8-2020

Validation of a Brief Prosody Rating Scale for Children with Autism Spectrum Disorder

Sarai S. Holbrook
Utah State University

VALIDATION OF A BRIEF PROSODY RATING SCALE FOR CHILDREN WITH AUTISM SPECTRUM DISORDER

by

Sarai S. Holbrook

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

DOCTOR OF PHILOSOPHY

in

Disability Disciplines

Approved:

_____	_____
Sandra Gillam, Ph.D. Major Professor	Teresa Ukrainetz, Ph.D. Committee Member
_____	_____
Ron Gillam, Ph.D. Committee Member	Sarah Schwartz, Ph.D. Committee Member
_____	_____
Stephanie Borrie, Ph.D. Committee Member	Cindy Jones, Ph.D. Committee Member

_____

Richard S. Inouye, Ph.D.
Vice Provost for Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

2020

ABSTRACT

Validation of a Brief Prosody Rating Scale for Children with Autism Spectrum Disorder

by

Sarai S. Holbrook, Doctor of Philosophy

Utah State University, 2020

Major Professor: Dr. Sandra Gillam
Department: Special Education and Rehabilitation

Speech prosody differences have been noted in persons with autism spectrum disorder (ASD) since its presentation as a clinical entity. Yet, despite many studies, no universal characterization of speech prosody in ASD has been identified. Evidence can be found for speech prosody in ASD being under modulated, not different from typical, or overly variable. For those persons whose speech prosody is atypical and interferes with daily functioning, valid, reliable, and efficient assessments of speech prosody are needed. Currently, there are only three validated assessments for speech prosody specific to ASD and none of them are simultaneously valid, reliable, and efficient.

The purpose of this study was to design, validate, and establish sufficient reliability of a one-item, 7-point continuous analogue rating scale for screening the speech prosody of children with ASD. Additionally, I investigated whether a brief, online training would improve reliability. The rating scale ranged from 1-7 with anchors at 1 (monotonous), 4 (typical), and 7 (overly modulated). Thirty-five 30-second audio clips

were chosen from archival databases of children with ASD and neurotypical development who participated in a previous narrative intervention study and the normative samples of both versions of the Test of Narrative Language. Three expert speech-language pathologists (SLPs) selected clips for the end and mid points of the scale and developed "gold standard" ratings. Three of these were used in a short, online training. A total of 42 ASHA-certified SLPs with experience in treating children with ASD rated 20 of the audio clips at two time points. Twenty of the SLPs participated the online training prior to rating.

Analyses were conducted using linear mixed-effects modeling to account for lack of independence between audio clips and participants. Models were built using a research-question, theory-based modeling approach. Results indicated moderate levels of inter- and intra-rater reliability, with the exception of intra-rater reliability for the trained group, which was good (ICC = 0.76). The results also partially supported the validity of the scale in its content, its relations to other variables, raters' response processes, and the consequences of testing. Currently, this prosody rating scale requires further validation before wide use.

(133 pages)

PUBLIC ABSTRACT

Validation of a Brief Prosody Rating Scale for Children with Autism Spectrum Disorder

Sarai S. Holbrook

Differences in the speech prosody, or "melody" of speech, of persons with autism spectrum disorder (ASD) have long been noted by researchers. Yet, despite many studies, researchers have not identified a universal description of speech prosody in ASD. It may be flat or monotonous, not different from typical, or overly variable. However, atypical speech prosody can immediately set someone apart from their peers. This distinction could negatively social, academic, and vocational interactions. For those persons with ASD whose speech prosody is different from typical and interferes with daily functioning, valid, reliable, and efficient assessments of speech prosody are needed. Currently, there are only three validated assessments for speech prosody specific to ASD and none of them are simultaneously valid, reliable, and efficient.

The purpose of this study was to design, validate, and establish sufficient reliability of a one-item, 7-point continuous rating scale for screening the speech prosody of children with ASD. Additionally, I investigated whether a brief, online training would improve reliability. The rating scale ranged from 1 (monotonous) to 7 (overly variable). Thirty-five 30-second audio clips from previous studies were chosen from children with ASD and neurotypical development. Three expert speech-language pathologists (SLPs) selected clips for the end and mid points of the scale and developed "gold standard"

ratings. A total of 42 ASHA-certified SLPs with experience in treating children with ASD rated 20 of the audio clips at two time points. Twenty of the SLPs participated the online training prior to rating.

Analyses were conducted using linear mixed-effects modeling, which were built using a research-question, theory-based modeling approach. Results indicated moderate levels of reliability, except for intra-rater reliability in the trained group, which was good (ICC = 0.76). The results also partially supported the validity of the scale; however, this prosody rating scale requires further study and development before wide use.

ACKNOWLEDGMENTS

I would like to thank Dr. Sandi Gillam for her unwavering support of my doctoral work and personal development, my unique circumstances notwithstanding. My appreciation also goes to Dr. Ron Gillam for his tutelage and for making available the audio files from the normative database of the TNL. I would not have had a study without them. To Dr. Sarah Schwartz, Dr. Tyson Barrett, and Jeremy Haynes I give my gratitude for guiding me through the forests of multilevel modeling, REDCap, and R. I thank Dr. Stephanie Borrie for helping me wrangle the acoustic aspects of this study. I thank Drs. Teresa Ukrainetz and Cindy Jones for their feedback, advice, and optimism throughout.

To my lab mates I give my gratitude. Thanks go for the laughs, the cries, the theoretical wanderings, the comics, the late night/early morning WhatsApps, the tips and tricks, the peer group. You kept me sane.

Special thanks go to my parents, siblings, husband, daughter and God. I would not be where I am without you. Especially, I thank the Grantsville bunch for being proud of me and helping with Clara. I thank my mom for the phone calls and dropping everything to come when I needed help – and CW and Jay for baching it when she did. I thank Tab for encouraging me, especially at the very end.

To Clara, you helped me do things I didn't think I could, motivated me, and cheered me up. You loved me no matter what. To Camron, thanks for being the string to my kite forever. I couldn't fly without you. To my Heavenly Father and Jesus my thanks go for allowing me the option to quit, but then giving me the vision and strength not to.

Sarai S. Holbrook

CONTENTS

Page

ABSTRACT iii

PUBLIC ABSTRACT v

ACKNOWLEDGMENTS vii

LIST OF TABLES ix

LIST OF FIGURES x

CHAPTER

1. INTRODUCTION 1
Speech Prosody in Autism Spectrum Disorder 1
Acoustic vs. Perceptual Measurements of Speech Prosody 5
Reliability and Validity 6
Perceptual Tools to Assess Prosody in ASD 12
Effects of Training on Ratings 17
Purpose 19
Research Questions 19

2. METHOD 25
Participants 25
Materials 31
Procedure 39
Analysis Plan 41
Validity of the Prosody Rating Scale 46

3. RESULTS 53
Reliability 53
Validity 54

4. DISCUSSION 81
Overview 81
Reliability 82
Content Validity and Response Process Validity 83
Relations to Other Variables 85
Consequences of Testing: Scale Representativeness 88
Limitations and Future Directions 89
Conclusion 94

REFERENCES 95

CURRICULUM VITAE 112

LIST OF TABLES

Table Page

1 Characteristics of Children in the Audio Clips 27

2 SLP Demographics and ASD experience According to Group 29

3 Best Fit MLM Predicting Prosody Ratings.....69

4 Best Fit MLM Predicting Social Acceptability Ratings 72

5 Best Fit MLM Predicting ASD Ratings.....75

LIST OF FIGURES

Figure		Page
1	Screenshot of the Prosody Rating Scale	33
2	Screenshot of the Social Acceptability Scale	35
3	Screenshot of Rating Scale of Presence/Absence of ASD	36
4	Screenshot of Online Training	38
5	Visualization of Data Collection Structure	43
6	MLM Structure for Discriminant Validity	35
7	MLM Structure for Convergent Validity	49
8	MLM Structure for Social Acceptability	50
9	MLM Structure for Ratings of Presence/Absence of ASD	51
10	MLM Structure of Best Fit Prosody Model	52
11	Histogram of SLP Survey Completion Time	58
12	Correlation Plot of Prosody Ratings with Other Variables	60
13	Correlation Plot of Prosody Ratings with F0 Measures	63
14	Scatter Plot of Mean F0 and Prosody Ratings	64
15	Bar Plot of Prosody Ratings Faceted by ASD Diagnosis	65
16	Histogram of Prosody Ratings Faceted by ASD Diagnosis	66
17	Violin Plot of Raw Interaction Between SLP Group, ASD Rating, and ASD Diagnosis.	67
18	Model Plot of Interaction Between ASD Rating and ASD Diagnosis	70
19	Social Acceptability Cut Point Plot	73

20 Moderating Effect of Social Acceptability Ratings on Prosody Ratings 74

21 ASD Rating Cut Point Plot 76

22 Observed Data Violin Plots of Prosody Ratings According to Race 77

23 Observed Data Violin Plots of Prosody Ratings According to Sex 78

24 Moderating Effect of Caseload ASD Percent on Prosody Ratings 80

CHAPTER 1

INTRODUCTION

Speech Prosody in Autism Spectrum Disorder

Atypical speech prosody often sets speakers with autism spectrum disorder (ASD) apart from peers who have neurotypical development (NTD; Filipe et al., 2014; Fusaroli et al., 2017; Nakai et al., 2017). Speech prosody is defined as a group of speech characteristics that exist above the level of words, phrases, and sentences in connected speech (Filipe et al., 2014; Fusaroli et al., 2017; Nakai et al., 2017). These characteristics may include rate, pitch or intonation, stress, pauses, intensity, and duration (Shriberg et al., 1992; Stevens et al., 1983; Szczepek Reed, 2011). While any one of these characteristics are related to the holistic conceptualization of speech prosody, the concept of speech prosody is not adequately captured by one or two characteristics, but rather the synthesis of all of them in emerging discourse (Szczepek Reed, 2011; Tseng et al., 2005). Atypical prosody can adversely affect interactional partners' perceptions of and reactions to persons with ASD and thus can have far-reaching effects academically, socially, and vocationally (Gordon et al., 2019; Peppé, 2009; Schölderle et al., 2016; Szczepek Reed, 2010; Wiklund, 2016; Wynn et al., 2018).

From the earliest conceptualizations of ASD, unusual speech prosody was included in descriptions of the disorder (Asperger, 1944, 1991; Kanner, 1943). However, accounts were not specific nor consistent about the nature of atypical speech prosody. Kanner included both parents' and therapists' observations about prosody in his 1943 report of 11 case studies. For example, descriptions included "he…uttered inarticulate sounds in a monotonous singsong manner" (p. 232), "her voice is peculiarly

unmodulated" (p. 241), and "she speaks well on almost any subject, though with something of an odd intonation" (p. 241). Asperger (1944, 1991) also reported on the prosody of cases he observed. Depending on the case, he described speech prosody as being monotonous as well as "sing-song" (p. 42). He specifically noted the variability of prosody across individuals:

> If one listens carefully, one can invariably pick up these kinds of abnormalities in the language of autistic individuals…. The abnormalities differ, of course, from case to case. Sometimes the voice is soft and far away, sometimes it sounds refined and nasal, but sometimes it is too shrill and ear-splitting. In yet other cases, the voice drones on in a sing-song and does not even go down at the end of a sentence. Sometimes speech is over-modulated and sounds like exaggerated verse-speaking. However, many possibilities there are, they all have one thing in common: the language feels unnatural (Asperger, 1944, 1991, p. 70).

As can be seen, from the very first descriptions of ASD, the speech prosody of persons with ASD did not follow a consistent pattern of being either under-modulated or overly variable, but both, sometimes in the same individual. Current research continues to reflect this apparent incongruity (Fusaroli et al., 2017; Kissine & Geelhand, 2019 - see discussion in particular). Despite continued efforts to characterize atypical speech prosody in persons with ASD as either overly variable or overly monotonous, there is still little consensus about the exact nature of prosody in ASD (Fusaroli et al., 2017; Redford et al., 2018).

The strongest research examining speech prosody in persons with ASD to date is

a systematic literature review and series of meta-analyses by Fusaroli and his colleagues (2017). This review included 30 studies which compared prosodic characteristics of persons with ASD and persons with NTD (univariate studies) and 15 machine-learning studies (multivariate studies). A total of 3,114 participants were included in the literature review and meta-analyses, 966 of whom had ASD. The univariate studies included analyses on various prosodic characteristics including mean pitch, pitch variability, pitch and severity of clinical features of ASD, mean intensity, intensity variability, duration, and vocal quality. The purpose of the machine learning studies was to detect differences in given data sets, then determine group membership based on patterns in the data. Meta-analyses were run on any prosody feature that had at least five studies with statistical estimates of the results to produce Cohen's *d* effect sizes. There were not enough data from the machine-learning studies to combine them into overall estimates of sensitivity/specificity.

The multivariate studies summarized in Fusaroli et al.'s review consistently differentiated between persons with ASD and NTD, however, they were not able to pinpoint a prosodic characteristic that was primarily responsible for distinguishing between the two groups. The univariate studies suggested that persons with ASD might have had elevated mean pitch F0 and a wider standard deviation of mean F0 than persons with NTD. Standard deviation of F0 is an acoustic correlate closely related to a perceived "singsong" prosody pattern noted by both Asperger and Kanner. No other individual prosody characteristics showed consistent differences between persons with ASD and those with NTD. However, Fusaroli and his colleagues strongly emphasized that their evidence was not robust enough to firmly conclude that high, variable mean F0 was a

universal marker of ASD. This is partly because the evidence from the articles included in the review was quite variable in the reported speech prosody characteristics of persons with ASD. There was evidence in the individual studies that speech prosody in persons with ASD was not different from those with NTD, that it was more modulated, and that it was less modulated. Research conducted since the Fusaroli et al. review and meta-analyses has been consistent with these patterns of inconsistency. Some research suggests speech prosody patterns may be less variable in some persons with ASD (Kissine & Geelhand, 2019; Nakai et al., 2017). Other studies have found no difference in prosody between persons with ASD and persons with typical development (Dahlgren et al., 2018).

The fact that research findings conflict over what characterizes speech prosody in ASD in general seems to fit with Asperger's original observation that, "abnormalities differ… from case to case" (Asperger 1944, 1991, p. 70). It is possible that after 75 years, we have fallen into the trap of "[letting] each child's unique personality vanish behind the type" (Asperger 1944, 1991, p. 70) when we seek to define speech prosody in ASD as having any universal presentation. Rather, it seems more productive to screen each individual with ASD to see if a) speech prosody is atypical for that individual and b) the individual's speech prosody patterns adversely affect their interpersonal interactions (Filipe et al., 2014; Grossman et al., 2013; Paul, Augustyn, et al., 2005). To effectively assess speech prosody patterns of persons with ASD to determine if prosody is an area for intervention, we need brief, yet valid and reliable screening tools of speech prosody for persons with ASD that reflect the range of patterns possible in this population (Diehl & Paul, 2009; Peppé, McCann, Gibbon, O'Hare, & Rutherford, 2006). The aim of this study was to develop and validate such an assessment tool.

Acoustic vs. Perceptual Measurements of Speech Prosody

Over the years, speech prosody has been measured both acoustically and perceptually. The main difference between acoustic and perceptual correlates of prosody is method of measurement. Acoustic analysis has been defined as, "measurement of the quality of phonation in terms of frequency, intensity, and time" (Nicolosi et al., 2004, p. 3). Acoustic correlates of prosody include fundamental frequency, intensity, duration, and rate. These correlates are objectively measured using a variety of instruments including acoustic analysis programs and software, sound-level meters, mobile applications, and stop watches (Edgerton & Wine, 2017; Simmons et al., 2016; Szczepek Reed, 2011). Perceptual analysis refers to "the description of sound after it has been received and interpreted in terms of pitch, loudness, and quality" (Nicolosi et al., 2004, p. 231-232). Perceptual correlates of prosody are subjectively rated by the people who perceive the signal (Szczepek Reed, 2011). These correlates of prosody may include pitch, intonation, emphasis, and phrasing.

As can be seen by the definitions of acoustic and perceptual analysis, these two methods of measuring speech prosody are related, but not identical. For instance, fundamental frequency (F0) is an acoustic measurement that refers to the "number of complete opening and closing cycles of the vocal folds per second" (Szczepek Reed, 2011, p. 25). It is measured in Hertz (Hz). The perceptual correlate of F0 is pitch and is characterized in terms of its relation to other parts of the discourse or in comparison to other persons (Nicolosi et al., 2004; Szczepek Reed, 2011). For example, if a man's typical F0 is around 120 Hz and he suddenly begins speaking at around 220 Hz, his pitch will be perceived as being "high"; however this same designation will not be given to a

woman who habitually speaks at 220 Hz, even though the number of opening and closing cycles of the vocal folds are the same between the two people. There are perceptual correlates for many other acoustic measures of prosody. Intensity is heard as loudness. Duration is heard as length. Rate is heard as speed or quickness of speech. While the best measures of perceptual and acoustic prosodic deficits in children with ASD is a matter of debate in the literature, both have been used reliably (Adams, 1992; Azizi et al., 2016; Bone et al., 2015; Filipe et al., 2014; Kargas et al., 2016; McCann & Peppé, 2003; Nadig & Shaw, 2012).

Reliability and Validity

Reliability and validity in measurement are of primary concern in behavioral testing, which includes the assessment of speech prosody. Together, the American Educational Research Association (AERA), American Psychological Association (APA), and the National Council on Measurement in Education (NCME) published updated guidelines in 2014 detailing standards that measurement instruments needed to meet for adequate reliability and validity.

Reliability. Regarding reliability, the 2014 AERA, APA, and NCME guidelines outlined one overall standard with 19 supporting standards grouped into eight "thematic clusters" (American Educational Research Association et al., 2014, p. 42). The overarching reliability standard reads, "Appropriate evidence of reliability/precision should be provided for the interpretation for each intended score use" (p. 42). Overall, measurement instruments should take into account a variety of possible sources of error that might affect the reliability or precision of the instrument across contexts and raters. In the present study, I addressed reliability both between (interrater reliability) and within

raters (intra-rater reliability) then examined the effects of a brief training on the inter- and intra-rater reliability coefficients.

Validity. The main validity standard published by AERA, APA, and NCME (American Educational Research Association et al., 2014) emphasized the need to validate measures for their intended purpose. If there are multiple purposes, the validity of the measure needs to be established for each purpose. The standards outline five areas of evidence that are needed to support the validity of a given measure. These areas are evidence of test content, response processes, internal structure, relations to other variables, and consequences of testing.

Test content evidence of validity. Evidence for a test's validity based on its content investigates the relationship between the content of the test and the theoretical construct it is designed to measure (American Educational Research Association et al., 2014). Validity evidence for content is evaluated using a combination of expert opinion, evidence from empirical studies, and logical analyses of the relationship between the test content and the theoretical construct being analyzed (American Educational Research Association et al., 2014; Furr, 2018). Validity evidence for the content of the current prosody rating scale was based on definitions of speech prosody proposed by previous authors (Szczepek Reed, 2011; 't Hart et al., 1990; Tench, 1996; Wichmann, 2000), and on descriptions of speech prosody in ASD in published literature (Dahlgren et al., 2018; Fusaroli et al., 2017; Kissine & Geelhand, 2019; Nakai et al., 2014, 2017; Paul, Augustyn, et al., 2005). Additionally, speech-language pathologists (SLPs) with extensive experience in working with persons with ASD gave their input into the construct of the scale and suggested improvements, which I incorporated.

Response processes evidence of validity. It is important to demonstrate that the cognitive processes participants engage in while completing a test or scale match those that the test developer(s) intended (American Educational Research Association et al., 2014; Furr, 2018). For instance, a multiple-choice test might be designed to assess content knowledge at the end of an instructional unit. The intended cognitive process would be for a test taker to think about what information is being requested and to retrieve the appropriate answers from memory. However, a person taking the test might be responding in such a way as to make a visual pattern on the scantron sheet. When gathering evidence for the validity of a test, it is important to determine what cognitive processes were *actually* employed and how well those processes match the intended process. Evidence that the cognitive processes used by participants match the processes expected by the researcher can be directly evaluated by asking participants to describe their thought processes when completing items or indirectly by evaluating associated behaviors such as response times, monitoring eye movements, or by asking judges to assess the processes the participants engaged in (American Educational Research Association et al., 2014; Furr, 2018).

I gathered evidence of validity based on response processes in the present study by asking raters to briefly describe how and why they arrived at the ratings they did. Additionally, I extracted the length of time it took each participant to complete the ratings and included it as a variable in relevant statistical models. If the participants took an extremely long time to complete the task (e.g. longer than 3 days for one survey), the validity of their response process might have been called into question. This could be especially true if they completed the training, because it is possible that they could have

forgotten the instructions and training provided, their consistency of responding may have drifted, and their engagement with the task could have decreased.

Internal structure evidence of validity. Evidence of validity based on the internal structure of a test refers to the idea that the items of proposed dimensions of a test should actually reflect the dimensions they are designed to represent. In other words, a test's dimensionality should reflect the structure of the underlying theory (or theories) upon which it's built (Furr, 2018). The internal structure evidence of validity for a measure is assessed through factor analyses. These procedures verify that items that should be correlated in theory are indeed correlated with each other. They are also used to see if test takers from different groups respond differently to certain items in certain dimensions (American Educational Research Association et al., 2014). For the current scale, given that there was only one item on the scale and that item was designed to measure prosody, the internal structure evidence of validity for the test overlapped with the content evidence of validity. While it may seem unnecessary to validate an assessment with only one item, Furr (2011) argued that it might be more necessary to investigate and establish the psychometric properties of single-item scales because they are more prone to poor reliability within and between raters.

Relations to other variables evidence of validity. Evidence of validity based on a measure's relations to other variables falls into four broad categories: discriminant, convergent, concurrent, and predictive evidence (American Educational Research Association et al., 2014; Furr, 2018). Discriminant evidence of validity accumulates when relations are weak between the test being validated and tests that evaluate different constructs. Convergent evidence of validity is obtained from examining relations between

the test being evaluated and other tests which measure the same construct. Concurrent evidence of validity is gathered when pertinent measures are assessed at the same time as the test being investigated. Predictive evidence of validity is assessed by examining the relations between the measure in question and other measures that were given at a later point in time.

In the present study, I obtained concurrent discriminant evidence of validity by evaluating the relations between prosody ratings and the narrative proficiency of the children who provided the audio clips. Narrative proficiency was evaluated using either the first or second version of the Test of Narrative Language (TNL; Gillam & Pearson, 2004, 2017) for all of these children. While the children exhibited speech prosody while telling the stories required by the test, it is unlikely that their prosody would have significantly influenced their narrative proficiency scores because the scoring of the TNL is based primarily on inclusion of macro- and micro-structural elements of narrative, not on prosodic elements (Gillam & Pearson, 2004, 2017).

I gathered concurrent convergent evidence of validity by examining the associations between the ratings on the prosody rating scale and acoustic measures of F0, ratings of social acceptability, and the ratings of presence/absence of ASD. Since prosody and F0 and F0 standard deviation are closely related but are not theoretically identical to the overall construct of speech prosody (Redford et al. 2018), I examined the relations between ratings on the prosody rating scale and acoustic measures of F0 and F0 standard deviation. Because atypical prosody has so long been associated with the recognition of ASD (Asperger, 1944, 1991; Filipe et al., 2014; Kanner, 1943)**,** I evaluated the relations between the ratings on the prosody rating scale and binary ratings of presence/absence of

ASD. Similarly, atypical prosody has previously been associated with reduced ratings of likability (Redford et al., 2018). Establishing predictive evidence of validity was beyond the scope of this study.

Consequences of testing validity. The consequences of testing, both intended and unintended, are considered part of the evidence for a test's validity. Test developers are responsible for validating the consequences of all the intended interpretations of their measures (American Educational Research Association et al., 2014). When consequences are unintentional, but stem from errors in construct underrepresentation or the inclusion of construct-irrelevant content, this must be evaluated and rectified by test developers. To the extent possible, consequences of testing should be anticipated and prepared for in test development (American Educational Research Association et al., 2014).

My intent was that this prosody rating scale would be used by SLPs to screen for prosody atypicalities as part of a larger language assessment battery. The intended interpretation of this scale extended only to that of screening. Therefore, the intended consequences of testing using this measure were binary: if prosody was determined to be atypical, follow-up assessments of prosody would be conducted; if prosody was determined to be typical, assessment and intervention resources could be allocated to other areas. The prosody rating scale was not intended to be interpreted as a complete assessment of prosody nor as a diagnostic measure for ASD. It was designed to be an addendum to an existing testing battery in a full speech and language assessment conducted by qualified SLPs. Nevertheless, it is also possible that demographic factors of SLP raters, and the children in the audio clips may have interacted in ways that could have been meaningful for the consequences of testing; therefore, I investigated the

relations of demographic factors with prosody rating scale scores.

In his chapter on diagnostic systems for ASD, Gillberg (2011) noted that combining a categorical instrument for initial diagnosis of ASD with several more assessments that utilize continuous scales to further specify symptomology may be most effective for assessing persons with ASD. Accordingly, as noted previously, the prosody scale I developed was designed to be used as part of a battery of assessments to serve the latter purpose, allowing assessment team members, including SLPs, to develop a precise view of the areas that are most pressing for intervention and those of relative strength for the person being assessed.

Perceptual Tools to Assess Prosody in ASD

In the present study, I developed and validated a continuous perceptual rating scale for assessing prosody in persons with ASD. The ability to quickly and reliably determine if the prosody of a person with ASD may be adversely affecting social, academic, and/or vocational functioning is important because intervention time is limited and there are many possible intervention targets for persons with ASD (Gillon et al., 2017; Smith & Gillon, 2004; Thurm et al., 2011). If prosody is not a significant problem for an individual with ASD, this needs to be ascertained quickly so that other areas of deficit may take precedence in assessment and intervention (Shriberg et al., 1992). On the other hand, if prosody atypicalities negatively affect functioning in social, academic, and/or vocational settings, this also needs to be detected quickly so that further assessment of prosody may be conducted and prosody targets can be incorporated into treatment planning.

