Society for the Study of Brain Impairment*, *19*(Spec Iss 1), 91–102. https://doi.org/10.1017/BrImp.2017.31

Kean, J., & Reilly, J. (2014). Classical test theory. *Handbook for clinical research: Design, statistics, and implementation*, 192-194.

Kean, J., & Reilly, J. (2014). Item response theory. *Handbook for clinical research: Design, statistics and implementation*, 195-198.

Koegel L. K. (2000). Interventions to facilitate communication in autism. *Journal of autism and developmental disorders*, *30*(5), 383–391. https://doi.org/10.1023/a:1005539220932

Koegel, L. K., Bryan, K. M., Su, P. L., Vaidya, M., & Camarata, S. (2020). Definitions of Nonverbal and Minimally Verbal in Research for Autism: A Systematic Review of the Literature. *Journal of Autism and Developmental Disorders*, *50*(8), 2957–2972. https://doi.org/10.1007/s10803-020-04402-w

Koegel, R. L., Vernon, T. W., & Koegel, L. K. (2009). Improving social initiations in young children with autism using reinforcers with embedded social interactions. *Journal of autism and developmental disorders*, *39*(9), 1240–1251. https://doi.org/10.1007/s10803-009-0732-5

Konishi, S., & Kitagawa, G. (2008). Information criteria and statistical modeling.

Krstovska-Guerrero, I., & Jones, E. A. (2016). Social-communication intervention for toddlers with autism spectrum disorder: Eye gaze in the context of requesting and joint attention. *Journal of Developmental and Physical Disabilities*, *28*(2), 289-316.

Kuhn, M. (2008). Building Predictive Models in *R* Using the caret Package. *Journal of Statistical Software*, *28*(5). https://doi.org/10.18637/jss.v028.i05

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* (Vol. 26, p. 13). New York: Springer

Lambert, D. (1992). Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. *Technometrics*, *34*(1), 1–14. https://doi.org/10.2307/1269547

Landa, R. J., Holman, K. C., & Garrett-Mayer, E. (2007). Social and Communication Development in Toddlers With Early and Later Diagnosis of Autism Spectrum Disorders. *Archives of General Psychiatry*, *64*(7), 853. https://doi.org/10.1001/archpsyc.64.7.853

Lawless, J. F. (1987). Negative binomial and mixed Poisson regression. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, 209-225.

Lawton, K., & Kasari, C. (2012). Teacher-implemented joint attention intervention: Pilot randomized controlled study for preschoolers with autism. *Journal of Consulting and Clinical Psychology*, *80*(4), 687–693.

Lindsey, J. K. (1999). Models for repeated measurements. *OUP Catalogue*.

Lord C. (2000). Commentary: achievements and future directions for intervention research in communication and autism spectrum disorders. *Journal of autism and developmental disorders*, *30*(5), 393–398. https://doi.org/10.1023/a:1005591205002

Lord, C., & Jones, R. M. (2012). Annual Research Review: Re-thinking the classification of autism spectrum disorders. *Journal of Child Psychology and Psychiatry*, *53*(5), 490–509. https://doi.org/10.1111/j.1469-7610.2012.02547.x

Lord, C., Luyster, R. J., Gotham, K., & Guthrie, W. (2012). *Autism Diagnostic Observation Schedule, second edition (ADOS-2) toddler module.* Los Angeles, CA: Western Psychological Services.

Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Jr, Leventhal, B. L., DiLavore, P. C., Pickles, A., & Rutter, M. (2000). The autism diagnostic observation schedule-generic: a standard measure of social and communication deficits associated with the spectrum of autism. *Journal of autism and developmental disorders*, *30*(3), 205–223.

Lord, C., Shulman, C., & DiLavore, P. (2004). Regression and word loss in autistic spectrum disorders. *Journal of Child Psychology and Psychiatry*, *45*(5), 936–955. https://doi.org/10.1111/j.1469-7610.2004.t01-1-00287.x

Loveland, K. A., & Landry, S. H. (1986). Joint attention and language in autism and developmental language delay. *Journal of Autism and Developmental Disorders*, *16*(3), 335–349. https://doi.org/10.1007/BF01531663

Luby, J., Mrakotsky, C., Stalets, M. M., Belden, A., Heffelfinger, A., Williams, M., & Spitznagel, E. (2006). Risperidone in preschool children with autistic spectrum disorders: an investigation of safety and efficacy. *Journal of child and adolescent psychopharmacology*, *16*(5), 575–587. https://doi.org/10.1089/cap.2006.16.575

Lüdecke, D., Ben-Shachar, M., Patil, I., Waggoner, P., & Makowski, D. (2021). performance: An R Package for Assessment, Comparison and Testing of Statistical Models. *Journal of Open Source Software*, *6*(60), 3139. https://doi.org/10.21105/joss.03139

Luyster, R., & Lord, C. (2009). Word learning in children with autism spectrum disorders. *Developmental psychology*, *45*(6), 1774–1786. https://doi.org/10.1037/a0016223

MacKenzie, D., Nichols, J., Lachman, G., Droege, S., Royle, J., & Langtimm, C. (2002). Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, 83, 2248–2255.

Maenner, M. J., Shaw, K.A., Baio, J.,Washington, A., Patrick, M., DiRienzo, M., Christensen, D.L., Wiggins, L.D., Pettygrove, S., Andrews, J.G., Lopez, M., Hudson, A., Baroud, T., Schwenk, Y., White, T., Rosenberg, C., Lee, L., Harrington, R.A., Huston, M., . . . Dietz, P.M. (2020). Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years—Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2016. *MMWR. Surveillance Summaries*, *69*. https://doi.org/10.15585/mmwr.ss6904a1

Maenner, M. J., Shaw, K. A., Bakian, A. V., Bilder, D. A., Durkin, M. S., Esler, A., Furnier, S. M., Hallas, L., Hall-Lande, J., Hudson, A., Hughes, M. M., Patrick, M., Pierce, K., Poynter, J. N., Salinas, A., Shenouda, J., Vehorn, A., Warren, Z., Constantino, J. N., … Cogswell, M. E. (2021). Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 8 Years—Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2018. *MMWR Surveillance Summaries*, *70*(11), 1–16. https://doi.org/10.15585/mmwr.ss7011a1

Magiati, I., Moss, J., Yates, R., Charman, T., & Howlin, P. (2011). Is the Autism Treatment Evaluation Checklist a useful tool for monitoring progress in children with autism spectrum disorders? *Journal of Intellectual Disability Research*, *55*(3), 302–312. https://doi.org/10.1111/j.1365-2788.2010.01359.x

Masi, A., DeMayo, M. M., Glozier, N., & Guastella, A. J. (2017). An Overview of Autism Spectrum Disorder, Heterogeneity and Treatment Options. *Neuroscience Bulletin*, *33*(2), 183–193. https://doi.org/10.1007/s12264-017-0100-y

Masters, G. N., & Wright, B. D. (1984). The essential process in a family of measurement models. *Psychometrika*, *49*(4), 529–544. https://doi.org/10.1007/BF02302590

Matson, J. L., & Kozlowski, A. M. (2011). The increasing prevalence of autism spectrum disorders. *Research in Autism Spectrum Disorders*, *5*(1), 418–425. https://doi.org/10.1016/j.rasd.2010.06.004

Matson, J. L., Mahan, S., Kozlowski, A. M., & Shoemaker, M. (2010). Developmental milestones in toddlers with autistic disorder, pervasive developmental disorder—Not otherwise specified and atypical development. *Developmental Neurorehabilitation*, *13*(4), 239–247. https://doi.org/10.3109/17518423.2010.481299

McConachie, H., Parr, J. R., Glod, M., Hanratty, J., Livingstone, N., Oono, I. P., Robalino, S., Baird, G., Beresford, B., Charman, T., Garland, D., Green, J., Gringras, P., Jones, G., Law, J., Le Couteur, A. S., Macdonald, G., McColl, E. M., Morris, C., … Williams, K. (2015). Systematic review of tools to measure outcomes for young children with autism spectrum disorder. *Health Technology Assessment*, *19*(41), 1–506. https://doi.org/10.3310/hta19410

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). London, UK: Chapman

& Hall.