There are few tools available for SLPs to use to assess prosody in ASD that have

adequate evidence of validity and reliability. While there are several validated assessments of speech prosody for other communication disorders (e.g. Hosokawa et al., 2017; Pernambuco, Espelt, & Costa de Lima, 2017; Strand, Duffy, Clark, & Josephs, 2014; Vaz Freitas, Pestana, Almeida, & Ferreira, 2014; Yiu, Chan, & Mok, 2007), at the time of writing, only three assessments of speech prosody had been validated for persons with ASD (de Villiers et al., 2007; McSweeny & Shriberg, 2001; S. Peppé & McCann, 2003; Shriberg et al., 1992). The first of these to be developed was the Prosody-Voice Screening Profile (PVSP; McSweeny & Shriberg, 2001; Shriberg, Kwiatkowski, & Rasmussen, 1990; Shriberg et al., 1992). Like the current prosody rating scale, the PVSP measure was designed as a screening tool to determine if prosodic and vocal characteristics of persons being evaluated were atypical enough to warrant further assessment. The test evaluates seven domains of prosody/voice: phrasing, rate, stress, loudness, pitch, vocal quality, and resonance. Evaluation is based on conversational speech with some utterances excluded if they meet certain criteria (e.g. they are single words, are unintelligible, have narrative or performance-like characteristics). Samples must be at least 12 qualifying utterances long, with 24 qualifying utterances required for more difficult cases or for research purposes, which equates to approximately 10 minutes of recording time. Coders rate each utterance and then the resulting codes are entered into a specialized computer program for graphing and analysis, a process which takes an average of 48 minutes, but which varies widely based on participant and rater characteristics. Initial training for novice users takes about 15 hours, on average (Shriberg et al., 1992). Validity was established through reviews of the literature, discussion with persons expert in voice and prosody analysis, pilot studies, acoustic analyses, and

comparisons of PSVP scores with expert perceptual analyses. Intra-rater reliability ranged from 71-100% and interrater reliability ranged from 74-100%. Relative to ASD, the measure was used to evaluate the prosody/voice characteristics of 30 persons with ASD who were high functioning as part of the validation process. McSweeny and Shriberg (2001) found that individuals with ASD did not differ from persons with NTD on the pitch section of the PSVP. Given that other research suggests that pitch of persons with ASD differs from those with NTD (Fusaroli et al., 2017; Kissine & Geelhand, 2019), it is possible that this measure was not fully reflective of pitch characteristics of persons with ASD. Additionally, the time required for training on and scoring this scale is likely to outweigh the proposed benefits of its use as a screening instrument by SLPs working with children with ASD. If reliability can be obtained on a shorter screener that requires less training, it would add substantially to the assessment arsenal of SLPs.

The second prosody assessment that was developed for persons with ASD is The Profiling Elements of Prosody in Speech-Communication test (Peppé, Maxim, & Wells, 2000; Peppé & McCann, 2003; Wells & Peppé, 2003; Wells, Peppé, & Goulandris, 2004). This test is a standardized, but not norm-referenced measure designed to assess prosodic function. The measure was designed to assess both receptive and expressive prosodic abilities in four functional categories: affect, interaction, focus, and chunking. Another two areas were added in the latest edition of the measure: phrase and lexical stress. The test was designed, as much as possible to isolate individual prosodic skills. However, in doing so, the authentic interactional contexts were sacrificed because most tasks in the test have a very prescribed context. The first revised edition (Peppé & McCann, 2003) included instructions to gather a semi-structured connected speech

sample but did not provide sufficient instructions for how to do so consistently across participants. The test takes approximately 45 minutes to administer and appears to be useful for selecting intervention targets but does not provide norms with which to compare participants' scores, except for approximately 30 persons with high-functioning ASD.

De Villiers and her colleagues (2007) developed a set of rating scales to assess the conversational skills of persons with ASD. They analyzed the conversational speech of 46, ten to thirteen-year-old children with ASD. Participants were all verbally fluent and participated in conversations with an adult from the research team for about 10 minutes. Conversations were listened to and transcribed to develop a set of nine initial codes of conversational breakdown. Interrater reliability ranged from fair to almost perfect, depending on the code (k = .34 - .82). The three codes most directly related to prosody (i.e. pausing, atypical stress and atypical intonation), had fair to moderate interrater reliability scores of k = .58, .43 and .79, respectively. The nine codes were later collapsed into five codes: atypical intonation, terse, semantic drift, perseveration, and pedantic speech. The resulting "Atypical intonation" and (less directly) "Terse" scales related to speech prosody and so I will limit further discussion of these rating scales to these areas. Atypical intonation was defined as a combination of speech that had "a flat, monotone or 'wooden' quality" (p. 1377) and "phonological stress (i.e., intonation and inflection) without contextual support for that stress" (p. 1377). The "terse" scale was a combination of the original "terse" and "pausing" scales and included minimal responsiveness and increased pause length.

While the atypical intonation scale in particular was similar in concept to our

scale, it differed in that it only covered the flat or monotone end of pitch variability in ASD, whereas the research on prosody in ASD suggests persons with ASD can present with a range of pitch variability, from monotone to typical to overly variable. Thus, the intonation rating scale presented by de Villiers et al. (2007) may be useful in some contexts but does not cover the range of prosody with which persons with ASD may present (Dahlgren et al., 2018; de Marchena & Miller, 2017; Fusaroli et al., 2017; Kissine & Geelhand, 2019). Also, the authors also did not calculate interrater reliability for the final version of their scale, which would be important for a full validation of the scales in question according to established standards for educational and psychological testing (American Educational Research Association et al., 2014).

All the three previous measures listed required significant training and familiarization time in order to use and/or did not have a full validation. Thus, valid, reliable, and *efficient* assessments of prosody in ASD are needed to support SLPs when making treatment decisions with individuals with ASD. As the previous assessments sought to evaluate specific prosody characteristics with varying results, for the purposes of screening, I determined that rating the broad construct of speech prosody with one item would result in the most efficient instrument. Additionally, given that evidence from the systematic review and meta-analyses performed by Fusaroli et al. (2017) on the univariate studies did not show consistent patterns in the sub-categories of prosody and even the machine-learning studies were not able to parse out what exactly what was "different" in the speech prosody of children with ASD, I did not expect SLPs to be able to reliably evaluate the individual factors of prosody for a screening with relatively little training. The goal of developing the present scale was to help streamline assessment for

practicing SLPs, so I tested the scale to see if no training or a very minimal amount of training would result in acceptable levels of inter- and intra-rater reliability.

Effects of Training on Ratings

Training can improve the inter- and intra-rater reliability of behavioral ratings scales (Barsties et al., 2017; Brinca et al., 2015; Eadie & Baylor, 2006; Eadie & Kapsner-Smith, 2011; Fay & Latham, 1982; Lee et al., 2009; Melchers et al., 2011; Støre-Valen et al., 2015; Tabuse et al., 2007). Brinca et al (2015) investigated the effect of anchors and training on the reliability of voice quality ratings for different types of speech stimuli. They found that, overall, intra-rater reliability was better than interrater reliability, and that training that included speech-sample anchors improved reliability. Similarly, Fay and Latham (1982) found that a 4-hour workshop-type training that focused on participation, feedback and practice improved participants' rating abilities.

Melchers et al. (2011) also investigated the effects of training on reliability measurements of a rating scale. While the rating scale in question differed significantly in the constructs it evaluated, the training methods used were pertinent for use in this context. Melchers and colleagues used a 2 x 2 between-subjects design to compare the effects frame-of-reference training to anchored rating scales on participants' abilities to rate performance of potential job candidates in the context of employment interviews. Frame of reference training referred to the process of giving instruction on a selected trait or concept for the purpose of developing a shared understanding of that trait or concept. Descriptively anchored rating scales referred to scales that provided explanations or examples associated with selected levels of the rating scale. The researchers filmed several responses to seven interview questions, then asked 199 participants to rate the

answers on a 5-point scale anchored at 1 (poor), 3 (average), and 5 (good). Participants received either frame of reference training or a control training prior to viewing the interview videos. For rating purposes, they received either a descriptively anchored rating scale or a graphically anchored rating scale. They found that both types of training improved reliability performance when compared to the control condition, but a combination of both types of training improved reliability the most. In other words, participants who developed a common conceptualization of a desired behavior (i.e. frame-of-reference) *and* who had the ability to reference examples and descriptions of the levels of the rating scale (i.e. anchors) were the most reliable on the scale.

For the current study, I modeled both the rating scale and the design of my training after the Melchers et al. (2011) study. All SLPs in my study had access to three "anchor" audio clips that represented the extremes and middle of the prosody rating scale during the rating process. The SLPs who were randomized to the training group participated in a brief, web-based frame of reference type training. I designed the training such that SLPs would finish with a common frame-of-reference for definitions of prosody and with an understanding of speech prosody in persons with ASD.

The efficacy of web-based trainings to improve reliability has been previously investigated. Støre-Valen et al. (2015) examined the efficacy of a web-based training program in improving reliability of ratings for an overall assessment of health functioning. The training included access to a manual and rating instructions, 20 vignettes with patients of varying symptom severity, visual feedback which compared their scores to gold-standard scores. The authors found that the participants with little to some experience with the rating scale benefitted the most from training in terms of improving

in reliability. Those with the most experience with the rating scale did not improve in their reliability because it was already very high. Because I validated a novel rating scale, I anticipated that the training would greatly improve reliability as SLPs gained familiarity with the construct of speech prosody and speech prosody characteristics in persons with ASD, practiced with the scale and received feedback on their performance compared to "gold standard" ratings.

Based on the findings and treatment designs of (Brinca et al., 2015; Fay & Latham, 1982; Melchers et al., 2011; Støre-Valen et al., 2015), the training in the present study incorporated principles of active learning (e.g. knowledge of performance), practice, and active participation. As noted previously, the training involved familiarization with the definition of prosody to create a common "frame-of-reference" for the construct. Additionally, the training included self-paced modules with feedback provided for correct or incorrect ratings based on comparisons with expert ratings. The scale itself was provided with examples of prosody for the end points and middle of the scale that could be referenced prior to listening to each audio clip to be rated for comparison and calibration purposes.

Purpose

The purpose of this study was to develop a brief rating scale of speech prosody that had adequate validity, reliability, and efficiency to be used to screen the prosody of children with ASD. A secondary purpose was to investigate whether a brief, web-based training would improve the reliability of this rating scale.

Research Questions

1. Is this prosody rating scale reliable?

a. What is the interrater reliability of the scale as measured by intraclass correlations (ICCs) between raters?

I predicted that the overall interrater reliability would be moderate to good when collapsed across all the SLPs since I anticipated the interrater reliability of the SLPs who do not receive training to pull the overall interrater reliability down. I hypothesized that interrater reliability would improve significantly after training because raters would have a stronger, more uniform idea of what each level of the rating scale represented, which would be less vulnerable to individual interpretation.

b. Does interrater reliability differ according to whether or not SLPs received training prior to using the scale?

I hypothesized that interrater reliability would be higher for those SLPs who received training because of their improved understanding of prosody generally and prosody in ASD specifically.

c. What is the intra-rater reliability of the scale as determined by ICCs between scores given by raters two weeks apart?

I hypothesized the intra-rater reliability would be moderate to good given that the interval between ratings would not be very long. I anticipated that the intra-rater ICCs for those who did not receive training would not be as stable as those who received training. I

d. Does intra-rater reliability differ according to whether or not SLPs received training prior to using the scale?

I hypothesized that intra-rater reliability would be higher for the SLPs who received training. I anticipated that intra-rater reliability would improve significantly after training because raters would develop a more stable conceptualization of each level of the rating

scale, insulating their ratings from other potential influences (e.g. mood).

2. Is this prosody rating scale valid?
    a. Does this prosody rating scale provide adequate evidence of validity in its test content?

I hypothesized that aligning my scale with current definitions of speech prosody, building it to reflect the range of prosodic presentation in children with ASD, and evaluating if SLPs' rationale for their responses related to the construct of speech prosody would sufficiently establish the evidence of validity in test content.

    b. Does this prosody rating scale demonstrate adequate evidence of validity in the response processes engaged in by raters?

The cognitive process I theorized the SLPs would engage in were as follows:

1. Listen to the audio clip,
2. Compare what they heard with their existing conceptualization of speech prosody,
3. Evaluate the variability of speech prosody in the clip with speech prosody in children with ASD and NTD based on their experience with these groups,
4. Provide a score based on that evaluation.

I expected that evidence of validity based on raters' response processes would accumulate if SLPs' rationales for their prosody ratings was relevant to speech prosody because this would reflect good alignment between the anticipated and actual cognitive processes. Evidence of validity in this area would be negatively impacted if SLPs gave irrelevant rationale for their ratings because this would reflect a lack of engagement with

the task and/or an incorrect or incomplete conceptualization of speech prosody and thus lack of alignment with the anticipated cognitive process.

I expected that response time would also have an effect on the evidence of validity based on raters' response processes. I hypothesized that if SLPs took longer than 3 days to complete a survey, their responses would decrease in validity because they might not develop and maintain a consistent internal conceptualization of the rating scale to apply across all audio clips, and if they were in the training group, they might not be able to recall the training.

c. Does this prosody rating scale demonstrate sufficient evidence of validity based on its relations to other relevant variables?

    i. What is the evidence for concurrent discriminant validity as measured by the associations of scores on this prosody rating scale with narrative proficiency?

Because the prosody rating scale was designed to measure a construct very different from narrative proficiency, I did not expect narrative proficiency to be strongly associated with the prosody rating scale. If there were no significant associations between narrative proficiency and the prosody rating scale, I hypothesized that it would be because the two measures were assessing separate constructs. Thus, the dissociation between the measures would provide evidence for the concurrent discriminant validity of the prosody rating scale, even though the audio samples were taken from a narrative discourse.

    i. What is the evidence for concurrent convergent validity as measured by the relations of scores on this prosody rating scale with F0 and F0 standard deviation measures?

I expected that acoustic measures of pitch, i.e. F0 and F0 standard deviation, would account for a moderate to large amount of the variance in prosody rating scale scores given that prosody is comprised of several acoustic measures of pitch, but may include elements that acoustic measurements fail to capture.

ii. What is the evidence for concurrent convergent validity as measured by identifying the prosody rating scale score at which raters consistently rated children as less socially acceptable, defined as a social acceptability score of 5 or more?

I anticipated social acceptability ratings would account for a moderate to large amount of variance in prosody rating scale scores. This is because previous research has shown that persons with atypical prosodic patterns are rated as less likeable (Redford et al., 2018). I expected that the less socially acceptable the child was rated, the more extreme the prosody rating scale score would be.

iii. What is the evidence for concurrent convergent validity as measured by identifying at what score on the prosody rating scale children had a 75% chance of being rated as having ASD?

This question was primarily exploratory in nature. I did not know at what point(s) on the scale children would be more likely to be rated as having ASD. I did expect there to be an association at both ends of the scale, however. I anticipated that this variable would be significantly associated with the prosody rating scale because sometimes the features that humans pick up on to make disorder judgements are not readily picked up by acoustic measures (Munson et al., 2003) and thus may be accurately measured by a perceptual scale. Further, speech prosody has been associated with “frank” impressions of ASD (de

Marchena & Miller, 2017; Fusaroli et al., 2017; Kissine & Geelhand, 2019), so I hypothesized that there would be a point at which children were more likely to be classified as having ASD. I expected that when the judgement of ASD was present the prosody rating scale scores would become more extreme.

d. Are the consequences of testing, both intended and unintended, adequately anticipated and accounted for in the development of this prosody rating scale?

I hypothesized that if more atypical prosody was associated with less social acceptability and increased impression of presence of ASD, the consequences of testing could be positive in that children might receive services for prosody when they otherwise might not. I posited that demographic variables such as geographic region, years of experience, level of expertise with ASD, and/or work setting (school, private practice, university clinic, etc.) and children's demographic characteristics could play a role in SLPs' prosody ratings and consequent assessment and intervention decisions. I did not anticipate significant differences in the speech prosody ratings based on demographic characteristics of the children in the audio clips.

CHAPTER 2

METHOD

Participants

Children with ASD and typical development. Audio samples were drawn from two sources. The first source was a set of narratives told by five children (three boys and two girls) with ASD who participated in a single-subject multiple-baseline across participants narrative intervention study (Gillam et al., 2015). The majority of these audio samples came from the pre- and post-test administration of the TNL. A total of eight pre- and post-test TNL McDonalds's retell stories were included because the post-test recording of participant 001 and the pre-test recording of participant 004 were not available. An additional three clips from the children in this study were used: one as an anchor clip and two for training purposes. These participants with ASD were recruited from a local autism clinic and ranged in age from 8 to 11 years old. All participants achieved a score of 70 or above on the screening portion of the Universal Nonverbal Intelligence Test (UNIT; Bracken & McCallum, 1998) and were classified as "verbally fluent" on the Autism Diagnostic Observation Schedule (ADOS-2; Lord et al., 2012). The language abilities of the participants varied widely with Core Language Scores on the Clinical Evaluation of Language Fundamentals, 4th Edition (CELF-4; Semel, Wiig, & Secord, 2003) ranging between 48 and 114, yet each participant demonstrated difficulty in narrative language ability, whether in all aspects of narrative or primarily in organizational components. Participants' narrative ability was assessed prior to and immediately after intervention using the first version of the TNL (Gillam & Pearson, 2004). The Narrative Language Ability Index scores, which have a similar interpretation

as standard scores, ranged widely at both pre- and posttest. At pre-test, they ranged from 55 to 115 (60 points). At post-test, the gap narrowed, but scores still ranged from 76 to 124 (48 points).

The second source of audio samples was a set of narratives from children who participated in the normative samples of both editions of the TNL (Gillam & Pearson, 2004, 2017). These children have been previously described in the TNL manuals (Gillam & Pearson, 2004, 2017); however, I have included more specific information on the subset of participants whose audio I used in this study (see Table 1). For eight of these audio clips, I matched participants as closely as possible to the sex, age, and narrative proficiency scores of the five participants from Gillam et al. (2015) at pre-and post-test assessment. If no close matches for all three characteristics emerged, I matched first on sex, then on age, and finally on narrative performance. Seven additional audio clips from this source were included: one for training, four for validity, and two for anchoring purposes. The sample of recordings in this study was relatively balanced in terms of males and females (3:2). Females were slightly over-represented in this sample when compared to national prevalence estimates (3:1; Loomes et al., 2017).

Table 1
Characteristics of Children in the Audio Clips

Characteristic	Female (n = 11)	Male (n = 15)
Age in Months[a]	101.1 (14.4)	121.4 (13.4)
Narrative Proficiency[b]	104.5 (11.9)	81.9 (14.6)
Race		
African American	1 (9.1%)	1 (6.7%)
White	10 (90.9%)	14 (93.3%)
Ethnicity		
Hispanic	3 (27.3%)	0 (0%)
Nonhispanic	8 (70%)	15 (100%)
Diagnosis		
ASD	5 (45.5%)	9 (60%)
No ASD	6 (54.5%)	6 (40%)

[a]Age and narrative proficiency numbers represent M(SD), the remainder of characteristics represent Count(percent of total).
[b]There were two children who did not have narrative proficiency scores, so those means and standard deviations were based on the reduced participant total.

SLP rating teams. There were two groups of SLPs: Expert SLPs and Rating SLPs.

Expert SLPs. Three SLPs with expertise in ASD were recruited to listen to audio clips and select anchors for each end and the middle of the prosody rating scale and to provide "gold standard" ratings for audio clips. The a priori recruitment criteria for these SLPs were that they were required to hold a current Certificate of Clinical Competence in Speech-Language Pathology (CCC-SLP) from the American Speech-Language-Hearing Association (ASHA), have at least five years of clinical experience, with at least three of those years working closely with children with ASD, and a working knowledge of the literature relating to prosody in ASD. As part of the rating process, demographic information such as sex, geographic region, professional experience, etc. were gathered

from these SLPs. None of the Expert SLPs participated in data collection or analysis of the studies from which the audio samples were drawn. Demographic information of all SLPs in the study, including Expert SLPs, is displayed in Table 2.

Table 2
SLPs' Demographic Information and ASD experience According to Group

Characteristic	Total n = 45	Experts n = 3	Not Trained n = 22	Trained n = 20
Sex[a]				
Female	43 (95.6%)	3 (100%)	21 (95.5%)	19 (95%)
Male	2 (4.4%)	0 (0%)	1 (4.5%)	1 (5%)
Race/Ethnicity				
Asian-American	2 (4.4%)	0 (0%)	2 (9.1%)	0 (0%)
White	43 (95.6%)	3 (100%)	20 (90.9%)	20 (100%)
Region				
Midwest	10 (22.2%)	1 (33.3%)	4 (18.2%)	5 (25%)
Northeast	5 (11.1%)	0 (0%)	4 (18.2%)	1 (5%)
South	10 (22.2%)	0 (0%)	3 (13.6%)	7 (35%)
West	20 (44.4%)	2 (66.7%)	11 (50%)	7 (35%)
Education				
Master's	42 (93.3%)	1 (33.3%)	21 (95.5%)	20 (100%)
Doctorate	3 (6.7%)	2 (66.7%)	1 (4.5%)	0 (0%)
Years Practicing[b]	15.6 (10.2)	21.3 (15.6)	18.0 (10.5)	12.1 (8.3)
# Years Worked w/ASD	14.9 (9.3)	21.3 (15.6)	17.2 (10.3)	11.3 (5.7)
% Total Clients w/ASD	37.7 (21.4)	63.3 (20.2)	35.5 (22.4)	36.2 (18.6)
% Clients w/ASD	42.8 (27.3)	51.7 (29.3)	49.2 (30.4)	34.4 (21.8)
% Current Clients w/ Level 3 ASD	25.7 (23.1)	36.7 (20.2)	22.1 (15.8)	28.1 (29.5)
% Current Clients w/ Level 2 ASD	28.9 (20.2)	35.0 (13.2)	35.0 (23.0)	21.2 (15.3)
% Current Clients w/ Level 1 ASD	29.8 (27.3)	28.3 (20.2)	28.6 (24.9)	31.3 (31.4)

[a]Values given as Count(Percent of Total).
[b]Values given as M(SD).

Rating SLPs. Once the Expert SLPs established anchors for the rating scale, 42 more SLPs were recruited through professional networks to rate the audio clips using the

prosody rating scale. To participate, these SLPs were required to hold the CCC-SLP from ASHA or be within one month of completing their clinical fellowship experience prior to certification. All needed to have at least one academic year of clinical experience beyond their graduate degree. Additionally, they needed to have provided regular individual or group services to at least one child with ASD who was between the ages of 7 and 12 within the last year. As part of the rating process, demographic information such as sex, geographic region, professional experience, etc. were gathered. None of these SLPs participated in data collection or analysis of the studies from which the audio samples were drawn.

While participants had the option of typing in their preferred gender category if it was something other than male or female, no participants took this option. No participants indicated they preferred not to answer. All participants marked either male or female. The percentage of males in my sample (4.4%) is comparable to that of the overall percent of male SLPs according to ASHA (4%; American Speech-Language-Hearing Association, 2019). All rating team participants were Asian American or European American. The percentage of SLPs who were European American (96%) in this study was slightly higher than the percentage of SLPs in ASHA who are European American (92%; American Speech-Language-Hearing Association, 2019). The percentage of SLPs who were Asian American was slightly higher in my sample (4.4%) than Asian American SLPs who are members of ASHA (3%; American Speech-Language-Hearing Association, 2019). No participants said they preferred not to provide their race/ethnicity, no participants typed in alternate race/ethnicity categories. There were no SLPs who were African American or Black, American Indian or Alaska Native, Hispanic or Latinx,

Middle Eastern or North African, or Native Hawaiian or Other Pacific Islander.

Materials

Speech samples. As noted previously, the audio clips in this study were taken from the pre- and post-test administrations of the TNL in Gillam et al. (2015), the normative samples of the TNL (Gillam & Pearson, 2004, 2017), and narrative retells and novel narrative generations of the children who participated in Gillam et al. (2015). For clips used in the direct validation of the prosody rating scale, I selected 30-second clips from the McDonald's subtest of the TNL for both children with ASD and NTD. By selecting 30-seconds from the same narrative context, I standardized the length of the samples as well as controlled for variability in linguistic content. I used a mixture of McDonald's retells, novel narrative generations, and narrative retells for indirect validation, training, and anchor clips. In all of the clips, the 30-seconds were not taken from the very beginning or very end of the narratives to ensure that the more stereotyped prosodic patterns associated beginnings and endings of stories were omitted. Similarly, I clipped all sections of dialogue from the audio clips. By excluding dialogue, I avoided the use of character voices and/or overly exaggerated prosody that might be appropriate for a narrative performance of a character but would not be appropriate in typical social interaction.

All audio clips were recorded using digital voice recorders. Audio was then uploaded to a secure server housed at Utah State University (USU). Because this was an analysis of secondary data, recorder type, placement, and distance from the participant were not standardized. Manipulations of the audio files were completed using *Adobe Audition.* To the extent possible, extraneous sounds (e.g. examiner voice, tapping), were

removed from the audio clips. All clips were converted to a single audio channel to accommodate the fact that some sound files had one channel and some had two. The amplitude of the clips was equalized to -20 dB prior to ratings being given.

Rating scales. SLPs rated each audio clip on three separate scales. The scale of primary interest for this study was the prosody rating scale being validated. The remaining two were a social acceptability scale and a binary rating of presence or absence of ASD. I used these scales to support questions related to the prosody rating scale. All scales are described below.

Prosody rating scale. This rating scale was a single-item, 7-point continuous analogue scale ranging from one to seven (see Figure 1). A rating of one represented an overly monotonous prosody pattern, which was defined as prosody that was flat, uninflected, and lacking in variation so much so that it frequently distracted from the message being communicated. A rating of four represented a prosody pattern that would be typically expected of a child's narrative retelling and did not draw attention to itself. A rating of seven represented overly variable or "sing-songy" prosody that was so overly modulated, variable, and exaggerated that it frequently distracted from the message being communicated. These descriptive labels were selected after reading published literature on the characteristics of prosody in ASD and listening to audio clips from the narratives of children with ASD and children with NTD. I developed the scale to reflect the range in prosody that I heard in the audio clips as well as the range of prosody patterns reflected in the published literature on prosody in ASD from very unmodulated to overly variable (Dahlgren et al., 2018; Diehl et al., 2009; Diehl et al., 2015; Fusaroli et al., 2017; Kissine & Geelhand, 2019).

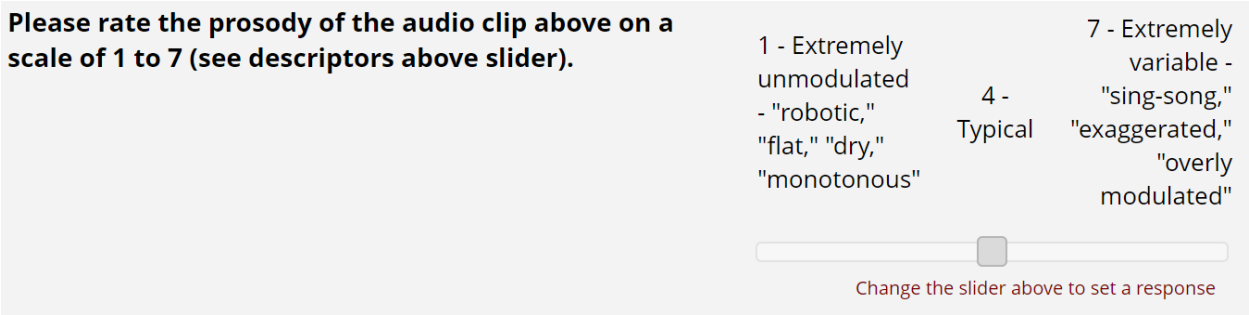

Figure 1. Screenshot of the prosody rating scale as the SLPs saw it. SLPs clicked and dragged the grey box along the light rectangle to indicate how unmodulated, over modulated, or typical a given child's prosody was. This scale was presented immediately below each audio clip to be rated on the survey.