Milborrow, S. (2016). Plotting rpart trees with the rpart.plot package. Available at http://www.milbo.org/rpart-plot/prp.pdf.

Miller, J. & Iglesias, A. (2010). Systematic analysis of language transcripts (SALT), Research Version 2010 [computer software]. Middleton, WI: Salt Software, LLC.

Mitchell, S., Brian, J., Zwaigenbaum, L., Roberts, W., Szatmari, P., Smith, I., & Bryson, S. (2006). Early Language and Communication Development of Infants Later Diagnosed with Autism Spectrum Disorder: *Journal of Developmental & Behavioral Pediatrics*, *27*(Supplement 2), S69–S78. https://doi.org/10.1097/00004703-200604002-00004

Mullen, E. (1997). Mullen Scales of Early Learning. Circle Pines, MN: American Guidance Service.

Mundy, P. (1995). Joint attention and social-emotional approach behavior in children with autism. *Development and Psychopathology*, *7*(1), 63-82. doi:10.1017/S0954579400006349

Mundy, P., (2016). Autism and joint attention: *Development, neuroscience, and clinical fundamentals.* New York, NY: Guilford Publications.

Mundy P. (2018). A review of joint attention and social-cognitive brain systems in typical development and autism spectrum disorder. *The European journal of neuroscience*, *47*(6), 497–514. https://doi.org/10.1111/ejn.13720

Mundy, P., Block, J., Delgado, C., Pomares, Y., Vaughan Van Hecke, A., & Parlade, M. V. (2007). Individual Differences and the Development of Joint Attention in Infancy. *Child Development*, *78*(3), 938–954.

Mundy, P., Delgado, C., Block, J., Venezia, M., Hogan, A., & Seibert, J. (2003). *A manual for the abridged Early Social Communication Scales (ESCS)*. University of Miami.

Mundy, P., Kasari, C., Sigman, M., & Ruskin, E. (1995). Nonverbal Communication and Early Language Acquisition in Children With Down Syndrome and in Normally Developing Children. *Journal of Speech, Language, and Hearing Research*, *38*(1), 157–167.

Mundy, P., Sigman, M., & Kasari, C. (1990). A longitudinal study of joint attention and language development in autistic children. *Journal of autism and developmental disorders*, *20*(1), 115–128. https://doi.org/10.1007/BF02206861

Mundy, P., Sigman, M., & Kasari, C. (1994). Joint attention, developmental level, and symptom presentation in autism. *Development and Psychopathology*, *6*(3), 389–401. https://doi.org/10.1017/S0954579400006003

Mundy, P., Sigman, M., Kasari, C., & Yirmiya, N. (1988). Nonverbal Communication Skills in Down Syndrome Children. *Child Development*, *59*(1), 235–249. JSTOR.

Mundy, P., Sigman, M., Ungerer, J., & Sherman, T. (1986). Defining the Social Deficits of Autism: The Contribution of Non-Verbal Communication Measures. *Journal of Child Psychology and Psychiatry*, *27*(5), 657–669. https://doi.org/10.1111/j.1469-7610.1986.tb00190.x

Mundy, P., Sigman, M., Ungerer, J., & Sherman, T. (1987). Nonverbal communication and play correlates of language development in autistic children. *Journal of Autism and Developmental Disorders*, *17*(3), 349–364.

Mundy, P., & Volkmar, F. R. (2013). RJA/IJA (Initiating/responding to joint attention). *Encyclopedia of autism spectrum disorders*, 2609-2616.

Munson, J., Dawson, G., Sterling, L., Beauchaine, T., Zhou, A., Elizabeth, K., Lord, C., Rogers, S., Sigman, M., Estes, A., & Abbott, R. (2008). Evidence for latent classes of IQ in

young children with autism spectrum disorder. *American journal of mental retardation : AJMR*, *113*(6), 439–452. https://doi.org/10.1352/2008.113:439-452

Murray, D. S., Creaghead, N. A., Manning-Courtney, P., Shear, P. K., Bean, J., & Prendeville, J. A. (2008). The relationship between joint attention and language in children with autism spectrum disorders. *Focus on autism and other developmental disabilities*, *23*(1), 5-14.

Murza, Schwartz, J. B., Hahs-Vaughn, D. L., & Nye, C. (2016). Joint attention interventions for children with autism spectrum disorder: a systematic review and meta-analysis. *International Journal of Language & Communication Disorders*, *51*(3), 236–251. https://doi.org/10.1111/1460-6984.12212

National Research Council, Committee on Educational Interventions for Children with Autism. (2001). Educating Children with Autism. National Academics Press; https://www.nap.edu/catal og/10017.html.

National Institute on Deafness and Other Communication Disorders (NIDCD). (2010). NIH workshop on nonverbal school-aged children with autism. Bethesda, MD.

Nordahl-Hansen, A., Fletcher-Watson, S., McConachie, H., & Kaale, A. (2016). Relations between specific and global outcome measures in a social-communication intervention for children with autism spectrum disorder. *Research in Autism Spectrum Disorders*, *29*, 19-29.

Nordin, V., & Gillberg, C. (1998). The long-term course of autistic disorders: update on follow-up studies. *Acta psychiatrica Scandinavica*, *97*(2), 99–108. https://doi.org/10.1111/j.1600-0447.1998.tb09970.x

Norrelgen, F., Fernell, E., Eriksson, M., Hedvall, Å., Persson, C., Sjölin, M., Gillberg, C., & Kjellmer, L. (2015). Children with autism spectrum disorders who do not develop phrase

speech in the preschool years. *Autism*, *19*(8), 934–943. https://doi.org/10.1177/1362361314556782

Oosterling, I., Roos, S., de Bildt, A., Rommelse, N., de Jonge, M., Visser, J., Lappenschaar, M., Swinkels, S., van der Gaag, R. J., & Buitelaar, J. (2010). Improved Diagnostic Validity of the ADOS Revised Algorithms: A Replication Study in an Independent Sample. *Journal of Autism and Developmental Disorders*, *40*(6), 689–703. https://doi.org/10.1007/s10803-009-0915-0

Ozonoff, S., Iosif, A.-M., Baguio, F., Cook, I. C., Hill, M. M., Hutman, T., Rogers, S. J., Rozga, A., Sangha, S., Sigman, M., Steinfeld, M. B., & Young, G. S. (2010). A Prospective Study of the Emergence of Early Behavioral Signs of Autism. *Journal of the American Academy of Child and Adolescent Psychiatry*, *49*(3), 256-66.e1-2.

Ozonoff, S., Hueng, K., & Thompson, M. (2011). Regression and other patterns of onset. In D. G. Amaral, G. Dawson, & D. Geschwind (Eds.), *Autism spectrum disorders*. New York: Oxford University Press.

Paparella, T., Goods, K. S., Freeman, S., & Kasari, C. (2011). The emergence of nonverbal joint attention and requesting skills in young children with autism. *Journal of Communication Disorders*, *44*(6), 569–583. https://doi.org/10.1016/j.jcomdis.2011.08.002

Pickard, K. E., & Ingersoll, B. R. (2015). Brief report: High and low level initiations of joint attention, and response to joint attention: Differential relationships with language and imitation. *Journal of autism and developmental disorders*, *45*(1), 262-268.