I chose a 7-point continuous analogue rating scale rather than an ordinal scale to more adequately capture the nuances of prosody and to facilitate more precise analyses (Finstad, 2010; Grant et al., 1999; Pfennings et al., 1995). When given too few response options, as in a 5-point ordinal scale, for example, raters have been shown to place ratings in-between response options (e.g. to put their response as "between a 2 and a 3") rather than definitively selecting one of the numbers (Finstad, 2010). Additionally, utilizing a continuous scale enabled the use of statistical procedures which required less power and could make more precise predictions. I elected to provide anchors at the ends and the middle of the scale to provide some guidance in rating, particularly in the middle range, since mid-scale ratings have been shown to be less consistent between and within raters (Kreiman & Gerratt, 1998).

As can be seen in Figure 1, raters saw the numbers "1", "4", and "7" with the descriptive labels "Extremely unmodulated – 'robotic,' 'flat,' 'dry,' 'monotonous'",

"Typical", and "Extremely variable – 'sing-song,' 'exaggerated,' 'overly modulated,'" placed beside each of the anchor numbers, respectively. A continuous slider appeared beneath the scale labels for participants to mark their ratings. Audio exemplars representative of the anchor points of the scale were available in the section immediately prior to each audio clip for referencing and calibration.

Rating of social acceptability. In conjunction with rating the prosody of each audio clip, SLPs rated the social acceptability of the child in the audio clip on a scale from one to seven. I used a rating scale of social acceptability adapted from the "likeability" scale presented in Redford et al. (2018). A score of one represented high social acceptability ("This child would definitely be socially well-accepted by peers and adults") and a score of 7 represented low social acceptability ("This child would definitely *not* be socially well-accepted by peers and adults"). I altered this scale by changing it from a "likeability" to a social acceptability scale and by adjusting the descriptive language at each end of the scale to reflect the change. I made these adjustments in the scale because SLPs' ratings of how much they personally like a child might not be reflective of the child's social acceptability to peers and adults in general.

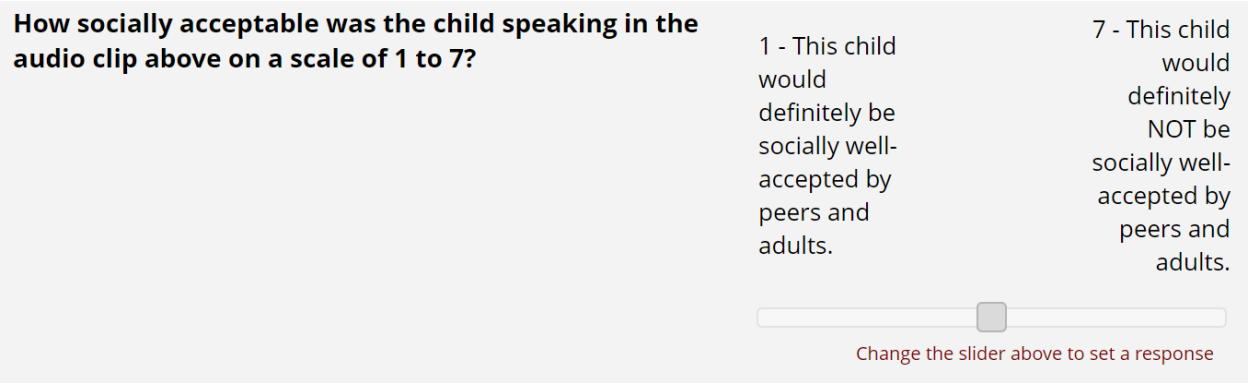

Figure 2. Screenshot of the question asking SLPs to rate the social acceptability of the child speaker in a given audio clip. SLPs clicked and dragged the grey box toward 1 or 7 to indicate how socially accepted they thought the child would be by peers and adults. This question was presented after the SLPs completed the prosody rating, but before they indicated their impression of presence or absence of ASD in the child.

Rating of presence or absence of ASD. Finally, participants rated their perception of the presence or absence of ASD for each clip. They answered the question "Would you classify the child speaking in the above audio clip with ASD?" by clicking radial buttons presented to the left of the words "yes" and "no". While SLPs typically do not independently diagnose ASD, I included this question to investigate the notion that atypical speech prosody contributes to the "frank" presentation of ASD (de Marchena & Miller, 2017).

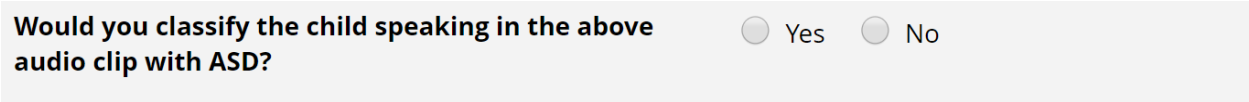

Figure 3. Screenshot of the question asking SLPs to rate the presence or absence of ASD of the child speaker in a given audio clip. The SLPs clicked the radial button next to their rating. This question was presented after the SLPs completed the prosody rating and the social acceptability rating.

Surveys. All ratings of audio clips were collected and managed using REDCap (Research Electronic Data Capture) tools hosted at Utah State University (Harris et al., 2009, 2019). REDCap is a HIPAA-compliant, web-based software platform designed to support data capture for research studies. REDCap offers a variety of response options including, but not limited to, multiple choice, multiple response, customizable sliding scales, and short and long response boxes. Longitudinal data are automatically linked together. Links, audio, photos, and video may be embedded into survey questions. I chose this system due to its security, customizability, and ability to track individuals' responses across multiple time points.

Training. To determine whether training improved reliability on the prosody rating scale, 20 of the Rating SLPs were randomly selected to participate in a 15-30-minute, self-paced on-line NearPod training module embedded within a REDCap survey (see Figure 4). NearPod is a cloud-based program that allows instructors to embed a variety of content (such as audio clips) and learning activities into a presentation to create an interactive, engaging learning experience (*NearPod*, n.d.). The survey sent to the Rating SLPs who were part of the training group included an embedded link to the

NearPod training module. SLPs received instruction on the concept of prosody in general and prosody in ASD specifically using text, audio, and visual supports. Then, the prosody rating scale was presented with one audio exemplar for each anchor of the scale to orient the SLPs to the levels of the scale. Finally, the SLPs practiced rating three different audio clips and checked their ratings against the “gold standard” ratings of the Expert SLPs.

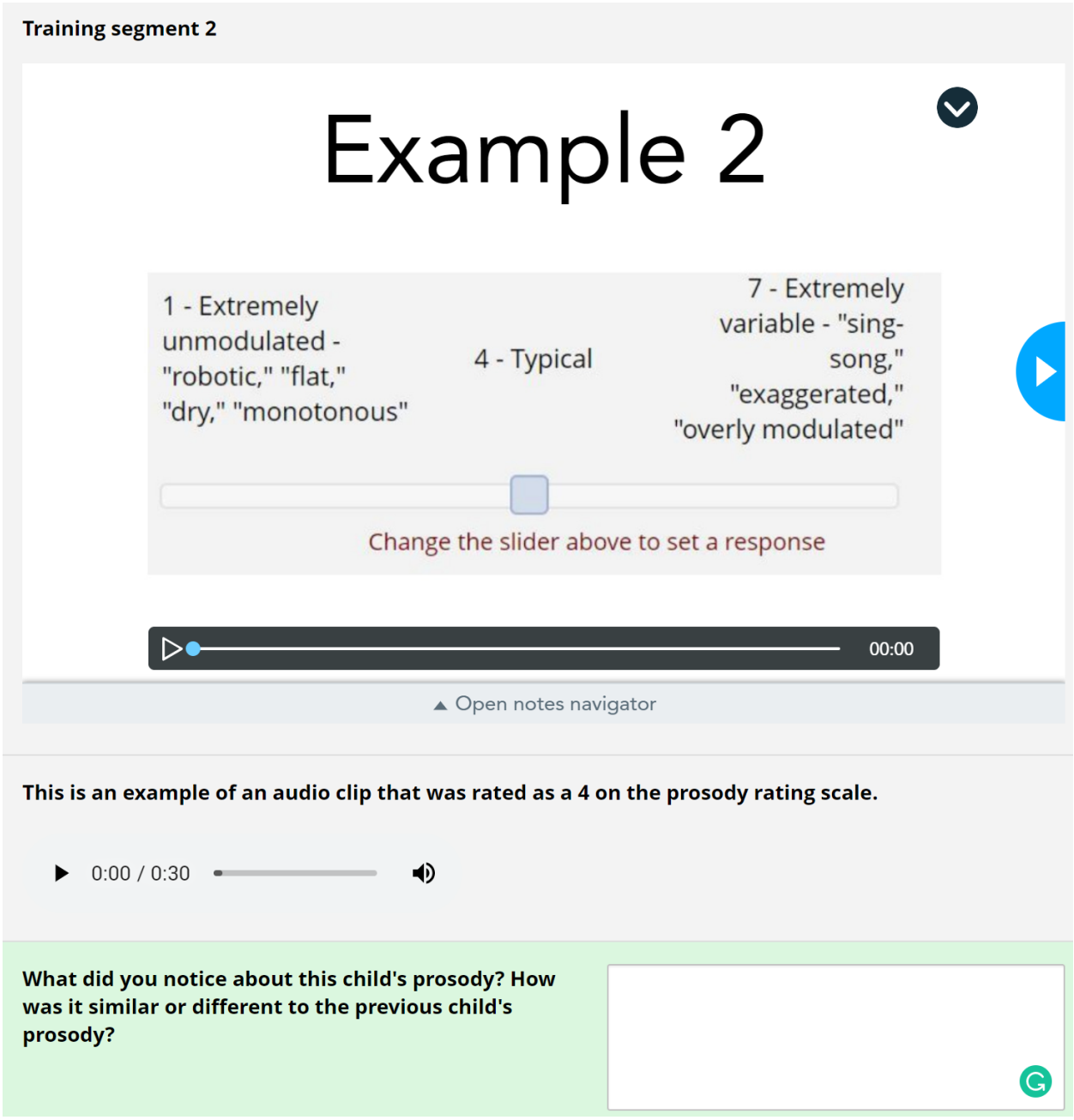

Figure 4. Screenshot of part of the online training module. The white box at the top is where the SLPs saw the NearPod presentation. SLPs clicked on the white triangle to listen to pre-recorded training materials, then clicked on the black triangle below to listen to audio. Any comments were entered in the free-response areas.

Procedure

I validated the prosody rating scale in the following steps:

1) tentative audio clips were selected for SLPs to rate and to serve as anchors for the scale,

2) Expert SLPs selected anchor clips and developed associated consensus ratings,

3) a preliminary version of the REDCap survey was developed and sent to the Expert SLPs,

5) audio clips were selected to use in the training,

6) the NearPod training and the REDCap surveys were finalized,

7) the REDCap survey was sent to Rating SLPs,

8) audio clips were acoustically analyzed in Praat (Boersma & Weenink, 2013), and results of the ratings were analyzed.

Audio clip selection and validation. I first clipped 30 seconds from the pre- and post-test McDonald's retells from the children in the Gillam et al. (2015) study. I also selected other narrative retell and generation clips from this study that could serve as potential anchors or training clips. Next, I identified children from the TNL normative databases who were closely matched on sex, age, and narrative proficiency to the children in the Gillam et al. (2015) study and clipped 30 second segments from their McDonald's retells. I compiled a list of potential anchor clips from these two groups of children and presented the resulting 15 audio clips to the Expert SLPs to rate. None of the potential anchor clips came from the pre- or post-test administration of the TNL in Gillam et al. (2015). The Expert SLPs independently rated each of 15 potential anchor stories on the 7-point rating scale. I selected three clips that the SLPs rated from this list

to use in the online training. Consensus scores on these three clips were based on averages of the Expert SLPs' individual scores.

Once the Expert SLPs identified the anchor clips for the scale, they rated the sixteen clips from the pre- and post-test McDonald's retells of the eight children in Gillam et al. (2015) and the age, sex, and narrative proficiency matched children from the TNL normative sample. Additionally, the Expert SLPs re-rated four of the clips from the first survey, for a total of 20 clips rated in this second survey. Shortly after completing these second ratings, the Expert SLPs met together came to consensus on a final "gold standard" score for each of the audio clips.

Rating SLP survey completion. Once the anchor stories were chosen, I created a REDCap survey for the Rating SLPs to complete. Rating SLPs were assigned to the training or no-training group based on a pre-defined list of participant numbers with a random assignment of training/no-training next to each number. Rating SLPs randomized to the training group received the REDCap survey with the NearPod training module embedded. Those randomized to the no-training group received the same REDCap survey, but without the NearPod training included.

After instructions and/or training, Rating SLPs listened to 20 audio clips. They indicated their prosody rating of the clip, their rating of the social acceptability of the child in the clip, and their impression of whether the child had ASD or not. The Rating SLPs provided their reasoning behind each rating before proceeding to the next audio clip. Rating SLPs were asked to complete the survey again after approximately two weeks had elapsed to obtain intra-rater reliability information. Once all the Rating SLPs completed the survey, the data were downloaded from REDCap and stored in a secure

folder on USU's Box system.

Acoustic parameter extraction. I used Praat to extract mean, minimum, maximum, and standard deviation of F0 for each audio clip. All clips had a sampling rate of 44.1 kHz and a sampling window length of 0.005 seconds. I set the minimum pitch to 100Hz and the maximum to 700Hz. I sought to cover a broad range of possible fundamental frequencies because previous research suggested it can be in the higher F0 ranges where differentiation between ASD and NTD occurs (Diehl & Paul, 2013). All acoustic parameters were collected in Hz because this is the globally accepted measurement of frequency (Taylor & Thompson, 2008). All pitch data must be interpreted with caution because of the overall low quality of recordings available.

Analysis Plan

Power and sample size analysis. Hox et al.'s (2010) observation that "optimal design is a question of balancing statistical power against data collection costs" (p. 235) applies in this case. Regarding sample size in multi-level modeling, Hox, citing Kreft (1996), suggested following the 30/30 rule of thumb if fixed parameters in the models are of most interest. That is, ideally, researchers have a sample size of 30 groups with 30 observations in each group when conducting multi-level model analyses where the primary interest is in the fixed parameters of the models. If researchers are interested in cross-level interactions, the sample size needs to be closer to 50 groups with 20 observations in each group. As I was interested in both the fixed parameters and any relevant cross-level interactions, my goal was to have 50 groups with 20 observations in each group. Given my budget of $2,169, I could afford to pay three SLPs $100 each to serve as the expert raters and the 42 SLPs who rated the 20 audio clips $40 each. I

gathered 60 initial ratings by the Expert SLPs, which was much less than the recommended sample size, so their ratings were used for anchor selection and in developing the “gold standard” ratings of prosody, not in the multi-level model analyses. A total of 833 ratings each of prosody, presence/absence of ASD, and social acceptability were collected from the 42 Rating SLPs at time 1 and 800 ratings were collected from the 40 SLPs at time 2 (differences due to incomplete surveys or attrition). Ultimately, 760 ratings were used for the final models at time 1 and 741 ratings at time 2. These numbers came close to the 50/20 rule at 40/19 for time 1 and 39/19 at time 2, with each SLP serving as a “group” and each rating serving as an “observation”. For a visual of the data collection procedure and sample structure, see Figure 5.

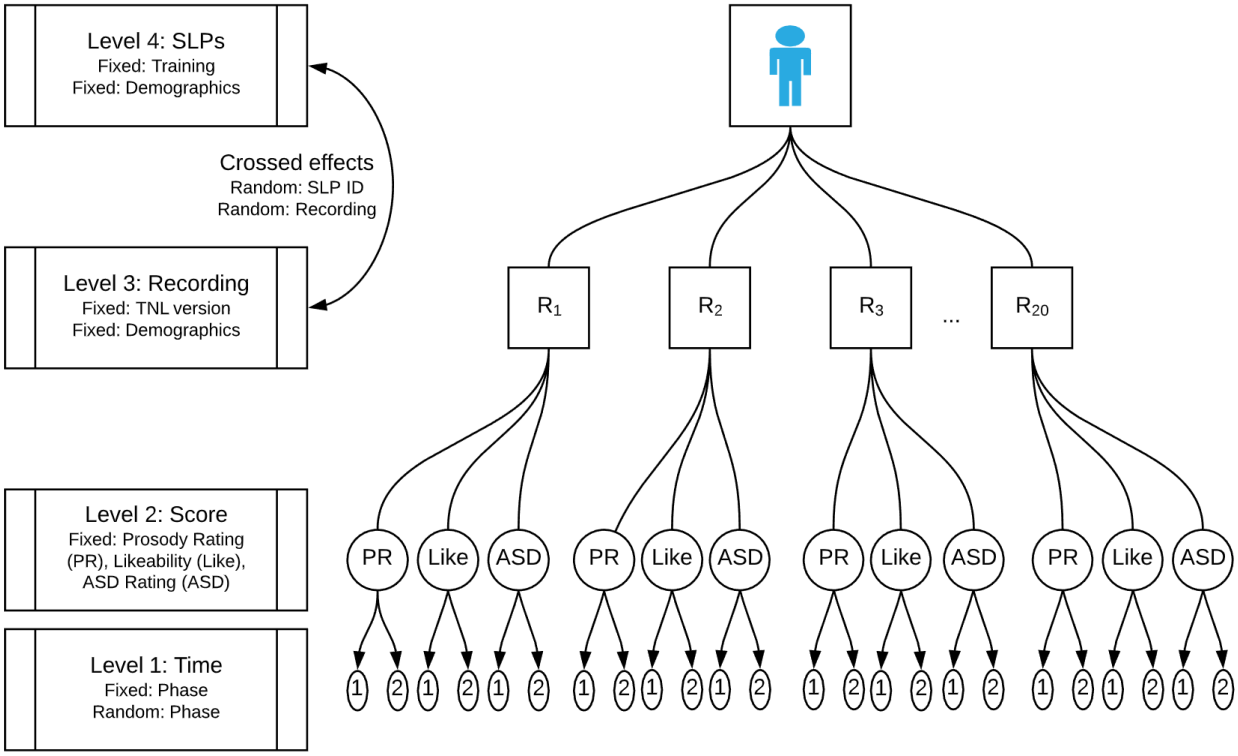

Figure 5. Visualization of the complete data collection structure with model parameters listed in boxes at the appropriate level. SLPs who rated the audio clip and their characteristics are represented at level 4, then the 20 audio recordings are represented at level 3. The three ratings that were given to each recording are at level 2. Level 1 represents the ratings for each scale given at time 1 and time 2.

Model fitting. Model fitting was conducted using a research question, theory driven modeling approach. I started building models in the order of my research questions. I compared initial models to the null model, then, significant predictors were added in a nested approach to the models for the next research question. Subsequent models were compared with the most representative model previously constructed. Comparisons between models were performed using a likelihood ratio test. Non-

significant terms were trimmed based on feedback from the likelihood ratio tests of nested models until the most parsimonious model that adequately represented variance was found. Variables that demonstrated significant skewness and/or prevented convergence in the models were transformed as appropriate.

Missingness. Participants who completed less than half of the survey at either time point were dropped from analyses. One participant rated just over half of the recordings (i.e. 13/20) at time 1. Their responses were included in analyses, but for recordings 14-20, data were missing at the lowest level of analysis. One participant contacted me to say that they had rated the first half of their first survey with an incorrect conceptualization of the scale, so their responses were dropped from analyses utilizing data from the first time point (i.e. all analyses except intra-rater reliability). In all, one participant's data were excluded from analyses using data from the first time point. Three participants' data were excluded from analyses using data from the second time point (i.e. intra-rater reliability) due to attrition.

When investigating the discriminant validity of the scale, I excluded all audio clips where the children had incomplete TNL data. Because the TNL data did not significantly improve the models when ASD diagnosis was controlled for, all subsequent modeling was conducted using data sets that included participants with incomplete TNL data. I hypothesized that the Expert SLPs represented a different underlying population of SLPs from that of the Rating SLPs and there were too few of them to analyze as an independent group, so their responses were not included in calculations of interrater and intra-rater reliability nor for model fitting.

After reading feedback from several SLPs who stated that Recording 8 did not

have enough speech to get a sense of the child's prosody, this recording was removed from the datasets for the final model fitting, resulting in one less child with ASD whose speech was rated.

Reliability of the prosody rating scale. To calculate reliability estimates, I used functions within the *irr* (Gamer et al., 2019) and *psych* (Revelle, 2019) packages in R. the Participants with missing data were excluded from reliability estimates because the methods used to calculate reliability did not tolerate missingness. Intra-class correlations (ICC) were used for both interrater and intra-rater reliabilities. For interrater reliability, only data from the first time point were utilized. Comparisons were made between the raters. For intra-rater reliability, data from both time points were utilized and comparisons were made between them.

Interrater reliability. My participants were drawn from a larger population of SLPs who worked with children with ASD, so I chose the model option of a "twoway" model over the "oneway" option in the *"icc()"* function of the *irr* package (Gamer et al., 2019). Since I was interested how consistent the group of Rating SLPs were with each other for interrater reliability (instead of how consistent they were independently), I set the type option as "agreement" instead of "consistency." Due to the fact that I used only the first time point ratings to calculate interrater reliability (rather than a mean of all ratings), I set the unit to "single" instead of "average." I calculated the overall interrater reliability and the training and no-training group-level interrater reliability for of the Rating SLPs. In all cases, values of .75 or greater indicated "good" interrater reliability (Koo & Li, 2016).

Intra-rater reliability. To evaluate Rating SLPs' consistency in prosody ratings over time, I used the "*testRetest()*" function in the *"psych"* package in R 3.6.1 to calculate the ICCs of this multi-level dataset (Revelle, 2019). Similar to interrater reliability, in all cases values of .75 or more were interpreted as "good" intra-rater reliability (Koo & Li, 2016).

Effects of training. To determine the effects of training, I interpreted a fully crossed, multi-level mixed effects model (MLM) with SLPgroup added as a fixed effect to a fully nested model with fixed effects that had previously been found significant to see if group membership improved the model. In addition to the omnibus intra-rater reliability ICC, separate inter- and intra-rater reliability ICCs were calculated for SLPs who were trained and those who were not.

Validity of the Prosody Rating Scale

Content and rater response processes validity. Evidence of validity based on the test content was determined by aligning the scale with current published definitions of prosody (Szczepek Reed, 2011; 't Hart et al., 1990; Tench, 1996; Wichmann, 2000) and descriptions of prosody and pitch variability in children with ASD in published literature (Dahlgren et al., 2018; Fusaroli et al., 2017; Kissine & Geelhand, 2019; Nakai et al., 2014, 2017; Paul, Augustyn, et al., 2005) as recommended in the standards set out by AERA, APA, and NCME (2014). This area of validity was also established by determining if SLPs' rationales for giving prosody ratings related to the overall construct of prosody, as described below in conjunction with SLPs' response processes.

Evidence of validity based on raters' response processes was measured by asking all raters to describe their rating process by telling why they rated audio clips as they did

and by measuring survey completion time. As long as the participants gave a relevant response for their rationale, I coded it as correct. If they gave an off-topic or irrelevant response (e.g. XX or sldkfjso), I coded it as incorrect. If they did not provide a rationale, I counted it as missing.

If participants took longer than 3 days to complete any one survey, it called into question the validity of their response process because they might not have been able to remember training items (if applicable) and would have been less able to calibrate with themselves and maintain intra-rater reliability from item to item. To investigate the effect of lengthy completion times, this variable was included as a predictor in validation MLMs for the prosody rating scale (see below).

Evidence of validity based on relations to other variables. Evidence of validity based on the relations to other variables was assessed through MLMs. The use of these type of models allowed for assessing associations between the relevant variables while simultaneously controlling for random variation between and within audio clips and raters. Multi-level modeling has been proposed as a desirable alternative to more traditional types of analyses due to its ability to handle missing data and to account for relatedness of covariates (Baayen et al., 2008; Quené & van den Bergh, 2004).

To assess the discriminant evidence of validity, I built an MLM predicting prosody rating scale scores based on narrative proficiency as measured by the TNL (Gillam & Pearson, 2004, 2017) while controlling for audio recording, SLP raters, and test edition variability. Because variability due to test edition was better accounted for by ASD diagnosis, this variable was used instead of TNL edition (see Figure 6).

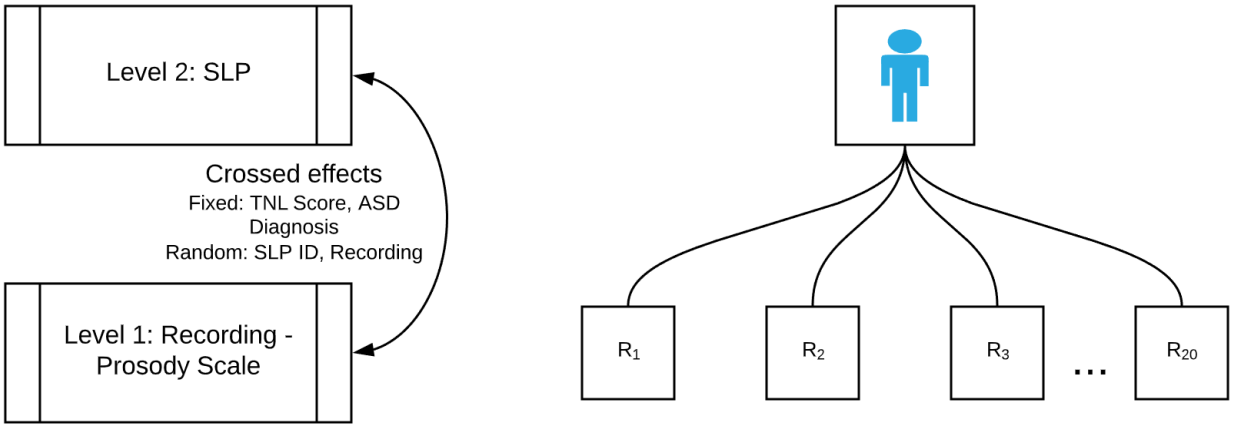

Figure 6. MLM figure for evidence of discriminant validity as indicated by associations between prosody ratings and TNL total quotient scores while controlling for ASD diagnosis with random effects for individual SLPs and recordings.

Evidence for convergent validity was assessed by examining relations between the prosody rating scale and variables including acoustic measures of F0 and F0 standard deviation, ratings of social acceptability, and ratings of presence/absence of ASD. After collecting the SLPs' prosody rating scale scores, I explored patterns in these variables in the data during the preliminary data explorations, then built the MLM (see Figure 7).