Pickett, E., Pullara, O., O'Grady, J., & Gordon, B. (2009). Speech acquisition in older nonverbal individuals with autism: a review of features, methods, and prognosis. *Cognitive and*

behavioral neurology : official journal of the Society for Behavioral and Cognitive Neurology, *22*(1), 1–21. https://doi.org/10.1097/WNN.0b013e318190d185

Pickles, A., Anderson, D. K., & Lord, C. (2014). Heterogeneity and plasticity in the development of language: A 17-year follow-up of children referred early for possible autism. *Journal of Child Psychology and Psychiatry*, *55*(12), 1354–1362. https://doi.org/10.1111/jcpp.12269

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rapin, I., Dunn, M. A., Allen, D. A., Stevens, M. C., & Fein, D. (2009). Subtypes of language disorders in school-age children with autism. *Developmental neuropsychology*, *34*(1), 66–84. https://doi.org/10.1080/87565640802564648

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests* (Expanded Ed.). Chicago: University of Chicago Press.

Remington, B., Hastings, R. P., Kovshoff, H., degli Espinosa, F., Jahr, E., Brown, T., Alsford, P., Lemaic, M., & Ward, N. (2007). Early Intensive Behavioral Intervention: Outcomes for Children With Autism and Their Parents After Two Years. *American Journal on Mental Retardation*, *112*(6), 418.

Reynell, J. K. (1977). Reynell Developmental Language Scales (Revised). Windsor, England: NFER Publishing.

Rimland, B. (1978). Savant capabilities of autistic children and their cognitive implications.

Rizopoulos, D. (2022). *GLMMadaptive: Generalized Linear Mixed Models using Adaptive Gaussian Quadrature*. R package version 0.8-5. https://CRAN.R-project.org/package=GLMMadaptive

Romski, M., Sevcik, R. A., Adamson, L. B., Cheslock, M., Smith, A., Barker, R. M., & Bakeman, R. (2010). Randomized comparison of augmented and nonaugmented language interventions for toddlers with developmental delays and their parents. *Journal of speech, language, and hearing research : JSLHR*, *53*(2), 350–364. https://doi.org/10.1044/1092-4388(2009/08-0156)

Rose, V., Trembath, D., Keen, D., & Paynter, J. (2016). The proportion of minimally verbal children with autism spectrum disorder in a community-based early intervention programme. *Journal of Intellectual Disability Research*, *60*(5), 464–477. https://doi.org/10.1111/jir.12284

Rutter M. (1978). Diagnosis and definition of childhood autism. *Journal of autism and childhood schizophrenia*, *8*(2), 139–161. https://doi.org/10.1007/BF01537863

Rutter, M., Greenfeld, D., & Lockyer, L. (1967). A five to fifteen year follow-up study of infantile psychosis. II. Social and behavioural outcome. *The British journal of psychiatry : the journal of mental science*, *113*(504), 1183–1199. https://doi.org/10.1192/bjp.113.504.1183

Russell, G., Mandy, W., Elliott, D., White, R., Pittwood, T., & Ford, T. (2019). Selection bias on intellectual ability in autism research: A cross-sectional review and meta-analysis. *Molecular Autism*, *10*(1), 9. https://doi.org/10.1186/s13229-019-0260-x

Sallows, G. O., & Graupner, T. D. (2005). Intensive behavioral treatment for children with autism: four-year outcome and predictors. *American journal of mental retardation : AJMR*, *110*(6), 417–438.

Sandbank, M., Bottema-Beutel, K., Crowley, S., Cassidy, M., Dunham, K., Feldman, J. I., Crank, J., Albarran, S. A., Raj, S., Mahbub, P., & Woynaroski, T. G. (2020). Project

AIM: Autism intervention meta-analysis for studies of young children. *Psychological Bulletin*, *146*(1), 1–29. https://doi.org/10.1037/bul0000215

Sandri, M., & Zuccolotto, P. (2008). A bias correction algorithm for the Gini variable importance measure in classification trees. *Journal of Computational and Graphical Statistics*, *17*(3), 611-628.

Schietecatte, I., Roeyers, H., & Warreyn, P. (2012). Exploring the nature of joint attention impairments in young children with autism spectrum disorder: Associated social and cognitive skills. *Journal of Autism and Developmental Disorders*, *42*(1), 1-12.

Schreibman, L., Dawson, G., Stahmer, A. C., Landa, R., Rogers, S. J., McGee, G. G., Kasari, C., Ingersoll, B., Kaiser, A. P., Bruinsma, Y., McNerney, E., Wetherby, A., & Halladay, A. (2015). Naturalistic Developmental Behavioral Interventions: Empirically Validated Treatments for Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders*, *45*(8), 2411–2428. https://doi.org/10.1007/s10803-015-2407-8

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461–464.

Shih, W., Patterson, S. Y., & Kasari, C. (2016). Developing an Adaptive Treatment Strategy for Peer-Related Social Skills for Children With Autism Spectrum Disorders. *Journal of Clinical Child & Adolescent Psychology*, *45*(4), 469–479. https://doi.org/10.1080/15374416.2014.915549

Shih, W., Shire, S., Chang, Y. C., & Kasari, C. (2021). Joint engagement is a potential mechanism leading to increased initiations of joint attention and downstream effects on language: JASPER early intervention for children with ASD. Journal of Child Psychology and Psychiatry. https://doi.org/10.1111/jcpp.13405

Shumway, S., & Wetherby, A. M. (2009). Communicative Acts of Children with Autism Spectrum Disorders in the Second Year of Life. *Journal of Speech, Language, and Hearing Research : JSLHR*, *52*(5), 1139–1156. https://doi.org/10.1044/1092-4388(2009/07-0280)

Sigman, M., & McGovern, C. W. (2005). Improvement in cognitive and language skills from preschool to adolescence in autism. *Journal of autism and developmental disorders*, *35*(1), 15-23.

Sigman, M., Ruskin, E., Arbeile, S., Corona, R., Dissanayake, C., Espinosa, M., Kim, N., López, A., & Zierhut, C. (1999). Continuity and change in the social competence of children with autism, Down syndrome, and developmental delays. *Monographs of the Society for Research in Child Development*, *64*(1), 1–114. https://doi.org/10.1111/1540-5834.00002

Siller, M., & Sigman, M. (2008). Modeling longitudinal change in the language abilities of children with autism: Parent behaviors and child characteristics as predictors of change. *Developmental Psychology*, *44*(6), 1691–1704. https://doi.org/10.1037/a0013771

Steidtmann, D., Manber, R., Blasey, C., Markowitz, J. C., Klein, D. N., Rothbaum, B. O., Thase, M. E., Kocsis, J. H., & Arnow, B. A. (2013). Detecting Critical Decision Points in Psychotherapy and Psychotherapy + Medication for Chronic Depression. *Journal of Consulting and Clinical Psychology*, *81*(5), 783–792.

Sparrow, S. S., Cicchetti, D. V., Balla, D. A., & Doll, E. A. (2005). *Vineland Adaptive Behavior Scales: Survey forms manual*. American Guidance Service.

Tager-Flusberg, H. (1999). The challenge of studying language development in children with autism. In L. Menn & N. B. Ratner (Eds.), *Methods for studying language production*. Mahwah, NJ: Erlbaum.