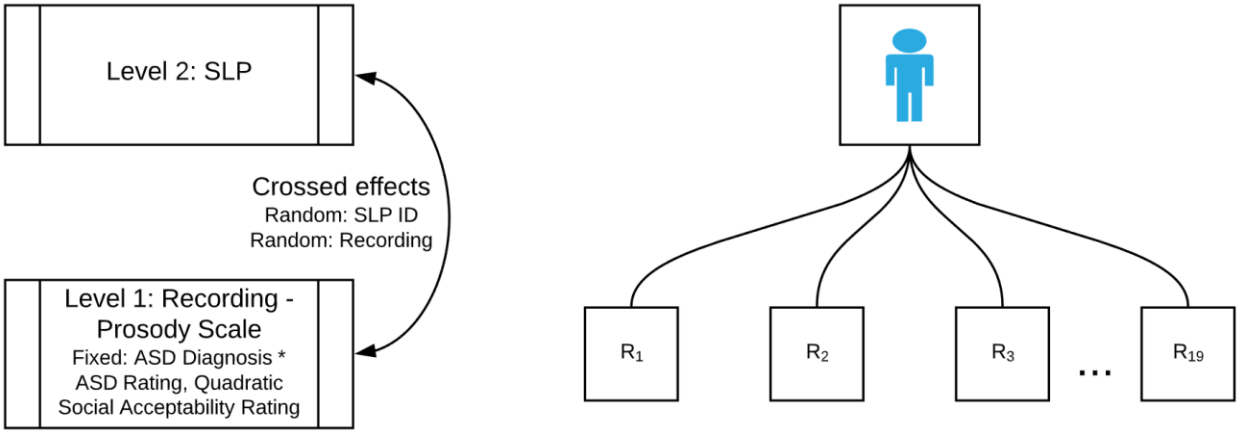

Figure 7. MLM figure for associations between prosody ratings and significant convergent validity measures including the interaction between ASD diagnosis and ratings of presence/absence of ASD social acceptability and its quadratic term with random effects for individual SLPs and recordings.

Additionally, to determine the prosody scores at which children were rated as having a 5 on the social acceptability rating scale, I fit MLMs examining the associations between social acceptability ratings and the same variables, added in the same order as in the MLMs for the prosody rating scale (see Figure 8).

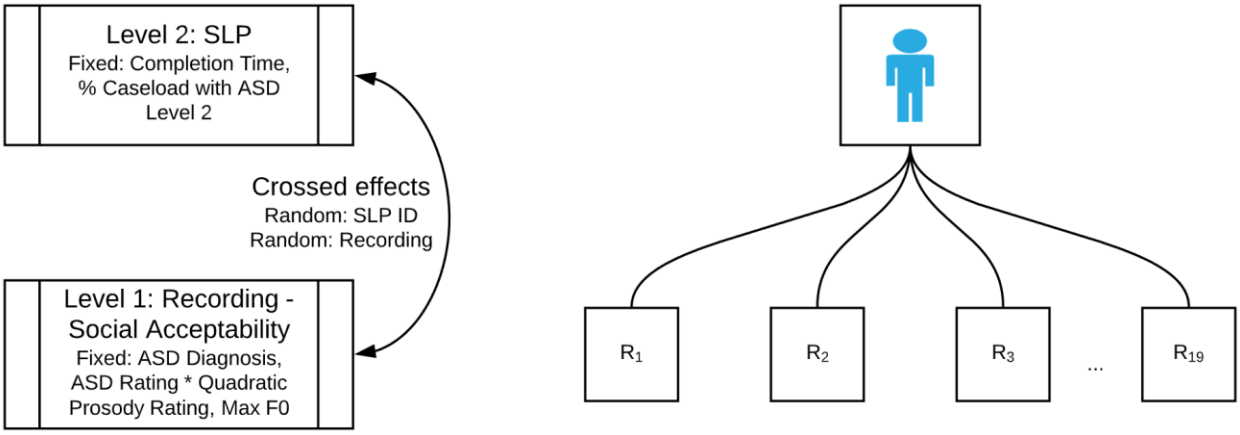

Figure 8. MLM figure for associations between social acceptability and significant fixed and random variables, including ASD diagnosis, the interactions between ratings of presence/absence of ASD and prosody ratings and the associated quadratic term and maximum F0 in recordings with random effects for individual SLPs and recordings.

To establish the point on the prosody rating scale at which children had a 75% chance of being rated as having ASD, I fit MLMs investigating associations between judgements of presence/absence of ASD and the same variables as the models for the prosody rating scale, with ratings of prosody taking the place of ratings of ASD in the model (see Figure 9).

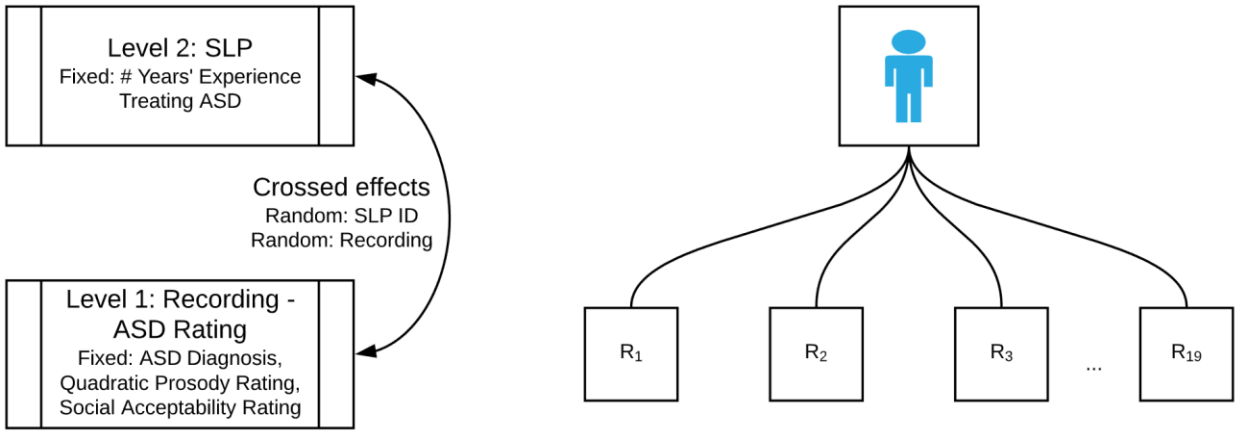

Figure 9. MLM figure for associations between ratings of presence/absence of ASD and significant variables including ASD diagnosis, prosody rating scale scores and the associated quadratic term, and social acceptability ratings with random effects for individual SLPs and recordings.

Evidence of validity based on consequences of testing. After determining the associations between the prosody rating scale and variables relating to discriminant and convergent validity, I investigated variables that provided evidence for the validity of the scale based on the intended and unintended consequences of testing.

Consequences of testing based on demographic variables such as sex and race of both the rater and the children in the audio clips were explored. I explored the effects of variables such as geographic region, years of experience, level of expertise with ASD, and/or work setting (school, private practice, university clinic, etc.) and children's demographic characteristics in the observed data during the preliminary data exploration phase. When meaningful were established, I added the relevant demographic variables as fixed effect predictors in my statistical models (see Figure 10).

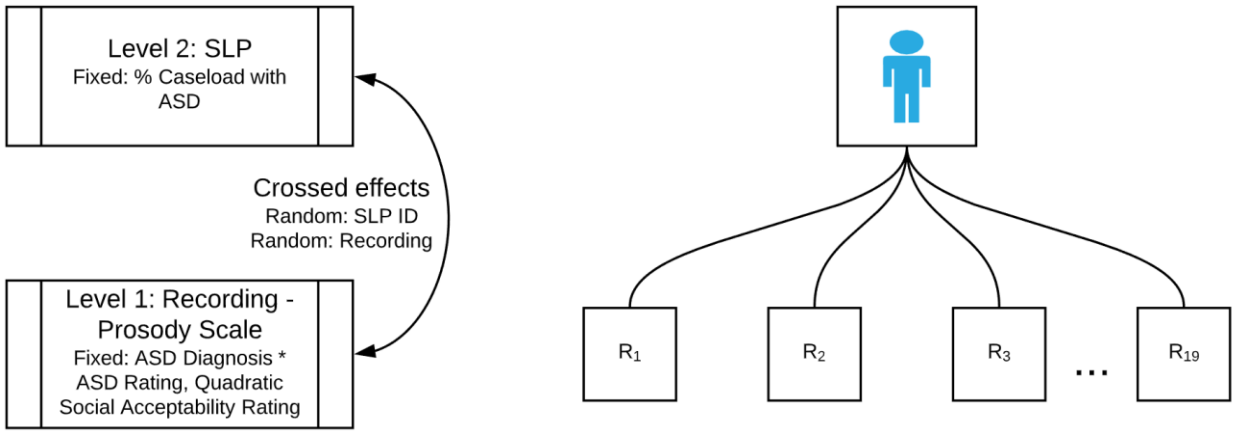

Figure 10. MLM figure for associations between prosody ratings and significant fixed and random variables including the interaction between ASD diagnosis and ratings of presence/absence of ASD, social acceptability and its quadratic term, and the percent of children with ASD that Rating SLPs had on their current caseload with random effects for individual SLPs and recordings.

All analyses were performed using the *"lme4," "irr,"* packages in R 3.6.1 (Bates et al., 2015; Gamer et al., 2019; R Core Team, 2019; Revelle, 2019; Wickham, 2009). The results of these explorations are given in the section below.

CHAPTER 3

RESULTS

The purpose of this study was to develop and validate a brief prosody screening instrument to be used as part of a larger battery of assessments to evaluate the strengths and challenges of children with ASD. To this purpose, I asked two main research questions. The first research question related to the reliability of the scale and whether a brief, web-based training would improve reliability. The second question asked whether the scale was valid, with four specific sub-areas of evidence of validity investigated, namely, test content, response process, relations to other variables (e.g. convergent and discriminant validity), and consequences of testing. Results are presented in tables and graphically to highlight the meaningful relations I found between variables and predictions (Bell, 2002; Pike & Rocconi, 2012).

Reliability

All reliability estimates were calculated on the dataset without Recording 8 in it.

Interrater reliability. These calculations were based on the scores of 40 raters, 19 who were trained and 21 who were not trained. Using the above specifications, the overall interrater reliability for both the Trained and Not Trained SLP groups together was .586, $F(18, 707) = 60.8$, $p < .001$, 95% CI [0.441, 0.759]. This level of interrater reliability is considered moderate (Koo & Li, 2016), but was lower than my a priori target number of .75. The interrater reliability for the Not Trained group was .546, $F(18, 364) = 27.7$, $p < .001$, 95% CI [0.395, 0.731] which was again moderate, but was lower than the overall interrater reliability and below my a priori target of .75. Interrater reliability for the Trained group was better than that of the Not Trained group at .631, $F(18, 327) =$

35.3, $p < .001$, 95% CI [0.482, 0.793], but it was still moderate and below my a priori target number of .75.

Intra-rater reliability. A total of 39 participants completed both surveys, 20 who were trained and 19 who were not. The overall mean within person, across item reliability was 0.73, (*SD* = 0.18). This was slightly below my a priori number of .75 and still moderate reliability, similar to interrater reliability. The mean within person, across item reliability for the group of SLPs who were Not Trained was moderate as indicated by an ICC of 0.68, *SD* = 0.2. The training group's reliability was higher at an ICC of 0.76, *SD* = 0.15, which was considered "good" according to Koo & Li (2016) and was above my a priori number of .75.

Validity

I next addressed the validity of this prosody rating scale in the areas recommended by the Standards for Educational and Psychological testing published by the AERA, APA, and NCME (2014).

Content validity. Validity evidence for the content of the prosody rating scale followed a primarily logical process based on definitions of speech prosody and intonation proposed by previous authors (Nicolosi et al., 2004; Shriberg et al., 1990, 1992; 't Hart et al., 1990; Tench, 1996; Wichmann, 2000), descriptions of speech prosody in ASD in published literature (Dahlgren et al., 2018; Fusaroli et al., 2017; Kissine & Geelhand, 2019; Nakai et al., 2014, 2017; Paul, Augustyn, et al., 2005) and consultations with expert SLPs well-versed in prosody measures. Additional support for the content validity of this measure was gathered by analyzing SLPs' rational for why they gave the prosody ratings they did for each recording, described in the next section.

In my initial conceptualizations of the scale, I used a definition of "prosody" that was too narrow; it essentially mimicked measures of F0, as one of my expert consultants pointed out. So, I adjusted the scale to reflect a broader definition of prosody embraced in much of the literature on prosody in ASD (see, for example, Fusaroli et al., 2017; Peppé, 2009). Ultimately, I created a working definition of prosody which represented a synthesis of the definitions I found in the literature (Diehl & Paul, 2009; Nicolosi et al., 2004; Shriberg et al., 1990, 1992; Szczepek Reed, 2011; 't Hart et al., 1990; Tench, 1996; Wichmann, 2000). The resulting amalgamated definition is as follows:

> Prosody, broadly defined, is the melody of speech. Prosody involves the suprasegmental aspects of speech. It is a conglomerate of elements that exist above the level of phonemes, morphemes, words, and sentences. These elements include intonation or pitch, stress, loudness, rate, rhythm, and, at times, vocal quality. Prosody lacking in variation is unmodulated, flat, dry, or robotic sounding. Overly variable prosody may be exaggerated, "sing-song", and/overly modulated.

Response processes validity. I used two metrics to assess response process validity: SLPs' given rationale for each audio clip and survey completion time. The first analysis also provided support for the content validity of the scale. To assess content and response process validity, I gathered SLPs' rationale for why they gave the audio clips the prosody rating they did, then coded the answers as either relevant, non-relevant, or missing. There were no missing rationales for prosody ratings given. Seven of 1532 (0.46%) responses across both time points were coded as not relevant. The responses coded as non-relevant were not related to prosody or demonstrated a misunderstanding of

the purpose of the scale (e.g. commented on how pleasant a child's prosody was to listen to instead of evaluated the holistic characteristics of the speech prosody). These responses tended to be evaluations of the narrative performance of the child (e.g. "Somewhat slow and laborious narrative. Question if the student is disinterested or if he requires extra language formulation time"), the likeability of the child or their prosody (e.g. "She made her story interested and I wanted to listen to it, " "I like her prosody. I think she could read a story or tell a story well. It would keep my attention") or were unclear (e.g. "Nice job by student"). All other rationales were coded as relevant to the question asked and ranged from quite short, as in "typical" or "WNL" to lengthier, as in:

> This child's prosody was mostly typical but her expressive response was slow, as if she could not formulate a response or recall the story. As she had a response or recalled the story, her rate of speech increased as did the intonational contour of her speech.

Thus, the majority of SLPs' rationale fit within my overarching definition of speech prosody and related to the specific question asked (e.g. "What was your reasoning for giving the score you did for this child's prosody?").

While evaluating the relevance of SLPs' rationales, I noted that several SLPs commented on the paucity of speech in Recording 8, so they were not able to get a good sense of the child's prosody. Of 81 responses from the SLPs across both time points for this recording, 19 (23.5%) mentioned something about the audio quality of this recording affecting their ratings. Illustrative examples included:

- "The sample was too small to get a good feel…",

- “Based on the limited sample (seemed incomplete), it was monotonous….Not a good sample to judge from.”
- “This one is hard to hear…”
- “This one was hard, as there was so little speech in the clip.”

I built many of my models with this recording included, but ultimately decided to exclude it because the validity of the SLPs’ responses was called into question when they didn’t have enough audio to work with. I ran the final models using a data set without this recording in it. The results reported for the final models were based on this smaller data set.

I investigated the effects of completion time in observed data explorations. There were no limits on how long the SLPs could take to complete the survey. The overall range of completion time was very large (.68 to 578.95 hours), but highly skewed. Most participants took between one and two hours complete the study (see Figure 11). SLPs in the trained group took an average of 120.8 (SD = 162.4) hours to complete the survey while the not trained group took an average of 34.1 (SD = 77.8) hours to complete it.

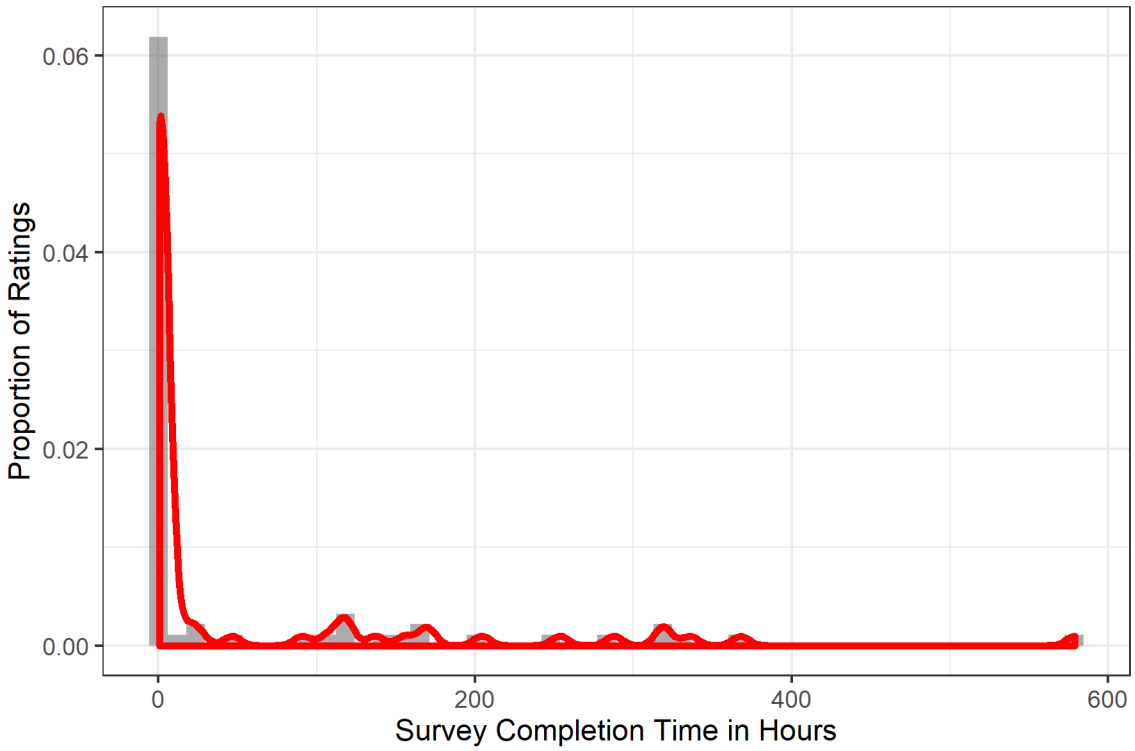

Figure 11. Observed data histogram of completion time of SLPs at both time points. Red line is an overlaid density plot highlighting the general patterns of data.

Recall that in my method I planned to transform variables that demonstrated skewness and/or resulted in model convergence problems. To that end, completion time was log transformed because it was skewed toward the shorter completion times. I classified completion time as an "SLP" factor, so I added it onto the previous best fit model that included the diagnosis of ASD of the child in the recording (ASD Diagnosis), mean F0, maximum F0 and the associated quadratic term, and social acceptability rating and its quadratic term (see Figure 10). This model was built with the dataset including Recording 8. Ultimately, the log likelihood ratio test between the model with completion time and the nested model without it indicated that completion time did not significantly

improve the model, $b = 0.01$, SE = 0.02, $\chi^2(11) = 0.42$, $p = .516$. Thus, I did not include this factor when I re-ran the models using the reduced dataset without Recording 8. The lack of significant differences between the two nested models suggested that even though there was a wide range in how long it took SLPs to complete the survey, the ratings of children's prosody were not significantly associated with survey completion time.

Ultimately, the process that the SLPs engaged in to rate the prosody of the children in this data set appears to be consistent with what I expected because the large majority of reasons given for the SLPs' prosody ratings were relevant to prosody and the question asked. Further, completion time was not significantly associated with the prosody ratings, which suggested that SLPs' response process, whether short or long, did not affect the validity of their responses.

Relations to other variables. Relations to other variables were evaluated in terms of evidence for concurrent discriminant validity, concurrent convergent validity, and consequences of testing.

Concurrent discriminant validity. To evaluate this aspect of validity, I examined the correlations between children's narrative proficiency and prosody ratings given by the Rating SLPs in the observed data (see Figure 12). These correlations revealed a negligible negative correlation (-0.19) between narrative proficiency and prosody ratings.

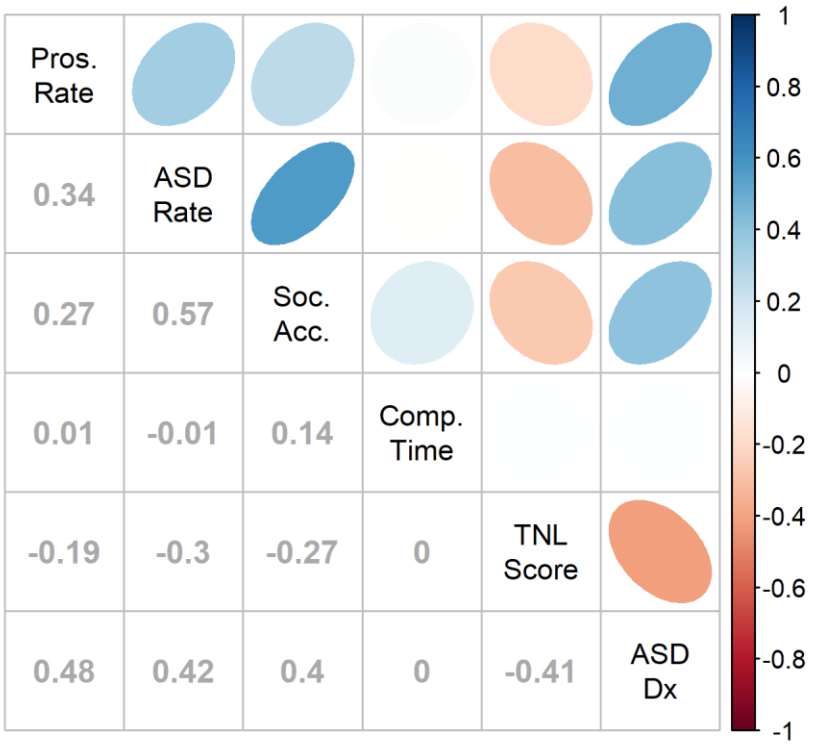

Figure 12. Correlation plot between the prosody rating scale (Pros. Rating), SLPs' ratings of children's ASD diagnosis (ASD Rate), the social acceptability rating scale (Socacc), the time it took SLPs to complete the surveys (Comp. Time), children's narrative proficiency scores (TNL score), and children's actual diagnosis of ASD or NTD. Negative correlations are shown in variations of red. Positive correlations are shown in variations of blue. Darker colors indicate stronger correlations, as do narrower ovals. Actual correlation numbers are presented to the left of the graph in shades that represent the direction and strength of the correlation.

To confirm what I discovered in the observed data, I built an MLM with children's narrative proficiency as the main fixed effect variable. Children missing a score for narrative proficiency (n = 2) were excluded from this MLM analysis. Because the TNL2 was published in 2017 and the audio from the children with ASD were collected prior to 2015, there were two versions of the TNL in my dataset. Therefore, I included TNL version as a control variable in this MLM. When I compared the model with narrative proficiency and TNL version in it to the model with just TNL version, there was no significant difference in the models, $b = 0.0001$, SE = 0.01, $\chi^2(6) = 0.00$, $p = .995$, suggesting that the TNL measured a separate construct from the prosody rating scale.

Interestingly, while the overall measure of narrative proficiency was not significantly associated with prosody ratings, TNL version was. This did not make theoretical sense because the versions should have been roughly equivalent in their relation to prosody. After some investigation, I realized that all but two of the children with ASD had been given the first version of the TNL, so it was likely that the variability attributed to the TNL version was actually a function of ASD diagnosis. Consequently, I added ASD Diagnosis into the model at this stage in the model building process rather than later as I had planned. Comparisons between models with just ASD Diagnosis and ASD Diagnosis plus TNL version were not significantly different when subjected to a likelihood ratio test, $b = .42$, SE = 0.90, $\chi^2(6) = 0.21$, $p = .643$), supporting the notion that the variance attributed to the TNL version previously was more accurately attributed to ASD diagnosis. I included ASD Diagnosis in all subsequent models.

Concurrent convergent validity. I first examined the relations of several variables with the prosody rating scale scores in the observed data, beginning with F0 measures, then moving to comparisons with the social acceptability scale ratings and the ratings of presence/absence of ASD. After I investigated relations in the observed data, I built MLMs to further elucidate relevant relations between variables.

Observed data. I investigated the relations between the prosody rating scale and four measures of F0: mean F0, minimum F0, maximum F0, and standard deviation of F0. The raw Pearson product moment correlations between these variables are shown in Figure 13. The measure of F0 that correlated most strongly with the prosody ratings was mean F0 at $r = 0.51$, which was weaker than I expected. Also surprising was the fact that standard deviation of F0, an acoustic measure of pitch variability, demonstrated a very weak correlation with the prosody ratings ($r = -0.08$).

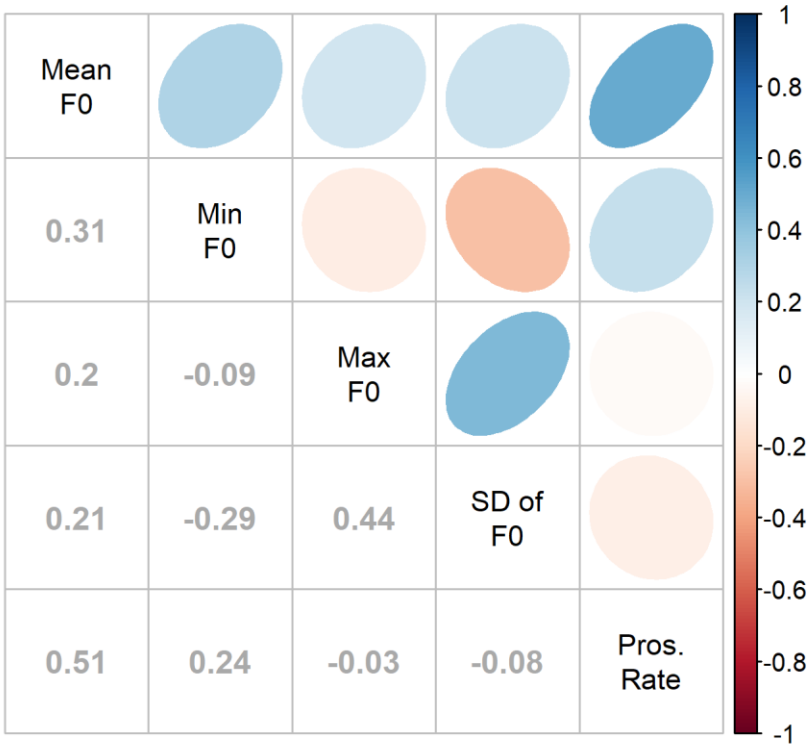

Figure 13. Pearson's correlation plot between the prosody rating scale (Pros. Rating) and children's mean F0, minimum F0, maximum F0, and standard deviation of F0 (SD of F0). Negative correlations are shown in variations of red. Positive correlations are shown in variations of blue. Darker colors indicate stronger correlations, as do narrower ovals. Actual correlation numbers are presented to the left of the graph in shades that represent the direction and strength of the correlation.

I investigated the relations between the prosody rating and mean F0 further by looking at how it may have been related to SLPs' ratings of presence/absence of ASD and the children's actual diagnosis of ASD, as shown in Figure 14. Interestingly, children

who were rated as not having ASD and truly did not have ASD were generally rated as having typical prosody on the prosody rating scale, regardless of their mean F0. The children who were rated as having ASD, regardless of whether they actually had ASD or not, in general, had higher mean F0 values and were rated with more variety along the prosody rating scale (right facet of Figure 14).

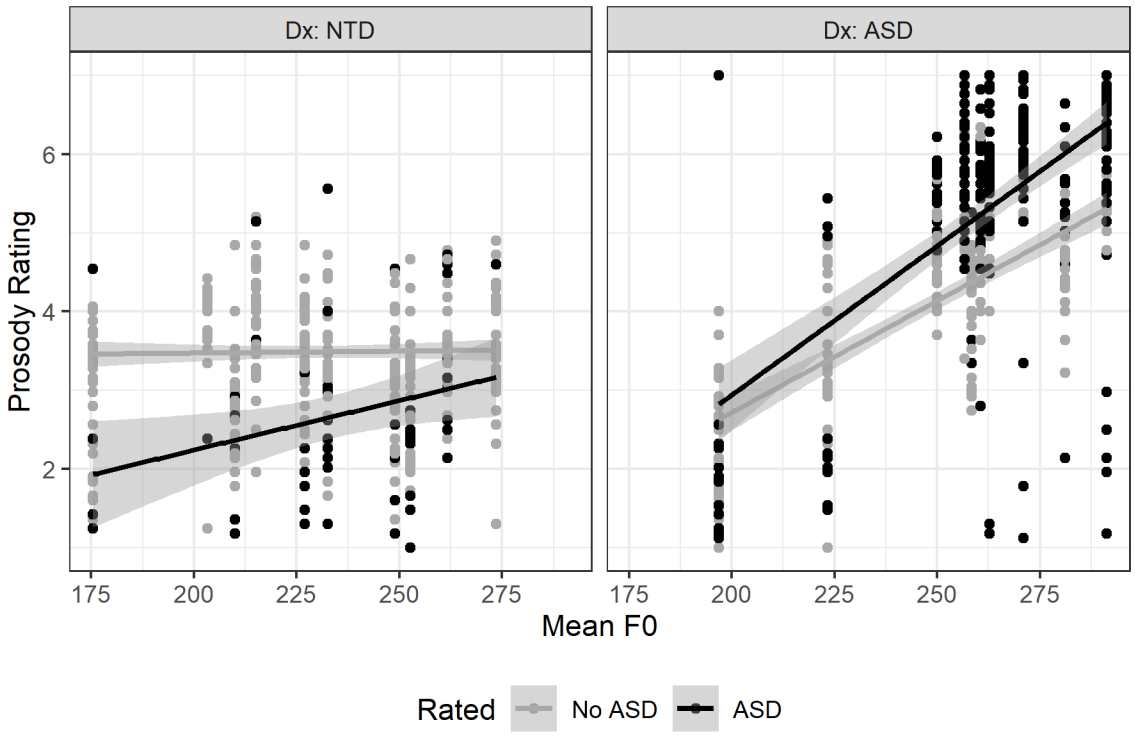

Figure 14. Observed data scatter plot of relations between mean F0 and prosody ratings given by 42 Rating SLPs on the original 20 audio recordings. A total of 833 ratings were represented due to one SLP's incomplete ratings. The plot is faceted by children's actual diagnosis of ASD and colored by SLPs' ratings of whether the children had ASD. Lines represent independently fit simple linear regressions with the naïve assumption of independence.

Along with measures of F0, I investigated relations between the prosody rating scale and SLPs ratings of presence or absence of ASD. As can be seen in Figure 15, SLPs were not accurate in their ratings of whether or not children had ASD, particularly for children who actually had a diagnosis of ASD. Their accuracy was only about 50-60% for children who had ASD. They were more accurate for children who did not have ASD, with the percentage being more around 75% accurate for this group.

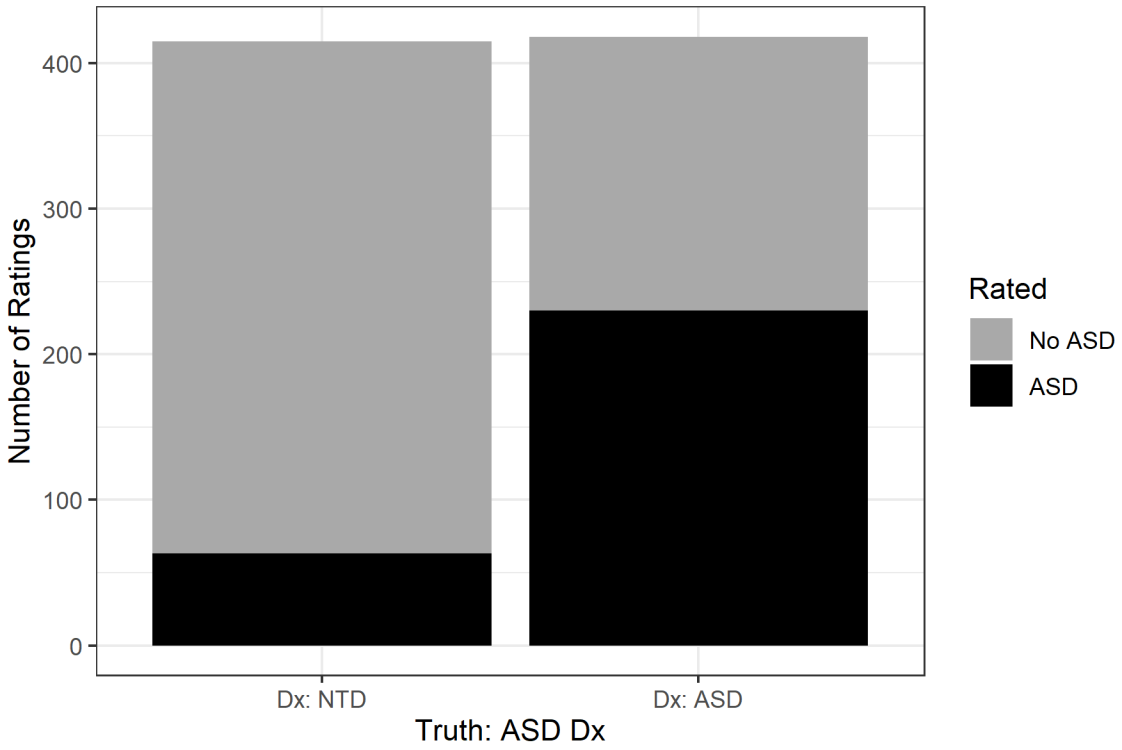

Figure 15. Observed data bar plot of the total number of ratings SLPs gave for the children's audio clips (n = 833), faceted by children's actual diagnosis (ASDn = 418, NTDn = 415). Colors within the bars represent SLPs' ratings of presence/absence of ASD. Uneven rating numbers were due to one participant not completing ratings for recordings 14-20.

Further investigation into the relations between prosody ratings and children's actual diagnosis is displayed in the histograms found in Figure 16. Most of the prosody ratings of children with ASD were generally in the more variable range, but ratings were present along the range of the prosody rating scale. The shape of the curve is much different for the children with NTD. The bulk of their ratings were clustered around 3.5-4, suggesting that children with NTD were generally more typical sounding with some sounding more monotonous. There were no prosody ratings for children with NTD beyond about a 5 on the prosody rating scale.

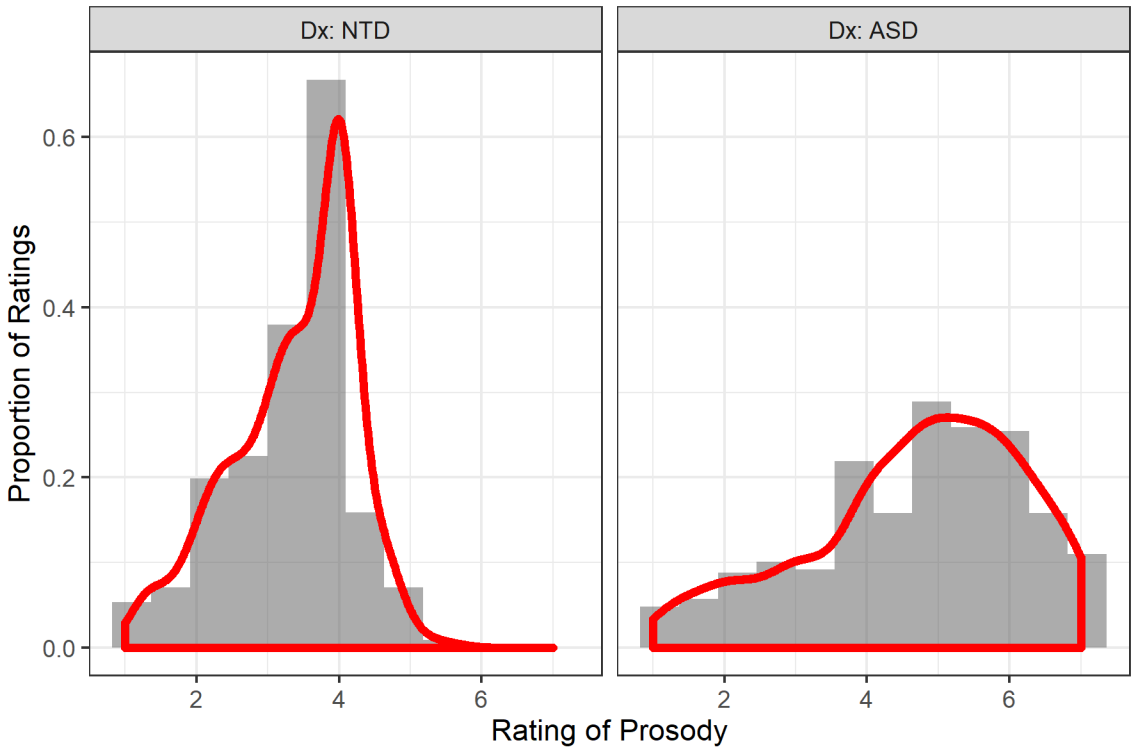

Figure 16. Observed data histogram of prosody ratings (total = 833) given by Rating SLPs at Time 1 faceted by actual ASD diagnosis of the children in the audio recordings. Red lines are an overlaid density plot highlighting the general patterns of data.

As I continued to investigate relations between variables in the data, a 3-way interaction emerged between SLP group, ASD Rating, and actual ASD Diagnosis, see Figure 17. Notice how the Expert SLPs were the only group of SLPs that did not inaccurately rate children as having ASD when they did not have it (see right panel). It also seemed as if the Expert SLPs had a more solidified conceptualization of what they thought ASD "sounded" like. The Expert SLPs rated all of the children that had ASD who they correctly rated as having ASD as having more variable prosody.

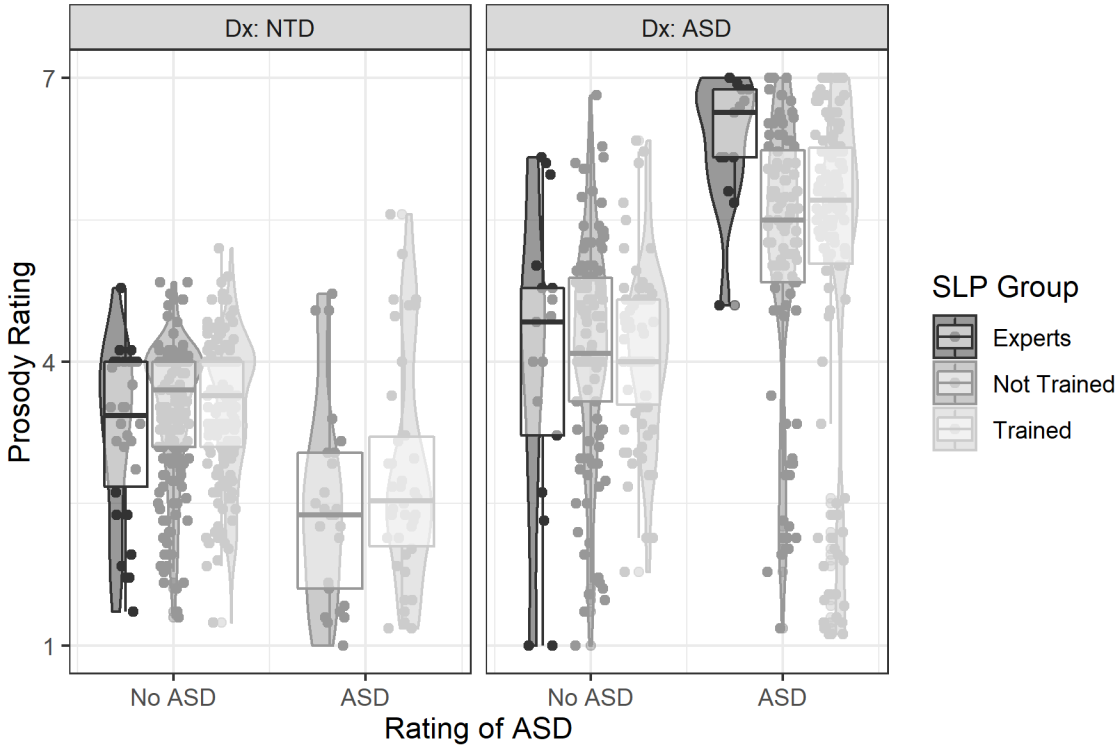

Figure 17. Observed data violin plot with overlaid box plot visualization of the 3-way interaction between SLP group, ASD Rating, and ASD Diagnosis.

Model building. To further assess the relations of the prosody rating scale with other measures, I built a series of nested MLMs with fixed effects variables drawn from constructs directly related to or similar to speech prosody. These included mean, minimum, maximum F0, standard deviation of mean F0, SLPs' ratings of the children's social acceptability, and SLPs' ratings of whether they thought the child in the audio clip had ASD or not. I began this section of model building with ASD Diagnosis only as a fixed effect and added fixed effects individually, keeping those with significant associations with prosody ratings as they surfaced. I investigated quadratic terms of all fixed effects except the binary rating of presence or absence of ASD because I constructed the prosody rating scale with typical prosody in the middle of the scale and hypothesized that the relations between the variables would be curvilinear. The best fit model of this stage of model building included ASD Diagnosis, mean F0, maximum F0 and its quadratic term, social acceptability and its quadratic term, and the percentage of children with ASD an SLP had on their caseload. When I refit this model using the adjusted dataset without recording 8 in it, mean F0, maximum F0 and its quadratic term ceased to be significant, so the final best-fit main-effects model included main effects of ASD Diagnosis, social acceptability and its quadratic term, and ASD percent.

After main effects modeling, I next investigated interactions. I began with the three-way interaction between SLP group, SLP's rating of children's ASD diagnosis (ASD Rating), and ASD Diagnosis that appeared in observed data explorations. Without the expert group included in the dataset anymore, the three-way interaction was not significant, $F(1, 739.94) = 0.36$, $p = 0.548$ according to the Wald test when I performed a class ANOVA on the model using the *"anova.merMod()"* function of the *"lme4"*

package in R 3.6.1 (Bates et al., 2015, p. 4). I subsequently tested the three, 2-way interactions associated with this three-way interaction using likelihood ratio tests between nested models. The fixed and random effects of most parsimonious model that significantly improved the main-effects only model is displayed in Table 3.

Table 3
Best Fit MLM Predicting Prosody Ratings

	Count	*b*	(SE)	*p*
Intercept		0.2090	0.2328	
Social Acceptability Rating – Linear		-0.375	0.0947	< 0.001
Social Acceptability Rating – Quadratic		0.0511	0.0133	< 0.001
% of Children with ASD on SLP Caseload		-0.0031	0.0015	< 0.05
True ASD Dx: No vs. Yes		1.1985	0.2486	< 0.001
Rating of ASD Dx: No vs. Yes		-0.7608	0.1256	< 0.001
Interaction: True ASD Dx vs. Rating of ASD Dx		1.3047	0.1598	< 0.001
Num. obs.	760			
Num. groups: SLP	40			
Num. groups: Recording	19			
Var.: SLP (Intercept)		0.0302		
Var: Recording (Intercept)		0.2572		
Var: Residual		0.6589		

This model contained a two-way interaction between ASD Rating and ASD Diagnosis, b = -1.30, SE = 0.16, χ^2 (10) = 63.70, $p < 0.001$. This interaction is graphically represented in Figure 18.

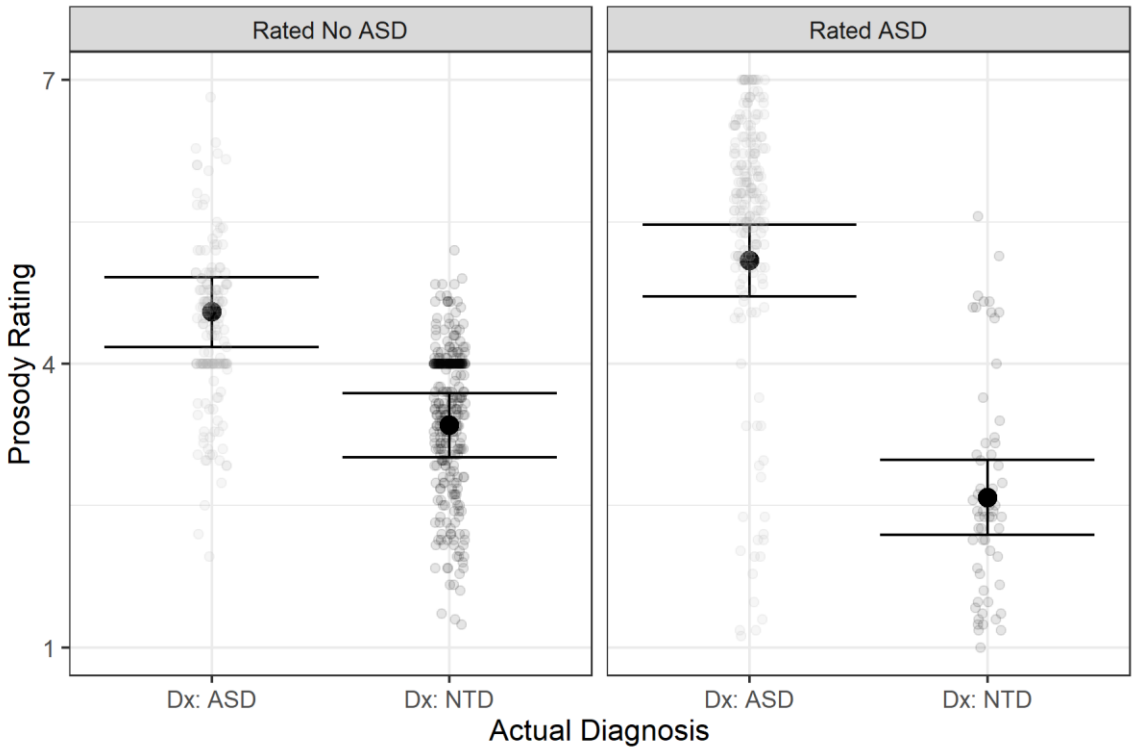

Figure 18. MLM interaction plot displaying two-way interaction between SLP ratings of ASD and the children's actual diagnosis from the best-fit model with all other predictors held constant at their means. Solid dots with 95% confidence interval error bars represent the model predictions. Light underlaid dots represent the raw SLP ratings.

As can be seen in the left panel, while holding all other variables constant at their means, the children who were rated as not have ASD, but who actually had ASD were rated as having more modulated prosody than those who were accurately rated as not having ASD. Looking at the right panel in the figure, the same pattern appeared, but more extremely so. The children who were rated as having ASD were rated as having more modulated prosody. The prosody of the children who were rated as not having ASD children was rated as less modulated.

Interestingly, when looking at the observed data (shown as light dots in the figure) in the left panel, there were clusters of ratings at a "4" for children who had ASD *and* children who had NTD. This same clustering of "typical" ratings of prosody did not appear in the right panel. The prosody ratings for children who were rated as having ASD seemed to cluster at the ends of the scale, regardless of whether they actually had ASD or not. So, even when the SLPs didn't know the child's diagnosis, they still rated children who they *thought* had ASD as having atypical prosody, whether overly modulated or under modulated. In general, the SLPs rated the children with ASD as having more modulated prosody than the children with NTD.

Social acceptability associations. To investigate the degree to which scores on this prosody rating scale were associated with ratings of social acceptability, I examined at what points on the prosody rating scale children were consistently rated as less socially acceptable (defined as a 5 or more on the social acceptability scale). I first built an MLM examining the association between the social acceptability ratings and fixed effect variables (see Figure 8). I used the same model building approach as with the prosody rating scale and the rating of presence or absence of ASD. The terms included in the best fit model can be seen in Table 4.

Table 4
Best Fit MLM Predicting Social Acceptability Ratings

	Count	*b*	(SE)	*p*
Intercept		2.4159	0.1638	< 0.001
True ASD Dx: No vs. Yes		0.4391	0.1377	< 0.01
Rating of ASD Dx: No vs. Yes		1.3167	0.1458	< 0.001
Prosody Rating – Linear		0.1068	0.0727	
Prosody Rating – Quadratic		0.3452	0.0436	< 0.001
Maximum F0 – Linear		0.3451	0.0833	< 0.001
Maximum F0 – Quadratic		0.1757	0.0648	< 0.01
SLPs' Survey Completion Time in Hours		0.0020	0.0006	< 0.001
% Children with ASD Level 2 on SLP Caseload		-0.0090	0.0038	< 0.05
Interaction: Rating of ASD Dx vs. Prosody Rating – Linear		-0.0267	0.0780	
Interaction: Rating of ASD Dx vs. Prosody Rating – Quadratic		-0.1484	0.0510	< 0.01
Num. obs.	760			
Num. groups: SLP	40			
Num. groups: Recording	19			
Var.: SLP (Intercept)		0.1846		
Var: Recording (Intercept)		0.0132		
Var: Residual		1.2262		

Using the results from the model in Table 4, I located the prosody ratings that were closest to the social acceptability ratings of 5. These "cut points" were at different values for children with ASD (1.35, 6.32) and children with NTD (1.07, 6.60) as can be seen in Figure 19. Children with ASD were rated as less socially acceptable at prosody ratings of 1.35 or less and 6.32 or more. Children with NTD were rated as less socially acceptable at prosody ratings of 1.07 or less and 6.60 or more. It would appear that the range of prosody presentation that adversely affected social acceptability was larger for children with ASD than for those with NTD.

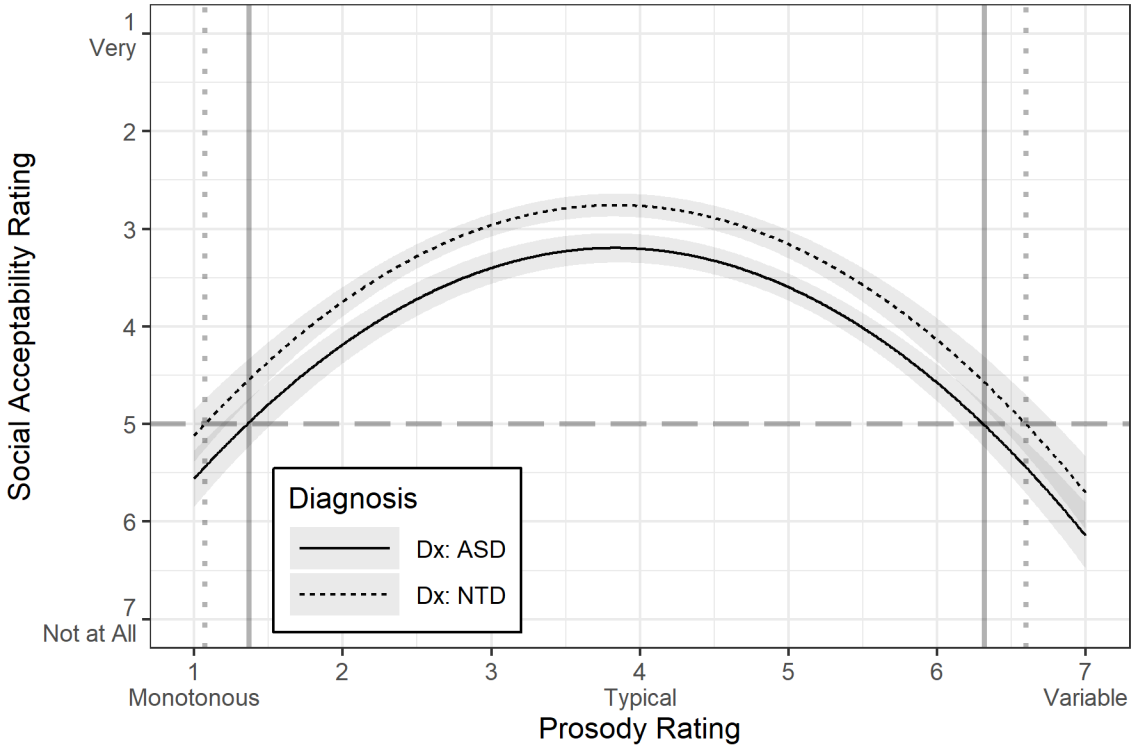

Figure 19. MLM "cut point" plot of SLPs' ratings of participants' social acceptability as a function of their prosody from the best-fit model faceted by diagnosis with all other predictors held constant at their means. Solid grey vertical lines represent the values on the prosody rating scale at which children with ASD were rated as less socially acceptable. Dotted grey vertical lines represent the values on the prosody rating scale at which children with NTD were rated as less socially acceptable. The dashed grey horizontal line represents the level of the social acceptability rating scale which was designated as the point where children were less socially acceptable than their peers.

Figure 20 shows the moderating effect of social acceptability on the interaction between ASD diagnosis and ASD rating in the best fit MLM of associations between prosody ratings and social acceptability ratings (for model parameters, see Table 3).

Children who were rated as moderately socially acceptable had lower prosody ratings than the children who were rated as either very socially acceptable or very socially unacceptable.