Tager-Flusberg, H., & Kasari, C. (2013). Minimally Verbal School-Aged Children with Autism Spectrum Disorder: The Neglected End of the Spectrum. *Autism Research*, *6*(6), 468–478. https://doi.org/10.1002/aur.1329

Tager-Flusberg, H., Paul, P., & Lord, C. (2005). Language and communication in autism. In F. R. Volkmar, A. Klin, R. Paul, & D. J. Cohen (Eds.), *Handbook of autism and pervasive developmental disorders* (3rd ed., pp. 335–364). Hoboken, NJ: Wiley

Tager-Flusberg, H., Rogers, S., Cooper, J., Landa, R., Lord, C., Paul, R., Rice, M., Stoel-Gammon, C., Wetherby, A., & Yoder, P. (2009). Defining Spoken Language Benchmarks and Selecting Measures of Expressive Language Development for Young Children With Autism Spectrum Disorders. *Journal of Speech, Language, and Hearing Research*, *52*(3), 643–652. https://doi.org/10.1044/1092-4388(2009/08-0136)

Thall, P. F., & Vail, S. C. (1990). Some Covariance Models for Longitudinal Count Data with Overdispersion. *Biometrics*, *46*(3), 657–671. https://doi.org/10.2307/2532086

Thurm, A., Manwaring, S. S., Swineford, L., & Farmer, C. (2015). Longitudinal study of symptom severity and language in minimally verbal children with autism. *Journal of child psychology and psychiatry, and allied disciplines*, *56*(1), 97–104. https://doi.org/10.1111/jcpp.12285

Tomasello, M., & Farrar, M. J. (1986). Joint Attention and Early Language. *Child Development*, *57*(6), 1454–1463. JSTOR. https://doi.org/10.2307/1130423

Vivanti, G., Prior, M., Williams, K., & Dissanayake, C. (2014). Predictors of outcomes in autism early intervention: why don't we know more?. *Frontiers in pediatrics*, *2*, 58.

Vuong, Q. H. (1989). Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica*, *57*(2), 307–333. https://doi.org/10.2307/1912557

Washington, S., Karlaftis, M., Mannering, F., & Anastasopoulos, P. (2020). *Statistical and econometric methods for transportation data analysis*. Chapman and Hall/CRC.

Wetherby, A.M. (2006). Understanding and measuring social communication in children with autism spectrum disorders. In: Charman T and Stone W (eds) *Social and Communication Development in Autism Spectrum Disorders: Early Identification, Diagnosis and Intervention.* New York: Guilford Press.

Wetherby, A.M., & Prizant, B. (1998). *Communication and Symbolic Behavior Scales Developmental Profile (Research Edition)*. Chicago, IL: Applied Symbolix. s

Wetherby, A.M., & Prizant, B. (2002). *Communication and Symbolic Behavior Scales Developmental Profile (First Normed Edition)*. Baltimore, MD: Paul H. Brooks.

Wodka, E. L., Mathy, P., & Kalb, L. (2013). Predictors of Phrase and Fluent Speech in Children With Autism and Severe Language Delay. *PEDIATRICS*, *131*(4), e1128–e1134. https://doi.org/10.1542/peds.2012-2221

Wright, B. D. (1977). Misunderstanding the Rasch model. *Journal of Educational Measurement*, 219-225.

Yoder, P. J., & Lieberman, R. G. (2010). Brief Report: Randomized Test of the Efficacy of Picture Exchange Communication System on Highly Generalized Picture Exchanges in Children with ASD. *Journal of Autism and Developmental Disorders*, *40*(5), 629–632.

Yoder, P., & Stone, W. L. (2006). A randomized comparison of the effect of two prelinguistic communication interventions on the acquisition of spoken communication in preschoolers with ASD. *Journal of speech, language, and hearing research : JSLHR*, *49*(4), 698–711. https://doi.org/10.1044/1092-4388(2006/051)

Yoder, P., Watson, L. R., & Lambert, W. (2015). Value-added predictors of expressive and receptive language growth in initially nonverbal preschoolers with autism spectrum disorders. *Journal of autism and developmental disorders*, *45*(5), 1254-1270.

Zhu, W., & Safrit, M. J. (1993). The Calibration of a Sit-ups Task Using the Rasch Poisson Counts Model. *Canadian Journal of Applied Physiology*, *18*(2), 207–219. https://doi.org/10.1139/h93-017