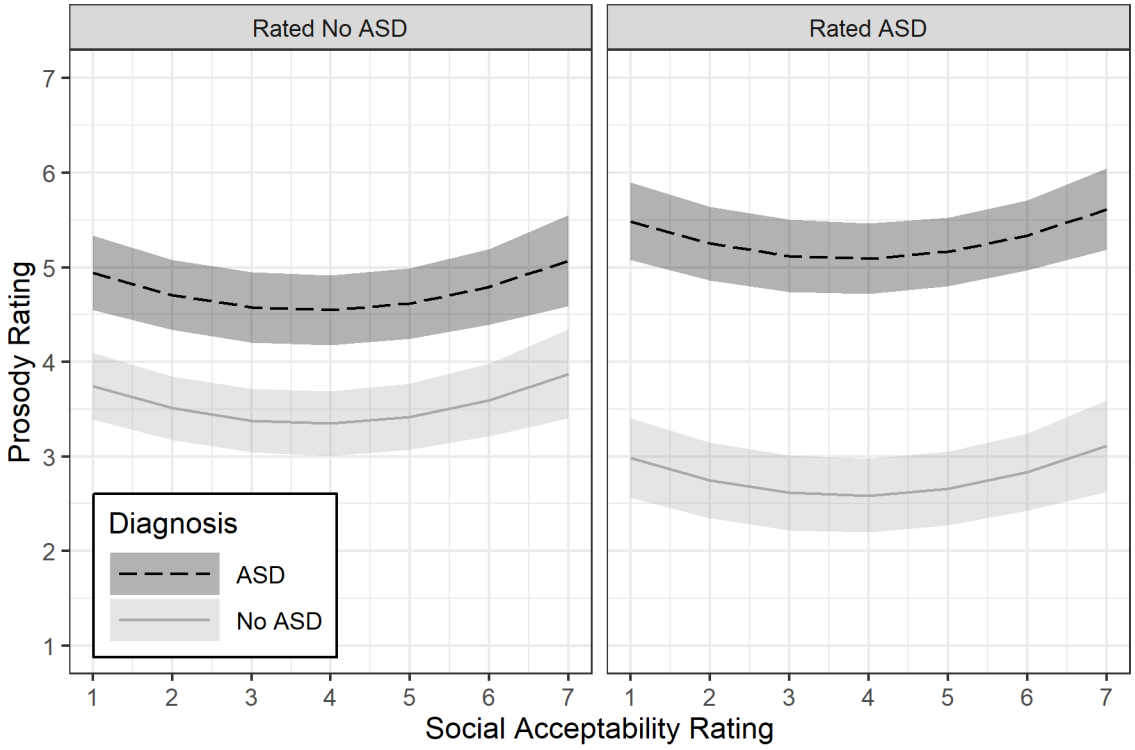

Figure 20. MLM interaction plot with moderating effect of SLPs' social acceptability rating of participants on their ratings of participants' prosody. Black dashed lines represent children with ASD. Grey solid lines represent children with NTD. Confidence bands represent 95% confidence intervals.

Presence or absence of ASD rating associations. To answer the question of what score(s) children had a 75% chance of being rated as having ASD, I examined at what points on the prosody rating scale children had a .75 probability of being rated as having ASD. I first built an MLM examining the association between the SLPs' binary rating of

whether the child had ASD or not and fixed effect variables. I used the same model building approach as with the prosody rating scale. The terms included in the best fit model can be seen in Table 5. I tested the interaction between social acceptability ratings and ASD Tx Years, which had been significant in models that were built on the dataset including recording 8. This interaction did not significantly improve the model using the smaller dataset, possibly because of lack of power. Using the results from the best fit model without interactions (see Table 5), I located the prosody ratings that were closest to .75 predicted probability. These "cut points" were different for children with ASD and those with NTD. Children with ASD had a .75 probability of being rated as having ASD at prosody ratings of 1.77 and 5.88. The prosody ratings at which children with NTD had a .75 probability of being rated as having ASD were 1.31 and 6.34. These "cut points" can be seen in Figure 21.

Table 5
Best Fit MLM Predicting ASD Ratings

	Count	*b*	(SE)	*p*
Intercept		-4.8721	0.6157	< 0.001
Social Acceptability Rating – Linear		1.3856	0.4475	< 0.01
Social Acceptability Rating – Quadratic		0.2299	0.1490	--
% of Children with ASD on SLP Caseload		0.6593	0.0872	< 0.001
True ASD Dx: No vs. Yes		0.7475	0.1067	< 0.001
Rating of ASD Dx: No vs. Yes		-0.0597	0.0248	< 0.05
Num. obs.	760			
Num. groups: SLP	40			
Num. groups: Recording	19			
Var.: SLP (Intercept)		1.0751		
Var: Recording (Intercept)		0.3065		

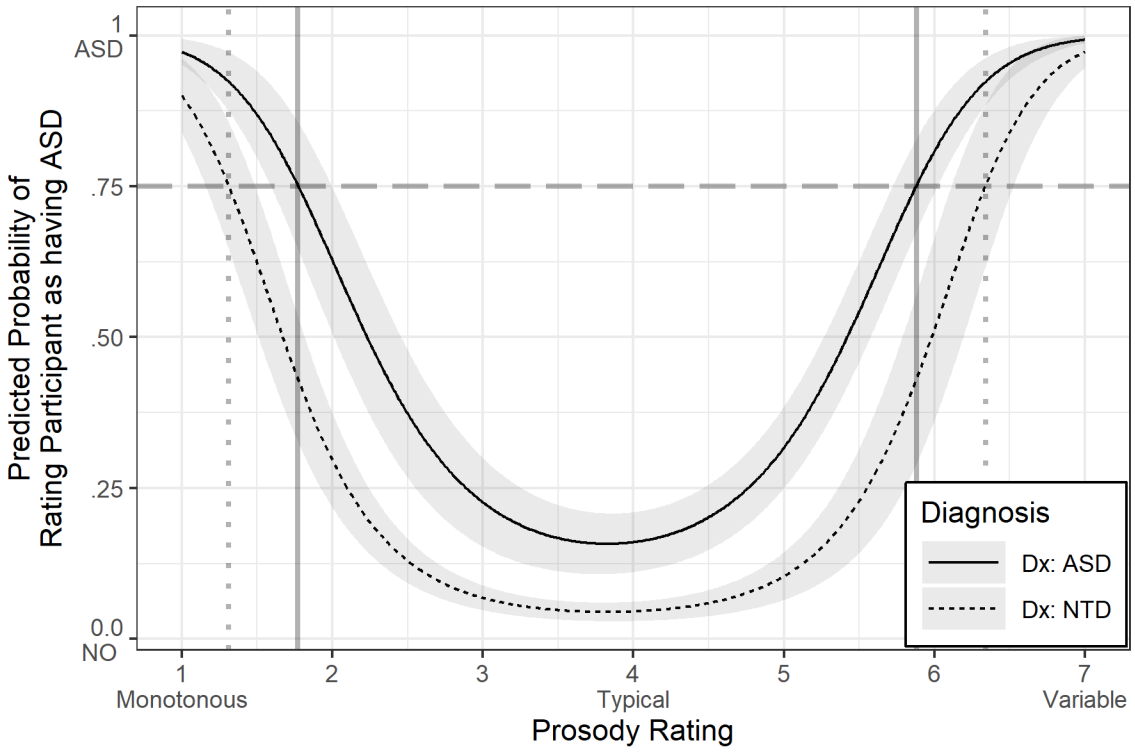

Figure 21. MLM plot of predicted probability of SLPs rating children as having ASD as a function of their prosody rating from the best-fit model with all other predictors held constant at their means. Dashed line indicates 75% probability of being rated as ASD. Solid grey vertical lines represent the values on the prosody rating scale at which children had a .75 probability of being rated ASD.

Consequences of testing. Evidence of validity based on the consequences of testing for different groups was first examined using observed data visualizations. Racial and ethnic representation are important to consider when investigating this area of validity (American Educational Research Association et al., 2014), so I examined these variables in my observed data. As can be seen in Figure 22, my data were not representative in these ways, so I was not able to include them in statistical modeling.

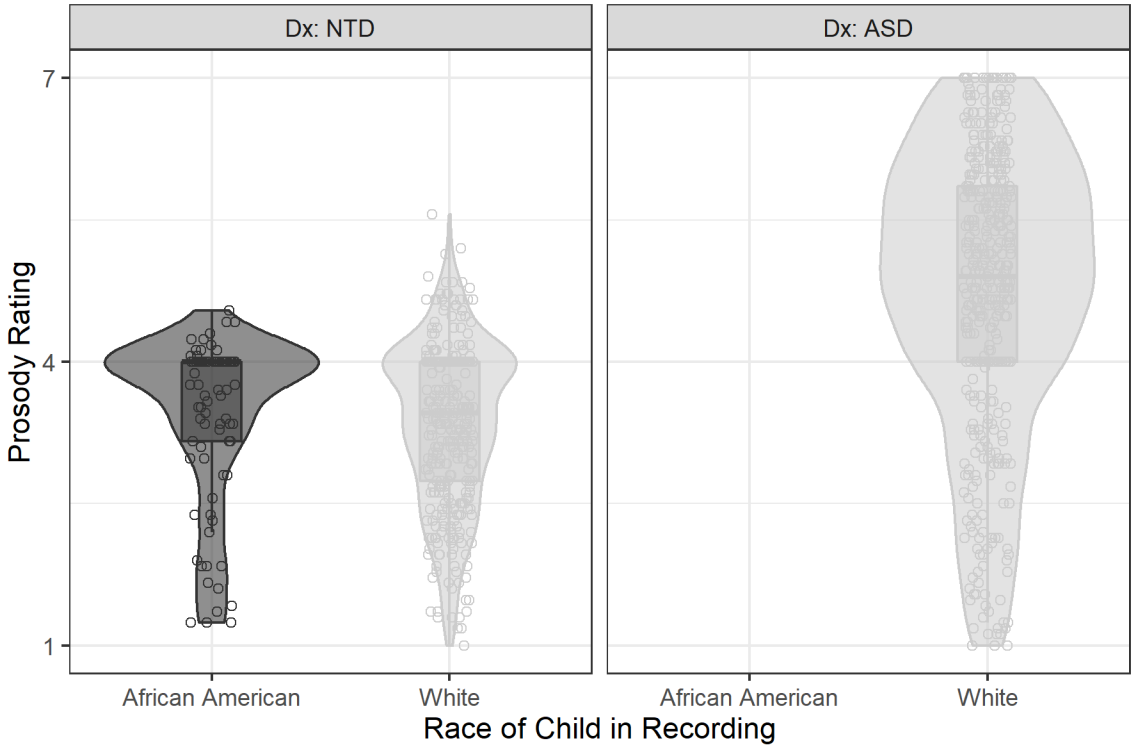

Figure 22. Observed data violin plots with box plots overlaid for prosody ratings according to children's race for children in the sample.

Child sex, however, was sufficiently balanced between the groups that I was able to investigate it statistically in my MLMs (see Figure 23).

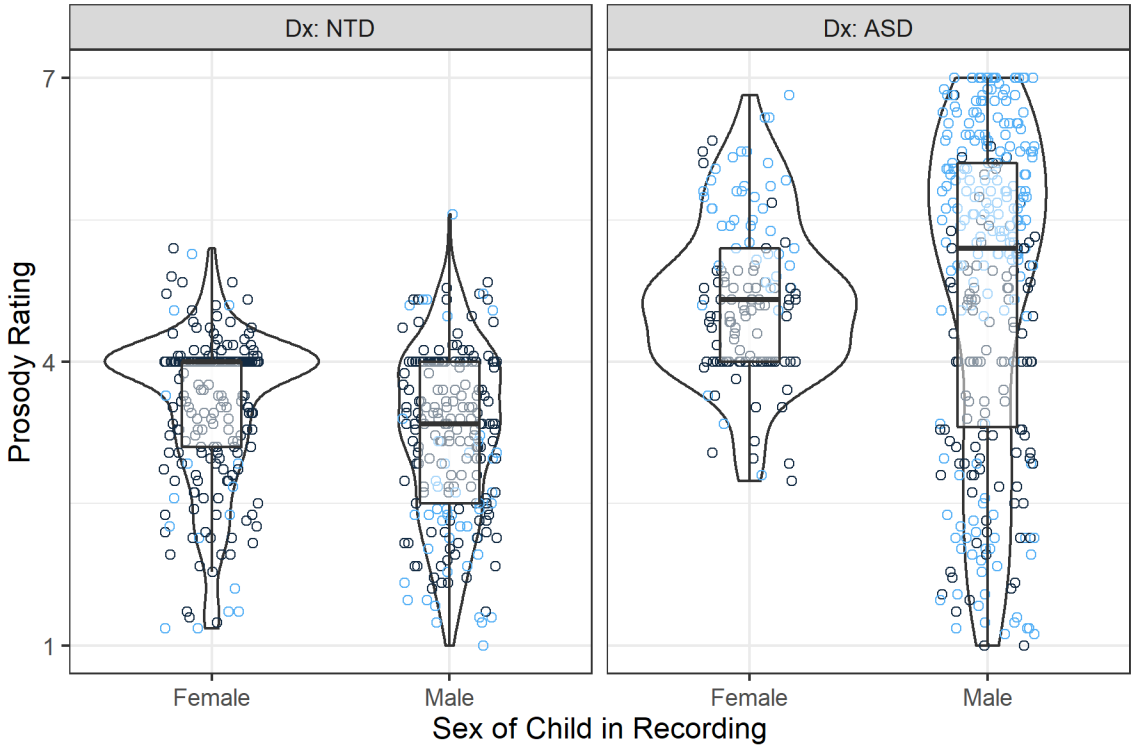

Figure 23. Observed data violin plots with box plots overlaid for the sex of children in the sample and prosody ratings, faceted by actual diagnosis. Blue and black dots represent the SLPs' ratings of presence or absence of ASD, respectively.

Results from observed data visualizations were further explored with MLMs. The process I followed built on the models previously described. After adding variables related to convergent and discriminant validity, I added variables related to participant characteristics, including children who provided audio and SLPs who rated the audio clips. I built models with SLP characteristics first, then investigated the effects of child characteristics. This part of the model-building was conducted before I trimmed Recording 8 from the dataset, but the final models were refitted using the adjusted dataset.

SLP characteristics. The previous best-fit model for the prosody rating scale prior to adding SLP characteristics included main effects for ASD Diagnosis, mean F0, maximum F0 and its quadratic term, and social acceptability ratings. Consequences of testing may be different for children evaluated by different SLPs, so I investigated the relations between several SLP characteristics and the prosody rating scale, namely (listed in the order they were built into the model): ASD Tx Years, the percent of the SLPs' clients who had ASD across their career, ASD Percent, the percent of clients on the SLPs' current caseloads with ASD who required very substantial support (ASD Percent Level 3), ASD Percent Level 2, the percent of clients on the SLPs' current caseloads who required support (ASD Percent Level 1), and completion time. The only SLP characteristic which significantly improved the model was ASD Percent, b = -0.003, SE = 0.002, χ^2 (11) = 3.94, p = .047. This remained true even after I refit the models with the dataset excluding recording 8 (See Table 3 and Figure 24). Those SLPs with more children with ASD on their caseload rated the prosody of children as less modulated, regardless of the children's diagnosis or whether they were rated as having ASD.

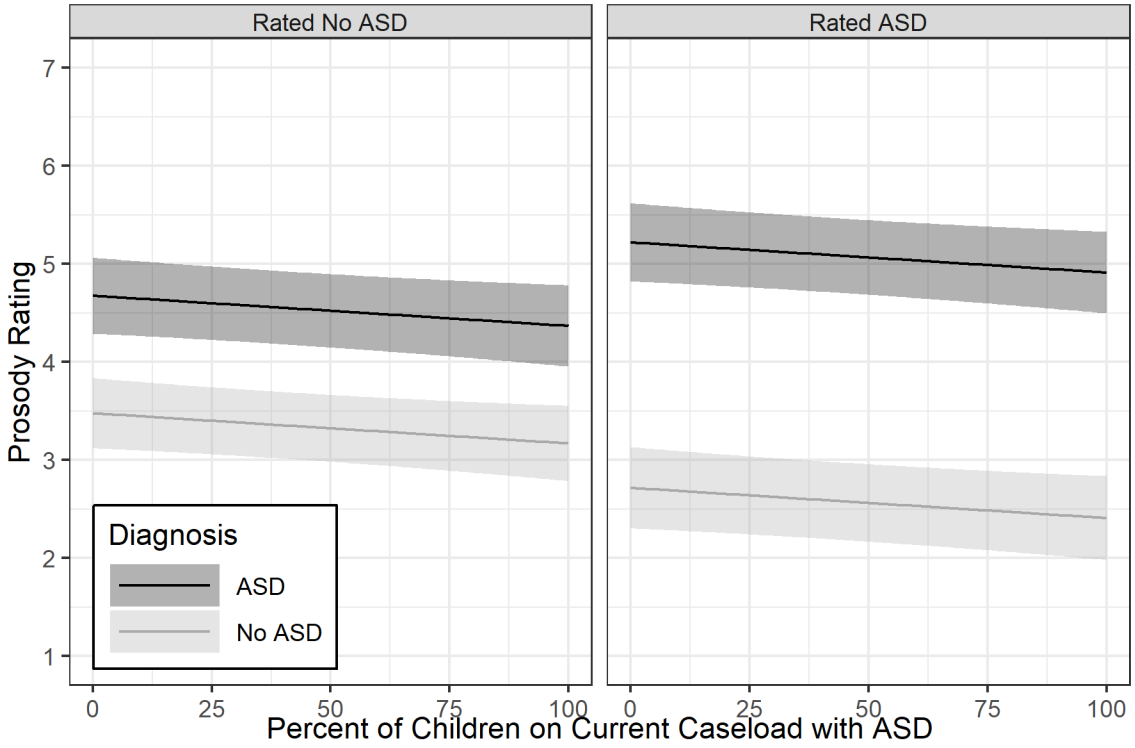

Figure 24. MLM interaction plot with moderating effect of percent of children with ASD on SLPs' current caseloads on ratings of participants' prosody. Black solid lines represent children with ASD. Grey solid lines represent children with NTD. Confidence bands represent 95% confidence intervals.

Child characteristics. Due to limitations in the existing sample data and resulting matching constraints, I was not able to fully assess the consequences of testing for those children with ASD of various racial and ethnic backgrounds. The two child characteristics I was able to investigate were ASD diagnosis, which was included in the models from the beginning for reasons described previously, and the sex of the children in the audio clips (Sex). I added Sex to the previous best fit model, but it did not significantly improve the model fit, b = 0.18, SE = 0.33, $\chi 2$ (12) = 0.31, p = .580.

CHAPTER 4

DISCUSSION

Overview

Speech prosody atypicalities have been noted in persons with ASD since its presentation as a clinical entity (Asperger, 1944, 1991; Kanner, 1943). Yet, despite many studies, including a systematic review and meta-analysis by Fusaroli and his colleagues (2017), no universal characterization of speech prosody in ASD has been identified. Evidence can be found to support characterizations of prosody in ASD as under modulated, not different from typical, or overly variable (Dahlgren et al., 2018; Filipe et al., 2014; Grossman et al., 2013; Irvine et al., 2016; Kissine & Geelhand, 2019; Nakai et al., 2014, 2017; Parish-Morris et al., 2017; Paul et al., 2005; Thurber & Tager-Flusberg, 1993; Wynn et al., 2018). For those persons whose speech prosody is atypical and interferes with daily functioning, valid, reliable, and efficient assessments of speech prosody are needed (Paul, Shriberg, et al., 2005). Currently, there are only three validated assessments for speech prosody specific to ASD and none of them are simultaneously valid, reliable, and efficient (de Villiers et al., 2007; S. Peppé & McCann, 2003; Shriberg et al., 1990, 1992).

In this study, I set out to validate a one-item, 7-point continuous analogue rating scale for screening the speech prosody of children with ASD. My intent was to create an easily accessible, valid, reliable, and efficient instrument for SLPs to include as part of an assessment battery for children with ASD. Additionally, I sought to establish “good” inter- and intra-rater reliability of the scale and to see if participants who engaged in a brief, online training would show higher levels of reliability (Koo & Li, 2016).

The rating scale ranged from 1-7 with anchor points at 1 (monotonous), 4 (typical), and 7 (overly modulated) to reflect the range of prosody presentation in children with ASD. I then selected twenty 30-second audio clips from short narrative discourse samples of children with ASD who participated in a narrative intervention study (Gillam et al., 2015) and age and sex matched peers with NTD who participated in the normative sample of the TNL (Gillam & Pearson, 2004, 2017). I selected 15 potential anchor and training clips that a group of three Expert SLPs rated to develop "gold standard" scores. These SLPs also selected clips to serve as the three anchors for the scale. Once the anchors were selected and "gold standard" ratings determined, I developed a brief web-based training for half of the SLPs to participate in. A total of 42 ASHA-certified SLPs who had children with ASD on their caseloads were recruited to rate the 20 audio clips used for validation purposes. They rated the clips at two time points, at least two-weeks apart. Half of the SLPs participated in the brief on-line training.

Results were explored then analyzed in R 3.6.1 (R Core Team, 2019). Observed data were plotted to examine any relevant relations. Then these explorations were confirmed using linear mixed-effects modeling. Model building followed a theory-based, research question driven approach. In all, some of my hypotheses were supported fully, others not at all, and some were inconclusive. A discussion of implications of these results follows along with an exploration of limitations of the study and possible future directions.

Reliability

One of the primary purposes of this investigation was to determine if this prosody

scale had adequate inter- and intra-rater reliability among and within SLPs so that it could be used as a screening instrument. A related purpose was to see if training improved reliability. I hypothesized that the overall inter- and intra-rater reliability would be moderate to good and that the reliability ICCs of the trained groups would be significantly higher than those who were not trained. These predictions mostly not supported by the data. Of the six reliability ICCs I calculated, only the intra-rater reliability of the trained group of SLPs met the criteria for "good" reliability, i.e. an ICC of .75 or more (Koo & Li, 2016). This was consistent with the results in Brinca et al. (2015). In all, these reliability findings suggested that, in its current form and with the current training, this prosody scale was not adequately reliable to be used as an initial screening tool for speech prosody in children with ASD. It is conceivable, however, that the scale might be used as a progress monitoring tool in its current form since the intra-rater reliability of SLPs who were trained was .76. Unfortunately, since this rating scale did not have sufficient interrater reliability to be used as a screening tool, the issues with initial assessment of prosody in ASD remain, namely the lack of fully validated, efficient tools.

Content Validity and Response Process Validity

Evidence that the constructs being measured are consistent with the intended constructs of a scale is critical in assessment development (American Educational Research Association et al., 2014). To ensure that the construct of speech prosody measured by this scale aligned with current conceptualizations of speech prosody and the most current research on the characteristics of speech prosody in ASD, I developed a synthesized working definition of speech prosody for use in this study. This definition

was based on multiple published definitions of speech prosody (Nicolosi et al., 2004; Shriberg et al., 1990, 1992; Szczepek Reed, 2011; ’t Hart et al., 1990; Tench, 1996). Additionally, I designed the scale to reflect the range of prosody demonstrated by persons with ASD according to the most current research (Dahlgren et al., 2018; Filipe et al., 2014; Fusaroli et al., 2017; Grossman et al., 2013; Irvine et al., 2016; Kissine & Geelhand, 2019; Nakai et al., 2014, 2017; Parish-Morris et al., 2017; Paul, Shriberg, et al., 2005; Thurber & Tager-Flusberg, 1993; Wynn et al., 2018). To verify that this alignment of content represented in the scale was successful, I analyzed the free response prosody rating rationales provided by SLPs. Results revealed that, in large part, SLPs gave rationales relevant to prosody, which suggested that the construct evaluated by the scale was consistent with the one that I originally intended, i.e. speech prosody.

Along similar lines as ensuring that the content evaluated in an assessment is valid, ensuring that raters engage in cognitive processes relevant to the task at hand is also crucial for evidence of a test’s validity (American Educational Research Association et al., 2014). Evidence to support the validity of the cognitive processes the SLPs engaged in while assessing speech prosody using this scale was also gathered through the analysis I performed to investigate the relevance of SLPs rating rationales to speech prosody. The vast majority of SLPs provided rationales were relevant to speech prosody, suggesting that SLPs were evaluating the children's prosody in relation to their previous conceptualizations of speech prosody. For instance, SLPs who put “WNL” or “within normal limits” evidenced by this short rationale that they, a) had a conception of “normal” prosody, b) compared this conception with the audio sample they were rating, c) determined that it matched their previous understanding of what “normal” speech

prosody would be, and d) gave a rating according to this evaluation. This process was closely aligned with what I expected SLPs' cognitive processes to be while rating. Further supporting the notion that the response processes of the SLPs was in line with what I expected was the fact that the time it took the SLPs to complete the survey did not have a significant association with their prosody rating scales and thus, it is possible that SLPs who completed more than half of the survey did not alter their conceptualizations of prosody nor their engagement with the task over time.

Relations to Other Variables

Discriminant validity: narrative proficiency. I included a measure of narrative proficiency in my statistical model evaluating the evidence for discriminant validity because I wanted to confirm that my scale measured a construct different from narrative proficiency even though all of the audio samples in the study were taken from narrative discourse. SLPs routinely evaluate children's narrative proficiency (Ukrainetz, 2015) and I reasoned that because speech prosody is overlaid on whatever speech it comes from (Shriberg et al., 1992), it may have been possible for the SLPs to unintentionally let their evaluation of the children's narratives color their assessment of the children's speech prosody. This did not appear to be the case, however. Narrative proficiency was not significantly associated with the prosody ratings, which suggested that the measure of narrative proficiency and the prosody rating scale evaluated two separate constructs.

Convergent validity: F0 and speech prosody constructs. Of particular interest in the validation of this scale were the relations between F0 measures and the prosody ratings scale. My original formulation of the scale was as an "intonation" rating scale, which is very closely tied to the construct of pitch ('t Hart et al., 1990; Tench, 1996;

Wichmann, 2000) and although the scale in its current form aimed to capture a wider definition of speech prosody than just what was captured by pitch, I hypothesized the constructs to still be very closely tied. Interestingly, I did not find this to be the case in the final models. None of the F0 measures were significantly associated with the prosody rating scale in the end. This could be due to a number of factors. Perhaps the scale did not measure a construct as closely related to speech prosody as I thought. Perhaps the scale was not calibrated appropriately as the observed data for ratings of the children with NTD seemed to imply. Importantly, though, the narrative proficiency scores of the children were not significantly associated with the prosody ratings, so the ratings were not a proxy for SLPs' impressions of the children's ability to tell a good story. It's possible that SLPs were influenced by factors such as grammar, vocabulary, and articulatory performance, although based on the majority of their reported rationales for their prosody ratings, if they were influenced by these factors, it was not purposefully so. Prior to rating the audio clips, the SLPs were instructed to focus only on the children's prosody and ignore other factors like vocabulary and articulation. Anecdotally, however, some SLPs commented on how difficult it was to focus only on the prosody of the children, so it is possible that part of the reason F0 measures were not significantly associated with prosody ratings was because SLPs were subconsciously rating features other than prosody. However, because I did not explicitly measure variables related to grammar, vocabulary, and articulation, the association of these variables with SLPs' prosody ratings remains unknown.

These potential influences notwithstanding, it is possible that the reason for the lack of association between F0 and the prosody rating scale was simply reduced sample size. Before I removed Recording 8, mean F0 and maximum F0 were both significantly

associated with the prosody rating scale scores. After I removed Recording 8, these variables were no longer significant. Thus, it seemed likely that at least a portion of the reason for lack of associations among these variables was the reduction in power to detect an effect resulting from the removal of Recording 8.

Convergent validity: social acceptability and ratings of ASD. While I did not have exact hypothesized numbers on the rating scale where I though children would be more likely to be perceived as less socially acceptable or more likely to be rated as having ASD, I did anticipate that there would be two distinct numbers on the prosody rating scale, one at the lower end and one at the higher end for both scales. These hypotheses were confirmed, but with modifications. The scores at which children were rated as less socially acceptable were different for children with ASD and children with NTD, even though SLPs were blind to which children had ASD and which had NTD. The same was true to an even greater extent for the ratings of presence or absence of ASD; children with ASD had different "cut points" at which they had a 75% chance of being rated as having ASD. The direction in which the "cut points" were different was the same between the two scales. For both the ratings of presence/absence of ASD and the social acceptability ratings, the "cut points" were more restrictive for the children with ASD than the children with NTD. For example, the "cut points" of the ratings of presence or absence of ASD were at the prosody ratings of 1.77 and 5.88, while the numbers for children with NTD were significantly more generous at 1.31 and 6.34 (see Figure 21). This suggested that SLPs heard *something* different between the two groups. Like the machine learning studies Fusaroli and his colleagues (2017) reported on, it would seem that the acoustic signal of the speech of children with ASD communicated a difference to

the SLPs that set them apart from their peers with NTD. Yet, as Fusaroli and his colleagues concluded, we simply don't have a clear idea of *what* that difference might be.

This conclusion was underscored by the fact that, even though SLPs heard differences between the groups, those differences did not translate into accurate ratings of the "true" diagnosis of the children. The SLPs only had audio information to make judgements, and as it turned out, they were not very accurate in predicting which children had ASD and which did not. This included the "expert" SLPs. While these SLPs were more accurate than the Rating SLPs, they still did not consistently accurately rate which children had ASD and which didn't. It could be tempting to attribute this lack of accuracy to the individual SLPs' lack of knowledge or preparation or to the fact that SLPs do not routinely diagnose ASD; however, recall that all of the SLPs in this sample had some, if not extensive experience in working with children with ASD. One of the inclusionary criteria for the study was that the SLPs had to be working with children with ASD on their current caseload. This fact, along with the fact that all of the participating SLPs had their Certificate of Clinical Competence from ASHA, suggested that the task of identifying the presence/absence of ASD based solely on speech prosody may be difficult, if not impossible, even for highly trained individuals. This could have been because of the variability of speech prosody characteristics in ASD or some other variable; it was not clear at this point. Testing for that reason of SLPs' inaccuracy was beyond the scope of this study, so the true reason remains unknown.

Consequences of Testing: Scale Representativeness

There was no association of Sex with the prosody rating scale based on my analyses. While there is some evidence to support that males and females exhibit

differences in prosodic characteristics at the ages of the children in my audio sample (Berger et al., 2019; Guzman et al., 2014), sex differences did not play a significant role in SLPs' prosody ratings in this study. This strengthens the case for the evidence of validity related to consequences of testing for my scale because it suggests that, were it sufficiently reliable, this scale could be used effectively with both male and female children with ASD. This of particular importance because many researchers suggest there are more females with ASD than prevalence estimates would suggest (Halladay et al., 2015; Parish-Morris et al., 2017; Van Wijngaarden-Cremers et al., 2014) and thus may be underserved. The underrepresentation of females in ASD estimates may in part be related to learned prosodic patterns that may serve to camouflage their ASD (Gamer et al., 2019; Parish-Morris et al., 2017), so an assessment that could be used equally well for males and females with ASD would be particularly useful.

Limitations and Future Directions

While the results of this validation study were promising, several limitations prevent the widespread use of this tool as it currently stands, leaving room for future development of the tool.

Audio clip quality. One of the primary limitations of to this study was the lack of control over the quality of the audio clips that the SLPs rated. While all of the clips were normalized to the same dB level, there were still variations in the perceived sound levels of the recordings according to the SLPs. Similarly, I clipped as much background noise out of the clips as possible, but I was unable to remove all of the background noise and fully account for microphone quality. Because the ratings using the scale depended on the audio samples, the validity of the scale was not as robust because of the lower quality

audio.

Recall from the Methods section that in order to avoid stereotyped utterances associated with the beginnings and endings of narratives and unnatural prosody resulting from dialogue, any instances of these types of speech were removed from the audio clips. Avoiding stereotyped and unnatural utterances created problems in itself, however, the main one of which was a loss of naturalness. Because I wanted to alter the audio samples as little as possible, when I removed portions of the clips, I did not remove any pause time either before or after the clipped segment. Consequently, often two short pauses were combined into one longer pause, which affected the SLPs' ratings of the clips. If the clipping did not result in a long pause, the remaining speech was sometimes stilted and unnatural, which also affected SLPs' ratings. An extreme, yet illustrative example of this was Recording 8 from my sample. The participant in this clip employed long pauses in his narrative even without any alteration to the clip, presumably to allow formulation and organization time (Hallin et al., 2016; Irvine et al., 2016; Thurber & Tager-Flusberg, 1993). Most of the speaking portion of his narrative was dialogue, so when I clipped the dialogue from his story, his long pauses combined into even longer pauses. The resulting 30 second clip had very little speech and the SLPs commented on the difficulty of rating this sample. Representative comments include: "very flat sample, but also not a lot of speech," "not a good sample to judge from," and "really not sure on this one - not enough of a speaking sample." While the problems caused by removing beginning and ending speech as well as dialogue were extreme in this case (I ultimately excluded the ratings of this clip from the final analysis), the same issues were present to a lesser extent in narratives from the other children. So, while this iteration of the validation of the scale

provided promise for its use, it would need to be validated on unmodified 30 second clips before it could be widely implemented.

A related issue was that this set of audio samples was taken from a narrative context only. Because this is a context that may be commonly available to SLPs as it is already part of their assessment battery, to gain a more complete picture of a students' prosody it may be necessary to collect a non-narrative connected speech sample or a conversational speech sample similar to the PVSP (McSweeny & Shriberg, 2001; Shriberg et al., 1990, 1992). This would also solve the problem of needing to remove sections of the audio file that contained dialogue and narrative beginnings and endings.

Sample representativeness and statistical power. One limitation of this study was the fact that the children in the audio samples were not very racially diverse. While this was a byproduct of using the data readily available, the validation of this rating scale would be strengthened by including more racially diverse children in the audio clips. Similarly, because I collected a convenience sample of SLPs to rate the audio clips, certain races and ethnicities were not represented in the SLP sample. While the representation of different genders and, to some extent, races in my sample of SLPs was fairly consistent with the overall demographic profile of SLPs in the United States (American Speech-Language-Hearing Association, 2019), including a more diverse sample of SLPs would reduce the threats to validity related to the consequences of testing.

As noted previously, to ensure adequate evidence for validity in response processes, I excluded recording 8 and its ratings from analysis. This reduced my statistical power and therefore my ability to detect relations between some of the

variables. The fact that mean F0 and maximum F0 were significantly associated with the prosody rating scale before recording 8 was removed, but ceased to be significantly associated afterward led me to believe that had I had more power in the final models, I would have been able to detect associations between these F0 variables and the prosody ratings scale as I had hypothesized.

Scale characteristics. I originally constructed the scale with three anchor points, one at each end of the scale and one in the middle; however, it is possible that reliability would have improved if I had given more anchors in the middle of the scale. This is due to evidence suggesting that ratings are less stable in the middle of scales, so providing more anchors in the mid points of the scale (e.g. at 2.5 and 5.5) may have improved reliability measures (Kreiman & Gerratt, 1998).

It is also possible that increasing the number and altering the placement of the example audio clips may have improved reliability. Audio exemplars representative of the anchor points of the scale were available in the section immediately prior to each audio clip for referencing and calibration. Had the anchor clips been embedded into the rating scale itself, SLPs would have been able to compare the anchors to the clip they were rating for more immediate referencing. This may have contributed to a firmer representation of each level of the scale. Similarly, one audio example was available for each anchor point of the scale. If participants had had access to more than one example of each anchor point of the scale, they may have been able to create a more solid mental representation of each level of the scale (Kreiman et al., 1993).

My results suggested that perhaps my scale was not calibrated correctly. The majority of children with NTD were rated around a 3.5 on the scale (see Figure 8) instead

of a 4, which I set as the "typical" score for the scale. If most of the kids with NTD were rated at around 3.5, then it seemed like my scale was not aligning with children's actual performance, although when the children with ASD were included, the mean prosody rating scale score was 4. Therefore, before altering the scale, I would want to test it in more discourse contexts. It is not unlikely that children telling a story during a standardized test might have reduced inflection due to lack of interest in the task. These same children may be quite animated in other speaking contexts, which underscores what Furr (2011) said about validating single-item scales over the course of multiple administrations. Further validation with different contexts will be needed in order to establish the "true" score on the scale that represents "typical" prosody.

Training. I included a training component to the validation of this scale because evidence suggests that training is needed for acceptable levels of reliability to be reached (Barsties et al., 2017; Eadie & Baylor, 2006; Fay & Latham, 1982; Støre-Valen et al., 2015). My original intention was for the training to be brief and easily implemented, which is why I created a short, web-based training. It was encouraging that intra-rater reliability for the trained group was good (ICC = .76) and this finding is consistent with Brinca and colleagues' (2015) finding that suggested that the inclusion of speech-sample anchors improves reliability. Nevertheless, it appears that the training as originally designed was not sufficient to achieve good levels of interrater reliability. As currently presented the training group received what Kreiman et al. (1993) characterized as "orientation" level training. This level of training involves basic definitions of scale levels, examples at anchor points on the scale and limited practice opportunities (less than 20). It is likely that if I increased the number of practice opportunities in my training

protocol and made the anchor examples available during the rating process (i.e. on the same REDCap page instead of the page previous to the clip to be rated), reliability would increase (Melchers et al., 2011). According to Gerratt and colleagues (1993), the most stable ratings are achieved when the levels of a ratings scale are anchored with an auditory exemplar. The vocal quality rating scale they tested had five anchor points, so it's could be that if I added at least two more anchors to my scale, the ratings would become more stable in a similar manner.

Conclusion

The initial validation of this rating scale resulted in a limited but encouraging proof of concept for this scale. Evidence of validity in the five areas recommended in American Educational Research Association et al. (2014) showed some promise, particularly in the area of evidence of validity based on relations to other variables. While the scale did not reach the levels of interrater reliability I had initially projected, interrater reliability was still moderate for both groups. Intra-rater reliability was moderate for the untrained group and good for the trained group, suggesting that training might improve the reliability of scale scores, especially if improvements were made to the current training. In all, this screening instrument is not ready for use in its current form, but given improvements in audio clip quality, representativeness of both children in the audio clips and SLPs, and improved training of SLPs, it has the potential to improve efficiency of speech prosody assessment for children with ASD.

REFERENCES

Adams, L. (1992). *Analysis of prosody in the speech of high-functioning autistic children.* (Doctoral dissertation). Retrieved from ProQuest. http://search.proquest.com.dist.lib.usu.edu/pqdtglobal/docview/304026396/abstract/44E8571DBF5E4B9FPQ/5

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Speech-Language-Hearing Association. (2019). A demographic snapshot of SLPs. *The ASHA Leader*, *24*(7), 32. https://doi.org/10.1044/leader.AAG.24072019.32

Asperger, H. (1944, 1991). 'Autistic psychopathy' in childhood. In U. Frith & Frith, Uta (Trans.), *Autism and Asperger Syndrome* (pp. 37–92). Cambridge, MA: Cambridge University Press. doi:10.1017/CBO9780511526770.002

Azizi, Z., Sharifi, S., & Nourbakhsh, M. (2016). The Tilt Model acoustic survey of intonation in children with severe autism. *International Journal of English Linguistics*, *6*(4), 78. https://doi.org/10.5539/ijel.v6n4p78

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Barsties, B., Beers, M., Ten Cate, L., Van Ballegooijen, K., Braam, L., De Groot, M., Van Der Kant, M., Kruitwagen, C., & Maryn, Y. (2017). The effect of visual

feedback and training in auditory-perceptual judgment of voice quality. *Logopedics Phoniatrics Vocology, 42*(1), 1–8. https://doi.org/10.3109/14015439.2015.1091036

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bell, J. F. (2002, July). *On the presentation of the results of multilevel analysis*. Paper presented at the 23rd Biennial Conference of the Society for Multivariate Analysis in the Behavioural Sciences, Tilburg, The Netherlands. Retrieved from https://www.cambridgeassessment.org.uk/images/109691-on-the-presentation-of-the-results-of-multilevel-analysis.pdf

Berger, T., Peschel, T., Vogel, M., Pietzner, D., Poulain, T., Jurkutat, A., Meuret, S., Engel, C., Kiess, W., & Fuchs, M. (2019). Speaking voice in children and adolescents: Normative data and associations with BMI, Tanner stage, and singing activity. *Journal of Voice*, *33*(4), 580.e21-580.e30. https://doi.org/10.1016/j.jvoice.2018.01.006

Boersma, P., & Weenink, D. (2013). *Praat: Doing phonetics by Computer* (Version 6.0.30) [Computer software]. http://www.fon.hum.uva.nl/praat/

Bone, D., Black, M. P., Ramakrishna, A., Grossman, R. B., & Narayanan, S. S. (2015). Acoustic-prosodic correlates of `awkward' prosody in story retellings from adolescents with autism. *Interspeech 2015*, 1616–1620. https://pdfs.semanticscholar.org/d0a6/3db67024f39d633865bddf5cb5f198f3bbdd.pdf

Bracken, B., & McCallum, R. S. (1998). *Universal nonverbal intelligence test*. The Riverside Publishing Co.

Brinca, L., Batista, A. P., Tavares, A. I., Pinto, P. N., & Araújo, L. (2015). The effect of anchors and training on the reliability of voice quality ratings for different types of speech stimuli. *Journal of Voice*, *29*(6), 776.e7-776.e14. https://doi.org/10.1016/j.jvoice.2015.01.007

Dahlgren, S., Sandberg, A. D., Strömbergsson, S., Wenhov, L., Råstam, M., & Nettelbladt, U. (2018). Prosodic traits in speech produced by children with autism spectrum disorders – Perceptual and acoustic measurements. *Autism & Developmental Language Impairments*, *3*, 239694151876452. https://doi.org/10.1177/2396941518764527

de Marchena, A., & Miller, J. (2017). "Frank" presentations as a novel research construct and element of diagnostic decision-making in autism spectrum disorder: Frank ASD. *Autism Research*, *10*(4), 653–662. https://doi.org/10.1002/aur.1706

de Villiers, J., Fine, J., Ginsberg, G., Vaccarella, L., & Szatmari, P. (2007). Brief report: A scale for rating conversational impairment in autism spectrum disorder. *Journal of Autism and Developmental Disorders*, *37*(7), 1375–1380. https://doi.org/10.1007/s10803-006-0264-1

Diehl, J. J., & Paul, R. (2009). The assessment and treatment of prosodic disorders and neurological theories of prosody. *International Journal of Speech-Language Pathology*, *11*(4), 287–292. https://doi.org/10.1080/17549500902971887

Diehl, J. J., Watson, D., Bennetto, L., Mcdonough, J., & Gunlogson, C. (2009). An acoustic analysis of prosody in high-functioning autism. *Applied Psycholinguistics*, *30*(03), 385. https://doi.org/10.1017/S0142716409090201

Diehl, J. J., Friedberg, C., Paul, R., & Snedeker, J. (2015). The use of prosody during syntactic processing in children and adolescents with autism spectrum disorders. *Development and Psychopathology*, *27*(3), 867–884. https://doi.org/10.1017/S0954579414000741

Diehl, J. J., & Paul, R. (2013). Acoustic and perceptual measurements of prosody production on the profiling elements of prosodic systems in children by children with autism spectrum disorders. *Applied Psycholinguistics*, *34*(01), 135–161. https://doi.org/10.1017/S0142716411000646

Eadie, T. L., & Baylor, C. R. (2006). The effect of perceptual training on inexperienced listeners' judgments of dysphonic voice. *Journal of Voice*, *20*(4), 527–544. https://doi.org/10.1016/j.jvoice.2005.08.007

Eadie, T. L., & Kapsner-Smith, M. (2011). The effect of listener experience and anchors on judgments of dysphonia. *Journal of Speech, Language, and Hearing Research*, *54*(2), 430–447. https://doi.org/10.1044/1092-4388(2010/09-0205)

Edgerton, L., & Wine, B. (2017). Speak up: Increasing conversational volume in a child with autism spectrum disorder. *Behavior Analysis in Practice*, *10*(4), 407–410. https://doi.org/10.1007/s40617-016-0168-2

Fay, C. H., & Latham, G. P. (1982). Effects of training and rating scales on rating errors. *Personnel Psychology*, *35*(1), 105–116. https://doi.org/10.1111/j.1744-6570.1982.tb02188.x

Filipe, M. G., Frota, S., Castro, S. L., & Vicente, S. G. (2014). Atypical prosody in Asperger syndrome: Perceptual and acoustic measurements. *Journal of Autism and Developmental Disorders*, *44*(8), 1972–1981. https://doi.org/10.1007/s10803-014-2073-2

Finstad, K. (2010). Response interpolation and scale sensitivity: Evidence against 5-point scales. *Journal of Usability Studies, 5*(3), 104-110. http://uxpajournal.org/wp-content/uploads/sites/8/pdf/JUS_Finstad_May_2010.pdf

Furr, R. M. (2011). *Scale construction and psychometrics for social and personality psychology*. London: SAGE.

Furr, R. M. (2018). *Psychometrics: An introduction* (3rd ed.). London: SAGE.

Fusaroli, R., Lambrechts, A., Bang, D., Bowler, D. M., & Gaigg, S. B. (2017). Is voice a marker for autism spectrum disorder? A systematic review and meta-analysis. *Autism Research*, *10*(3), 384–407. https://doi.org/10.1002/aur.1678

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2019). *Irr package documentation*. CRAN. https://www.r-project.org

Gerratt, B. R., Kreiman, J., Antonanzas-Barroso, N., & Berke, G. S. (1993). Comparing internal and external standards in voice quality judgments. *Journal of Speech, Language, and Hearing Research*, *36*(1), 14–20. https://doi.org/10.1044/jshr.3601.14

Gillam, R. B., & Pearson, N. A. (2004). TNL: Test of Narrative Language. *Austin, TX: Pro-Ed*.

Gillam, R. B., & Pearson, N. A. (2017). Test of Narrative Language, Second Edition. *Austin, TX: Pro-Ed.*

Gillam, S., Hartzheim, D., Studenka, B., Simonsmeier, V., & Gillam, R. (2015). Narrative Intervention for Children With Autism Spectrum Disorder (ASD). *Journal of Speech, Language & Hearing Research*, *58*(3), 920–933. https://doi.org/10.1044/2015_JSLHR-L-14-0295

Gillon, G., Hyter, Y., Fernandes, F. D., Ferman, S., Hus, Y., Petinou, K., Segal, O., Tumanova, T., Vogindroukas, I., Westby, C., & Westerveld, M. (2017). International survey of speech-language pathologists' practices in working with children with autism spectrum disorder. *Folia Phoniatrica et Logopaedica*, *69*(1–2), 8–19. https://doi.org/10.1159/000479063

Gordon, J. K., Andersen, K., Perez, G., & Finnegan, E. (2019). How old do you think I am? Speech-language predictors of perceived age and communicative competence. *Journal of Speech, Language, and Hearing Research*, 1–18. https://doi.org/10.1044/2019_JSLHR-L-19-0025

Grant, S., Aitchison, T., Henderson, E., Christie, J., Zare, S., McMurray, J., & Dargie, H. (1999). A comparison of the reproducibility and the sensitivity to change of visual analogue scales, Borg scales, and Likert scales in normal subjects during submaximal exercise. *Chest*, *116*(5), 1208–1217. https://doi.org/10.1378/chest.116.5.1208

Grossman, R. B., Edelson, L. R., & Tager-Flusberg, H. (2013). Emotional facial and vocal expressions during story retelling by children and adolescents with high-functioning autism. *Journal of Speech, Language, and Hearing Research*, *56*(3), 1035–1044. https://doi.org/10.1044/1092-4388(2012/12-0067)

Guzman, M., Muñoz, D., Vivero, M., Marín, N., Ramírez, M., Rivera, M. T., Vidal, C., Gerhard, J., & González, C. (2014). Acoustic markers to differentiate gender in prepubescent children's speaking and singing voice. *International Journal of Pediatric Otorhinolaryngology*, *78*(10), 1592–1598. https://doi.org/10.1016/j.ijporl.2014.06.030

Halladay, A. K., Bishop, S., Constantino, J. N., Daniels, A. M., Koenig, K., Palmer, K., Messinger, D., Pelphrey, K., Sanders, S. J., Singer, A. T., Taylor, J. L., & Szatmari, P. (2015). Sex and gender differences in autism spectrum disorder: Summarizing evidence gaps and identifying emerging areas of priority. *Molecular Autism*, *6*(1). https://doi.org/10.1186/s13229-015-0019-y

Hallin, A. E., Garcia, G. D., & Reuterskiöld, C. (2016). The use of causal language and filled pauses in children with and without autism. *Child Development Research*, *2016*, 11. https://doi.org/10.1155/2016/8535868

Harris, P., Taylor, R., Minor, B., Elliot, V., Fernandez, M., O'Neal, L., McLeod, L., Delacqua, G., Kirby, J., Duda, S., & REDCap Consortium. (2019). The REDCap consortium: Building an international community of software partners. *Journal of Biomedical Informatics*. https://doi.org/doi: 10.1016/j.jbi.2019.103208

Harris, P., Taylor, R., Thielke, R., Payne, J., & Gonzalez, J. (2009). Research electronic data capture (REDCap) – A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics*, *42*(2), 377–381. https://doi.org/10.1016/j.jbi.2008.08.010

Hosokawa, K., Barsties, B., Iwahashi, T., Iwahashi, M., Kato, C., Iwaki, S., Sasai, H., Miyauchi, A., Matsushiro, N., Inohara, H., Ogawa, M., & Maryn, Y. (2017).

Validation of the Acoustic Voice Quality Index in the Japanese language. *Journal of Voice*, *31*(2), 260.e1-260.e9. https://doi.org/10.1016/j.jvoice.2016.05.010

Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2010). *Multilevel analysis: Techniques and applications* (Second Edition). Routledge.

Irvine, C. A., Eigsti, I.-M., & Fein, D. A. (2016). Uh, Um, and Autism: Filler disfluencies as pragmatic markers in adolescents with optimal outcomes from autism spectrum disorder. *Journal of Autism and Developmental Disorders*, *46*(3), 1061–1070. https://doi.org/10.1007/s10803-015-2651-y

Kanner, L. (1943). Autistic disturbances of affective contact. *Nervous Child*, *2*(3), 217–250.

Kargas, N., López, B., Morris, P., & Reddy, V. (2016). Relations among detection of syllable stress, speech abnormalities, and communicative ability in adults with autism spectrum disorders. *Journal of Speech, Language & Hearing Research*, *59*(2), 206–215. https://doi.org/10.1044/2015_JSLHR-S-14-0237

Kissine, M., & Geelhand, P. (2019). Brief report: Acoustic evidence for increased articulatory stability in the speech of adults with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, *49*, 2572–2580. https://doi.org/10.1007/s10803-019-03905-5

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kreft, I. (1996). *Are multilevel techniques necessary?: An overview, including simulation studies*.

Kreiman, J., & Gerratt, B. R. (1998). Validity of rating scale measures of voice quality. *The Journal of the Acoustical Society of America, 104*(3), 1598–1608. https://doi.org/10.1121/1.424372

Kreiman, J., Gerratt, B. R., Kempster, G. B., Erman, A., & Berke, G. S. (1993). Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research. *Journal of Speech, Language, and Hearing Research*, *36*(1), 21–40. https://doi.org/10.1044/jshr.3601.21

Lee, A., Whitehill, T. L., & Ciocca, V. (2009). Effect of listener training on perceptual judgement of hypernasality. *Clinical Linguistics & Phonetics*, *23*(5), 319–334. https://doi.org/10.1080/02699200802688596

Loomes, R., Hull, L., & Mandy, W. P. L. (2017). What is the male-to-female ratio in autism spectrum disorder? A systematic review and meta-analysis. *Journal of the American Academy of Child & Adolescent Psychiatry*, *56*(6), 466–474. https://doi.org/10.1016/j.jaac.2017.03.013

Lord, C., Rutter, M., DiLavore, P. C., Risi, S., Gotham, K., & Bishop, S. (2012). *Autism diagnostic observation schedule - Second edition*. Torrance, CA.

McCann, J., & Peppé, S. (2003). Prosody in autism spectrum disorders: A critical review. *International Journal of Language & Communication Disorders*, *38*(4), 325–350. https://doi.org/10.1080/1368282031000154204

McSweeny, J. L., & Shriberg, L. D. (2001). Clinical research with the prosody-voice screening profile. *Clinical Linguistics & Phonetics*, *15*(7), 505–528. https://doi.org/10.1080/02699200110078159

Melchers, K. G., Lienhardt, N., Von Aarburg, M., & Kleinmann, M. (2011). Is more structure really better? A comparison of frame-of-reference training and descriptively anchored rating scales to improve interviewers' rating quality. *Personnel Psychology*, *64*(1), 53–87. https://doi.org/10.1111/j.1744-6570.2010.01202.x

Munson, B., Bjorum, E. M., & Windsor, J. (2003). Acoustic and perceptual correlates of stress in nonwords produced by children with suspected developmental apraxia of speech and children with phonological disorder. *Journal of Speech, Language, and Hearing Research*, *46*(1), 189–202. https://doi.org/10.1044/1092-4388(2003/015)

Nadig, A., & Shaw, H. (2012). Acoustic and perceptual measurement of expressive prosody in high-functioning autism: Increased pitch range and what it means to listeners. *Journal of Autism & Developmental Disorders*, *42*(4), 499–511. https://doi.org/10.1007/s10803-011-1264-3

Nakai, Y., Takashima, R., Takiguchi, T., & Takada, S. (2014). Speech intonation in children with autism spectrum disorder. *Brain and Development*, *36*(6), 516–522. https://doi.org/10.1016/j.braindev.2013.07.006

Nakai, Y., Takiguchi, T., Matsui, G., Yamaoka, N., & Takada, S. (2017). Detecting abnormal voice prosody through single-word utterances in children with autism spectrum disorders. *Perceptual And Motor Skills*, 31512517716855–31512517716855. https://doi.org/10.1177/0031512517716855

NearPod. (n.d.). nearpod.com

Nicolosi, L., Harryman, E., & Kresheck, J. (2004). *Terminology of communication disorders: Speech-language-hearing* (5th Edition). Riverwoods, IL: Lippincott Williams & Wilkins.

Parish-Morris, J., Liberman, M. Y., Cieri, C., Herrington, J. D., Yerys, B. E., Bateman, L., Donaher, J., Ferguson, E., Pandey, J., & Schultz, R. T. (2017). Linguistic camouflage in girls with autism spectrum disorder. *Molecular Autism*, *8*(1). https://doi.org/10.1186/s13229-017-0164-6

Paul, R., Augustyn, A., & Klin, A. (2005). Perception and production of prosody by speakers with autism spectrum disorders. *Journal of Autism & Developmental Disorders*, *35*(2), 205–220. https://doi.org/10.1007/s10803-004-1999-1

Paul, R., Shriberg, L. D., McSweeny, J., Cicchetti, D., Klin, A., & Volkmar, F. (2005). Brief report: Relations between prosodic performance and communication and socialization ratings in high functioning speakers with autism spectrum disorders. *Journal of Autism & Developmental Disorders*, *35*(6), 861–869. https://doi.org/10.1007/s10803-005-0031-8

Peppé, S. J. E. (2009). Why is prosody in speech-language pathology so difficult? *International Journal of Speech-Language Pathology*, *11*(4), 258–271. https://doi.org/10.1080/17549500902906339

Peppé, S., Maxim, J., & Wells, B. (2000). Prosodic variation in southern British English. *Language and Speech*, *43*(3), 309–334. https://doi.org/10.1177/00238309000430030501

Peppé, S., & McCann, J. (2003). Assessing intonation and prosody in children with atypical language development: The PEPS-C test and the revised version. *Clinical*

Linguistics & Phonetics, *17*(4–5), 345–354. https://doi.org/10.1080/0269920031000079994

Peppé, S., McCann, J., Gibbon, F., O'Hare, A., & Rutherford, M. (2006). Assessing prosodic and pragmatic ability in children with high-functioning autism. *Journal of Pragmatics*, *38*(10), 1776–1791. https://doi.org/10.1016/j.pragma.2005.07.004

Pernambuco, L., Espelt, A., & Costa de Lima, K. (2017). Screening for voice disorders in older adults (RAVI)—Part III: Cutoff score and clinical consistency. *Journal of Voice*, *31*(1), 117.e17-117.e22. https://doi.org/10.1016/j.jvoice.2016.03.003

Pfennings, L., Cohen, L., & van der Ploeg, H. (1995). Preconditions for sensitivity in measuring change: Visual analogue scales compared to rating scales in a Likert format. *Psychological Reports*, *77*(2), 475–480. https://doi.org/10.2466/pr0.1995.77.2.475

Pike, G. R., & Rocconi, L. M. (2012). Multilevel modeling: Presenting and publishing the results for internal and external constituents. *New Directions for Institutional Research*, *2012*(154), 111–124. https://doi.org/10.1002/ir.20017

Quené, H., & van den Bergh, H. (2004). On multi-level modeling of data from repeated measures designs: A tutorial. *Speech Communication*, *43*(1–2), 103–121. https://doi.org/10.1016/j.specom.2004.02.004

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Redford, M. A., Kapatsinski, V., & Cornell-Fabiano, J. (2018). Lay listener classification and evaluation of typical and atypical children's speech. *Language and Speech*, *61*(2), 277–302. https://doi.org/10.1177/0023830917717758

Revelle, W. (2019). *psych: Procedures for psychological, psychometric, and personality research* (Version 1.9.12) [R]. https://CRAN.R-project.org/package=psych

Schölderle, T., Staiger, A., Lampe, R., Strecker, K., & Ziegler, W. (2016). Dysarthria in adults with cerebral palsy: Clinical presentation and impacts on communication. *Journal of Speech Language and Hearing Research*, *59*(2), 216. https://doi.org/10.1044/2015_JSLHR-S-15-0086

Semel, E., Wiig, E. H., & Secord, W. A. (2003). *Clinical Evaluation of Language Fundamentals—Fourth Edition*. San Antonio, TX: Pearson.

Shriberg, L. D., Kwiatkowski, J., & Rasmussen, C. (1990). *Prosody-Voice Screening Profile (PVSP): Scoring forms and training materials.* Tucson, AZ: Communication Skill Builders.

Shriberg, L. D., Kwiatkowski, J., Rasmussen, C., Lof, G. L., & Miller, J. F. (1992). *The Prosody-Voice Screening Profile (PVSP): Psychometric data and reference information for children*. 54. Madison, WI: University of Wisconsin-Madison

Simmons, E., Paul, R., & Shic, F. (2016). Brief report: A mobile application to treat prosodic deficits in autism spectrum disorder and other communication impairments: A pilot study. *Journal of Autism & Developmental Disorders*, *46*(1), 320–327. https://doi.org/10.1007/s10803-015-2573-8

Smith, D. L., & Gillon, G. T. (2004). Autistic spectrum disorder: Caseload characteristics, and interventions implemented by speech-language therapists. *KAIRARANGA, 5*(2), 46-54.

Stevens, K. N., Nickerson, R. S., & Rollins, A. M. (1983). Suprasegmental and postural aspects of speech production and their effect on articulatory skills and

intelligibility. *Speech of the Hearing Impaired: Research, Training and Personnel Preparation*, 35–51.

Støre-Valen, J., Ryum, T., Pedersen, G. A. F., Pripp, A. H., Jose, P. E., & Karterud, S. (2015). Does a web-based feedback training program result in improved reliability in clinicians' ratings of the Global Assessment of Functioning (GAF) Scale? *Psychological Assessment*, *27*(3), 865–873. https://doi.org/10.1037/pas0000086

Strand, E. A., Duffy, J. R., Clark, H. M., & Josephs, K. (2014). The apraxia of speech rating scale: A tool for diagnosis and description of apraxia of speech. *Journal of Communication Disorders*, *51*, 43–50. https://doi.org/10.1016/j.jcomdis.2014.06.008

Szczepek Reed, B. (2010). Prosody and alignment: A sequential perspective. *Cultural Studies of Science Education*, *5*(4), 859–867. https://doi.org/10.1007/s11422-010-9289-z

Szczepek Reed, B. (2011). *Analysing conversation: An introduction to prosody*. Palgrave Macmillan.

't Hart, J., Collier, R., & Cohen, A. (1990). *A perceptual study of intonation*. Cambridge University Press.

Tabuse, H., Kalali, A., Azuma, H., Ozaki, N., Iwata, N., Naitoh, H., Higuchi, T., Kanba, S., Shioe, K., Akechi, T., & Furukawa, T. A. (2007). The new GRID Hamilton Rating Scale for Depression demonstrates excellent inter-rater reliability for inexperienced and experienced raters before and after training. *Psychiatry Research*, *153*(1), 61–67. https://doi.org/10.1016/j.psychres.2006.07.004

Taylor, B., & Thompson, A. (2008). *The international system of units (SI)*. (NIST Special Publication 811). Gaithersburg, MD: U.S. Department of Commerce.

Tench, P. (1996). *The intonation systems of English*. United Kingdom: Bloomsbury Publishing PLC.

Thurber, C., & Tager-Flusberg, H. (1993). Pauses in the narratives produced by autistic, mentally retarded, and normal children as an index of cognitive demand. *Journal Of Autism And Developmental Disorders*, *23*(2), 309–322. https://doi.org/10.1007/BF01046222

Thurm, A., Bishop, S., & Shumway, S. (2011). Developmental issues and milestones. In J. L. Matson & P. Sturmey (Eds.), *International Handbook of Autism and Pervasive Developmental Disorders* (pp. 159–173). Springer New York.

Tseng, C., Pin, S., Lee, Y., Wang, H., & Chen, Y. (2005). Fluent speech prosody: Framework and modeling. *Speech Communication*, *46*(3–4), 284–309. https://doi.org/10.1016/j.specom.2005.03.015

Ukrainetz, T. A. (2015). *School-age Language Intervention: Evidence-based practices*. Austin, TX: Pro-Ed.

Van Wijngaarden-Cremers, P. J. M., van Eeten, E., Groen, W. B., Van Deurzen, P. A., Oosterling, I. J., & Van der Gaag, R. J. (2014). Gender and age differences in the core triad of impairments in autism spectrum disorders: A systematic review and meta-analysis. *Journal of Autism and Developmental Disorders*, *44*(3), 627–635. https://doi.org/10.1007/s10803-013-1913-9

Vaz Freitas, S., Pestana, P. M., Almeida, V., & Ferreira, A. (2014). Audio-perceptual evaluation of Portuguese voice disorders—An inter- and intrajudge reliability

study. *Journal of Voice*, *28*(2), 210–215. https://doi.org/10.1016/j.jvoice.2013.08.001

Wells, B., & Peppé, S. (2003). Intonation abilities of children with speech and language impairments. *Journal of Speech, Language, and Hearing Research*, *46*(1), 5–20. https://doi.org/10.1044/1092-4388(2003/001)

Wells, B., Peppé, S., & Goulandris, N. (2004). Intonation development from five to thirteen. *Journal of Child Language*, *31*(4), 749–778. https://doi.org/10.1017/S030500090400652X

Wichmann, A. (2000). *Intonation in text and discourse: Beginnings, middles and ends*. United Kingdom: Taylor & Francis Ltd.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag.

Wiklund, M. (2016). Interactional challenges in conversations with autistic preadolescents: The role of prosody and non-verbal communication in other-initiated repairs. *Journal of Pragmatics*, *94*, 76–97. https://doi.org/10.1016/j.pragma.2016.01.008

Wynn, C. J., Borrie, S. A., & Sellers, T. P. (2018). Speech rate entrainment in children and adults with and without autism spectrum disorder. *American Journal of Speech-Language Pathology*, *27*(3), 965. https://doi.org/10.1044/2018_AJSLP-17-0134

Yiu, E. M. -L., Chan, K. M. K., & Mok, R. S. -M. (2007). Reliability and confidence in using a paired comparison paradigm in perceptual voice quality evaluation. *Clinical Linguistics & Phonetics*, *21*(2), 129–145. https://doi.org/10.1080/02699200600756355

CURRICULUM VITAE

SARAI S. HOLBROOK

(385) 414-3993, sarai.holbrook@aggiemail.usu.edu

EDUCATION

Doctor of Philosophy Anticipated: May 2020
Disability Disciplines
Utah State University, Logan, UT
Primary Mentor: Sandra Gillam, PhD., CCC-SLP

Dissertation Title: *Validation of a brief prosody rating scale for children with autism spectrum disorder*

- Overall 3.9 GPA
- Fellowship recipient 2017-2018

Master of Science December 2012
Communication Disorders
Brigham Young University, Provo, UT
Primary Mentor: Bonnie Brinton, PhD., CCC-SLP

Master's Thesis: *The effect of intervention using a robot on the social engagement behaviors of four children with autism in interaction with an unfamiliar adult*

- Overall 3.9 GPA
- Academic scholarship recipient

Bachelor of Arts June 2008
Major: Communication Sciences and Disorders
Minor: Music
Western Washington University, Bellingham, WA

- Overall 3.8 GPA
- Graduated cum laude

PUBLICATIONS

Articles

Holbrook, S., Israelsen, M. (2020). *Speech prosody interventions for persons with autism spectrum disorders: A systematic review*. Manuscript submitted for publication.

Gillam, S., **Holbrook, S**., Mecham, J., & Weller, D. (2018). *Pull the Andon rope on working memory capacity interventions until we know more*. Language, Speech and Hearing Services in Schools. *49* (July 2018), 434-448. doi:10.1044/2018_LSHSS-17-0121

Gillam, S., **Holbrook, S. S.**, & Westenskow, A. (2016). *The language of math: Part of the informational discourse continuum*. Perspectives of the ASHA Special Interest Groups, *1*(1),

118–127. https://doi.org/10.1044/persp1.SIG1.118

Book Chapters

Gillam, S., & **Holbrook, S.** (2019). Language intervention/therapy. In Damico, J.S. & Ball, M.J. (Eds.), *The SAGE Encyclopedia of Human Communication Sciences and Disorders*. Thousand Oaks, CA: SAGE

Gillam, R.B., Gillam, S.L., **Holbrook, S.**, & Orellana, C. (2017). Language disorder in children. In Goldstein, S. & DeVries, M. (Eds.), *Handbook of DSM-5 Disorders in Children and Adolescents* (pp. 57-76), NY: Springer. https://doi.org/10.1007/978-3-319-57196-6

In Preparation

Israelsen, M., Fox, C., **Holbrook, S.,** Gillam, S., Gillam, R. *Clinical decision making in narrative intervention: A data-driven approach*. Manuscript in preparation.

Gillam, S., **Holbrook, S.,** & Kamhi, A. *Specific language impairment*. Book chapter in preparation.

Holbrook, S., Orellana, C., Schwartz, S., & Gillam, S. *Developmental trajectory of narrative comprehension in literal and inferential questions*. Manuscript in preparation.

Holbrook, S. *Training the teacher: Creating self-aware, theory-based instructors in communication disorders classrooms.* Manuscript in preparation.

Holbrook, S. *Beyond the closet: Empowering language therapy through theory.* Manuscript in preparation.

Holbrook, S., Gillam, S., Passey, A. *Development of elaborated noun phrase usage in children with typical development ages 4-7.* Manuscript in preparation.

Holbrook, S., Gillam S., Beck, T., Reische, D., Froerer, C. *Playing in the SAND: Development of the Swift Analysis of Narrative Development for children ages 4-7.* Manuscript in preparation.

Wada, R. & **Holbrook, S.** *Religious views on disability: A summary of the literature*. Manuscript in preparation.

Meibos, A., Whicker, J., **Holbrook, S.** *Clinical supervision practices in graduate SLP/AuD training: More evidence needed.* Manuscript in preparation.

NATIONAL CONFERENCE PRESENTATIONS

Froerer, C., Siler, A., Green, N., **Holbrook, S.,** & Gillam, S. (April 2019). *Quick-SAND: Narrative quality rating of stories told by typically developing children ages 8-13.* Poster session presented at the annual National Conference on Undergraduate Research, Kennesaw State University, Kennesaw, GA.

Holbrook, S., Orellana, C., Schwartz, S., & Gillam, S. (November 2018). *Developmental trajectory of narrative comprehension in literal and inferential questions: Effects of demographics and context.* ePoster session presented at the annual meeting of the American Speech-Language-Hearing Association, Boston, MA.

Holbrook, S., Israelsen, M. (November 2018). *Speech prosody interventions for persons with autism spectrum disorder: A systematic review.* Poster session presented at the annual meeting of the American Speech-Language-Hearing Association, Boston, MA.

Beck, T., **Holbrook, S**., Gillam, S. (November 2018). *Holistic assessment of narrative discourse: A progress monitoring tool.* Oral session presented at the annual meeting of the American Speech-Language-Hearing Association, Boston, MA.

Gillam, S., **Holbrook, S**., Mecham, J., & Weller, D. (November 2018). *Improving working memory efficiency for children with developmental language disorders.* Oral session presented at the annual meeting of the American Speech-Language-Hearing Association, Boston, MA.

Holbrook, S., Orellana, C., Schwartz, S., & Gillam, S. (June 2018). *Developmental changes in the response to literal and inferential comprehension questions: Demographic and contextual effects.* Poster session presented at the 39th annual Symposium on Research in Child Language Disorders, Madison, WI.

Simms, G., Browning, S., Peterson, M., **Holbrook, S.,** Gillam, S. (April 2018). *Coding acoustic properties of discourse for students with autism spectrum disorders (ASD).* Poster session presented at the annual National Conference on Undergraduate Research, University of Central Oklahoma, Oklahoma City, OK.

Browning, S., Miller, A., Johnson, N., Southwick, S., **Holbrook, S.,** & Gillam, S. (April 2018). *Grammatical accuracy of narratives produced by typically developing children ages 4-7.* Poster session presented at the annual National Conference on Undergraduate Research, University of Central Oklahoma, Oklahoma City, OK.

Passey, A., Eggertson, K., Polson, B., **Holbrook, S.,** & Gillam, S. (April 2018). *Elaborative noun phrase use in narratives by school age children.* Poster session presented at the annual National Conference on Undergraduate Research, University of Central Oklahoma, Oklahoma City, OK.

Reische, D., Froerer, C., Mumford, S., Beck, T., **Holbrook, S.,** & Gillam, S. (April 2018). *Give us a HAND: Holistic narrative quality rating of stories told by typically developing children.* Poster session presented at the annual National Conference on Undergraduate Research, University of Central Oklahoma, Oklahoma City, OK.

Scott, K., Hampshire, T., Ashcroft, H., Israelsen, M., **Holbrook, S.,** & Gillam, S. (April 2018). *Narrative proficiency of stories produced by typically developing students ages 4-7.* Poster session presented at the annual National Conference on Undergraduate Research, University of Central Oklahoma, Oklahoma City, OK.

Holbrook, S., Israelsen, M., Winward, S., Zemke M., Lindstrom, M., Hammon-Stenquist, M., Gillam, S. (November 2017). *Narrative intervention with children with hearing loss: Facilitating complex discourse.* Oral session presented at the meeting of the American Speech-Language-

Hearing Association, Los Angeles, CA.

Holbrook, S. (November 2017). *Beyond the closet: Empowering language assessment & intervention through theory.* Poster session presented at the meeting of the American Speech-Language-Hearing Association, Los Angeles, CA.

Meibos, A., Whicker, J., **Holbrook, S.** (November 2017). *Clinical supervision practices in graduate SLP/AuD training: More evidence needed.* Poster session presented at the meeting of the American Speech-Language-Hearing Association, Los Angeles, CA.

Holbrook, S., Westenskow, A., Gillam, S., Long, S., Hansen, M., Zemke, M., Beck, T., Gibson, H., Forbes, R. (November 2016). *Improving math abilities in students with language impairments through interprofessional practice (IPP).* Poster session presented at the meeting of the American Speech-Language-Hearing Association, Philadelphia, PA.

Brinton, B., Fujiki, M., Robinson, L., Colton, M., Goodrich, M., Blanchard, K., **Dodge, S.**, Roueche, C., & Stabenow, A. (November 2013) *Using a humanoid robot to facilitate social interaction in children with ASD.* Poster session presented at the American Speech-Language-Hearing Association Convention, Chicago, IL.

REGIONAL/STATE CONFERENCE PRESENTATIONS

Froerer, C., Siler, A., Green, N., **Holbrook, S.,** & Gillam, S. (April 2019). *Quick-SAND: Narrative quality rating of stories told by typically developing children ages 8-13.* Poster session presented at the research week of Utah State University, Logan, UT.

Froerer, C., Siler, A., Green, N., **Holbrook, S.,** & Gillam, S. (February 2019). *Quick-SAND: Narrative quality rating of stories told by typically developing children ages 8-13.* Poster session presented at the 13th annual Utah Conference on Undergraduate Research, Weber State University, Ogden, UT.

Browning, S., Miller, A., Johnson, N., Southwick, S., **Holbrook, S.,** & Gillam, S. (February 2018). *Grammatical accuracy of narratives produced by typically developing children ages 4-7.* Poster session presented at the 12th annual Utah Conference on Undergraduate Research, Southern Utah University, Cedar City, UT.

Passey, A., Eggertson, K., Polson, B., **Holbrook, S.,** & Gillam, S. (February 2018). *Elaborative noun phrase use in narratives by school age children.* Poster session presented at the 12th annual Utah Conference on Undergraduate Research, Southern Utah University, Cedar City, UT.

Reische, D., Froerer, C., Mumford, S., Beck, T., **Holbrook, S.,** & Gillam, S. (February 2018). *Give us a HAND: Holistic narrative quality rating of stories told by typically developing children.* Poster session presented at the 12th annual Utah Conference on Undergraduate Research, Southern Utah University, Cedar City, UT.

Scott, K., Hampshire, T., Ashcroft, H., Israelsen, M., **Holbrook, S.,** & Gillam, S. (February 2018). *Narrative proficiency of stories produced by typically developing students ages 4-7.* Poster session presented at the 12th annual Utah Conference on Undergraduate Research, Southern Utah University, Cedar City, UT.

Holbrook, S. (April 2017). *Narrative intervention with children with hearing loss: Facilitating complex discourse*. Oral session presented at the research week of Utah State University, Logan, UT.

Beck, T., Huffaker, S., Israelsen, M., **Holbrook, S.**, & Gillam, S. (February 2017). *Holistic quality of the narratives produced by children with hearing loss before, during and after a narrative intervention.* Poster session presented at the *Utah Conference on Undergraduate Researc*h, Utah Valley University.

Southwick, S., Weller, D., Israelsen, M., & **Holbrook, S.** & Gillam, S. (February 2017). *Grammatical accuracy of narratives produced by students with hearing loss before, during and after participating in a narrative intervention.* Poster session presented at the *Utah Conference on Undergraduate Researc*h, Utah Valley University.

Winward, S., Forbes, R., Peterson, M., Israelsen, M., & **Holbrook, S.**, & Gillam, S. (February 2017). *Syntactic complexity of narratives produced by students with hearing loss before, during and after participating in a narrative intervention.* Poster session presented at the *Utah Conference on Undergraduate Researc*h, Utah Valley University.

Lindstrom, M., Williams, M., Scott, K., Israelsen, M., **Holbrook, S.**, & Gillam, S. (February 2017). *Narrative Proficiency of stories produced by students with hearing loss before, during and after participating in a narrative intervention.* Poster session presented at the *Utah Conference on Undergraduate Researc*h, Utah Valley University.

CURRENT PROJECTS

Development of the SAND: Swift Analysis of Narrative Development

- Developing a narrative quality rating scale of children's narratives adjusted from the McFadden and Gillam (1996) scale
- Creating age norms and reliability information for the scale using a large normative database

Narrative discourse proficiency in children and adolescents

- Coordinating detailed analyses of narratives produced by children between the ages of 4 and 15 who participated as part of the normative sample of the TNL-2 (Gillam & Pearson, 2016), including analyses of overall narrative proficiency and elaborated noun phrase use

RESEARCH EXPERIENCE

Intervention Coordinator August 2017 – Present
Child Language Research Lab, Logan, UT

- As part of the intervention team, orchestrate provision of intervention using the Supporting Knowledge in Language and Literacy (SKILL) narrative intervention

program as a part of a multi-site IES efficacy and replication (Goal 3) grant over 3 years of data collection
- Train and monitor interventionists under the direction of the Co-PI of intervention
- Provide SKILL intervention to small groups of children at risk for language and learning disabilities with high levels of fidelity
- Conduct screening, assessment, and scoring duties as needed

Research Lab Manager August 2015 – January 2019
Child Language Research Lab, Logan, UT
- Coordinate ongoing research projects including project planning, recruitment, data collection, and data analysis
- Supervise and train undergraduate and graduate research assistants in data collection and analysis
- Mentor undergraduate and graduate students as they develop projects for regional and national conferences such as UCUR, NCUR, and ASHA

Research Assistant July 2013 – August 2013
Jordan Family Education Center, South Jordan, UT
- Co-facilitated a class designed to teach siblings of students with autism about autism spectrum disorders to improve sibling interactions.
- Discussed research complications and solutions with primary investigator

Master's Thesis Researcher January 2011 – July 2012
Brigham Young University, Provo, UT
- Researched the effects of utilizing a humanoid robot to teach social engagement behaviors to children with low-functioning autism
- Assisted to develop, establish the reliability of, and utilize a behavioral coding system
- Participated in research design and planning with mechanical engineering, communication disorders, and computer science professors and students

TEACHING EXPERIENCE

Primary Instructor

Introduction to Communicative Disorders Fall Semester, 2016
COMD 2600, Utah State University, Logan, UT
- Planned and carried out instruction of 103 undergraduate students
- Organized and maintained web-based instructional site, Canvas
- Provided exposure to experts in multiple fields through invited guest lectures

Guest Lecturer

Autism Spectrum Disorder Overview and Assessment May 21, 2019
EDPS 6836/7836, University of Utah, Salt Lake City, UT
- Presented by invitation to a class of school psychology graduate students

- Highlighted potential for interprofessional practice (IPP) between school psychologists and speech-language pathologists
- Explained the role of the speech-language pathologist in assessment and intervention for children with ASD

Introduction to Communicative Disorders Fall Semesters, 2015, 2017, and 2018
COMD 2600, Utah State University, Logan, UT

- Gave lectures on several topics, including cultural and linguistic diversity in communication, developmental speech disorders, swallowing function and disorders, and fluency disorders
- Implemented flipped classroom techniques
- Showed multiple video exemplars of disorders and selections from relevant documentaries to contextualize content

Phonological Disorders Summer Semesters, 2017 and 2018
COMD 6150, Utah State University, Logan, UT

- Lectured on phonological awareness and speech sound disorders
- Utilized interactive lecture activities such as PollEv and small-group discussion

School-age Language Disorders Fall Semester, 2017
COMD 6020, Utah State University, Logan, UT

- Lectured on standardized diagnostic language measures
- Provided opportunities for hands-on experience with multiple standardized tests

Teaching Assistant

Brigham Young University, Provo, UT July 2011 – June 2012

- Planned and carried out review sessions for undergraduate and graduate level classes
- Provided detailed feedback to students both in person and in writing
- Assisted struggling students with content mastery

PROFESSIONAL SERVICE

Reviewer January – May 2020
ASHA Convention 2020 Topic Committee
Language and Learning in School-Age Individuals

Ad Hoc Reviewer September 2018
PLOS ONE

SUPERVISION

Clinical Instructor August 2016 – May 2017
Utah State University Speech-Language-Hearing Center, Logan, UT

- Supervised four master's level speech-language pathology students in clinical practicum experiences
- Supervised assessment and intervention of clients with hearing impairment, speech sound disorder, repaired cleft lip and palate, and expressive language delay

RELATED PROFESSIONAL EXPERIENCE

Speech-Language Pathologist August 2012 – June 2015
Jordan School District, Riverton, UT

- Organized and implemented speech, language, and social communication therapy for a caseload of over 60 students, ages 4 through 12, including 2 self-contained classrooms of students with high functioning autism
- Presented trainings for 2 continuing education in-service meetings for district speech-language pathologists
- Screened, evaluated, and made diagnostic decisions for incoming speech-language referrals
- Collaborated with general education teachers, special education teachers, and school psychologists to create unified intervention programs
- Utilized a variety of intervention models including push-in, pull-out, and home-based individual and group therapy
- Investigated and implemented intervention strategies for challenging cases, including students with persistent selective mutism and multiple disabilities necessitating AAC device usage
- Assisted in supervising a graduate student extern 1 day a week for 9 weeks

Part-Time Speech-Language Pathologist July 2014 – March 2015
Sandy Health and Rehab, Sandy, UT

- Implemented efficient dysphagia, cognitive rehabilitation, speech, and language assessment and intervention for adults ranging in age from 40 to over 90 years old with disabilities such as stroke, cerebral palsy, and dementia
- Coordinated with other speech therapists to maintain continuity of care
- Collaborated with occupational therapists and physical therapists to establish unified, functional objectives for optimal patient recovery
- Communicated with dietary staff to ensure patients with dysphagia received diet textures consistent with their level of impairment
- Maintained flexibility in working with other speech therapists, occupational therapists, and physical therapists

PRN Speech-Language Pathologist August 2013 – March 2015
Provo Rehabilitation and Nursing, Provo, UT

- Provided swallowing, cognitive rehabilitation, and language evaluation and therapy for a variety of adult patients, including those on ventilators and with tracheostomies
- Maintained detailed yet concise medical records consistent with the most current

medical billing and documentation standards

AWARDS AND SCHOLARSHIPS

Frederick Q. Lawson Fellowship Fall 2017 – Spring 2018

Student Scholar Award September 18-19, 2015
10th Annual Eleanor M. Saffran Cognitive Neuroscience Conference
Philadelphia, PA

PROFESSIONAL ORGANIZATION CERTIFICATIONS AND MEMBERSHIPS

Certificate of Clinical Competence in Speech-Language Pathology	Current
Active Speech Language Pathologist License, Utah Department of Commerce	Current
Member, American Speech-Language-Hearing Association	Current
Member, National Student Speech-Language-Hearing Association	Past