Stability of Variables Derived from Measures of Multisensory Function in Children with Autism Spectrum Disorder

Kacie Dunham, BA[a], Jacob I. Feldman, MS[b],
Yupeng Liu[c], Margaret Cassidy[c], Julie G. Conrad, BA[c,d], Pooja Santapuram, BA[c,e],
Evan Suzman, BS[f], Alexander Tu, BA[c,g], Iliza Butera, BA[a], David M. Simon, PhD[a,h],
Neill Broderick, PhD[i,j], Mark T. Wallace, PhD[a,b,i,k-m], David Lewkowicz, PhD[n],
*Tiffany G. Woynaroski, PhD[a,i,o]

[a] Vanderbilt Brain Institute, Vanderbilt University, Nashville, TN, USA

[b] Department of Hearing & Speech Sciences, Vanderbilt University, Nashville, TN, USA

[c] Neuroscience Undergraduate Program, Vanderbilt University, Nashville, TN, USA

[d] Present Address: College of Medicine, University of Illinois, Chicago, IL, USA

[e] Present Address: School of Medicine, Vanderbilt University, Nashville, TN, USA

[f] Department of Biomedical Sciences, Vanderbilt University, Nashville, TN, USA

[g] Present Address: College of Medicine, University of Nebraska Medical Center, Omaha, NE, USA

[h] Present Address: axialHealthcare, Nashville, TN, USA

[i] Vanderbilt Kennedy Center, Vanderbilt University Medical Center, Nashville, TN, USA

[j] Department of Pediatrics, Vanderbilt University Medical Center, Nashville, TN, USA

[k] Department of Psychology, Vanderbilt University, Nashville, TN, USA

[l] Department of Psychiatry and Behavioral Sciences, Vanderbilt University Medical Center, Nashville, TN, USA

[m] Department of Pharmacology, Vanderbilt University, Nashville, TN, USA

[n] Department of Communication Sciences & Disorders, Northeastern University, Boston, MA, USA

[o] Department of Hearing & Speech Sciences, Vanderbilt University Medical Center, Nashville, TN, USA

*Correspondence regarding this manuscript may be addressed to:

Tiffany Woynaroski, PhD, CCC-SLP

Email address: tiffany.g.woynaroski@vumc.org

Assistant Professor of Hearing and Speech Sciences

Vanderbilt University Medical Center

Vanderbilt Kennedy Center

Vanderbilt Brain Institute

1215 21st Ave South, Room 8310

Nashville, TN, 37232-8242, USA

Acknowledgments

This work was supported by NIH U54 HD083211, NIH/NCATS KL2TR000446, NIH/NIDCD R21 DC016144, NIH/NIDCD F31 DC015956 and NIH T32 MH064913. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the funding agencies. Results from this manuscript were previously presented at the 2018 International Multisensory Research Forum and the 2019 Gatlinburg Conference on Research and Theory in Intellectual and Developmental Disabilities.

Stability of Variables Derived from Measures of Multisensory Function in Children with Autism Spectrum Disorder

Abstract

Children with autism spectrum disorder (ASD) display differences in multisensory function as quantified by several different measures. This study estimated the stability of variables derived from commonly used measures of multisensory function in school-aged children with ASD. Participants completed: a simultaneity judgment task for audiovisual speech, tasks designed to elicit the McGurk effect, listening-in-noise tasks, electroencephalographic recordings, and eye tracking tasks. Results indicate the stability of variables derived from tasks tapping multisensory processing is variable. These findings have important implications for measurement in future research. Averaging scores across repeated observations will often be required to obtain acceptably stable estimates, and thus to increase the likelihood of detecting effects of interest, as it relates to multisensory processing in children with ASD.

Keywords: multisensory, autism, stability, reliability, psychometrics, measurement

Stability of Variables Derived from Measures of Multisensory Function in Children with Autism Spectrum Disorder

Recent literature and updated diagnostic criteria suggest that sensory abnormalities represent a core feature of autism spectrum disorder (ASD; American Psychological Association [APA], 2013). Children with ASD have been observed to display unusual responses to stimuli presented within a number of sensory modalities (e.g., audition and vision) and to demonstrate atypical responses to sensory stimuli presented across multiple sensory modalities (i.e., multisensory stimuli, such as audiovisual stimuli) in studies employing a broad range of measures, including psychophysical tasks, eye tracking, and electroencephalography (EEG). Differences in responding to multisensory stimuli are most consistently observed, seemingly, for stimuli that are social in nature, specifically audiovisual speech stimuli (Baum, Stevenson, & Wallace, 2015; Irwin & DiBlasi, 2017; Smith, Zhang, & Bennetto, 2017; Stevenson et al., 2014).

The aforementioned findings have fostered interest in exploring the extent to which variables derived from measures of audiovisual speech processing and perception may be valid for predicting other core and related symptoms of ASD, and the degree to which such indices are potentially malleable with targeted treatment. It is critical, however, to first ascertain the stability of variables derived from measures of multisensory function, as this influences the potential validity of these variables for predicting ASD and related symptomatology, as well as detecting training and intervention effects (Mccrae, Kurtz, Yamagata, & Terracciano, 2011). Indeed, discrepant findings across past studies exploring multisensory processing in children with ASD could be explained by instability of the measures that have been employed in prior work (Feldman et al., 2018; Magnotti & Beauchamp, 2018). Therefore, the present study explores the

stability of several variables derived from commonly used measures of multisensory function, in particular variables indexing attention to and integration of audiovisual speech.

Attention to and Integration of Audiovisual Speech in Typically Developing Individuals

Speech is inherently a multisensory process, wherein highly synchronized cues from the moving mouth accompany the dynamic acoustic signal (Chandrasekaran, Trubanova, Stillittano, Caplier, & Ghazanfar, 2009). The perception of speech is influenced by the presence of these complementary visual speech cues (Calvert & Campbell, 2003; Massaro & Palmer, 1998). This fact is illustrated by the McGurk Effect, a perceptual illusion wherein persons presented with incongruent auditory and visual speech cues (e.g., a visual “ga” paired with an auditory “ba”) often report perceiving an illusory percept (e.g., “da” or “tha”) purported to reflect a “fusion” of the mismatched multisensory information (McGurk & MacDonald, 1976).

Typically developing (TD) individuals attend to and integrate audiovisual speech cues very early in life (Soto-Faraco, Calabresi, Navarra, Werker, & Lewkowicz, 2012). Specifically, TD infants begin to look to the mouth of a speaker (the source of multisensory redundancy) by approximately 8 months of age (Lewkowicz & Hansen-Tift, 2012). Once this propensity to lipread emerges, TD individuals will continue to capitalize on the corresponding visual cues from the mouth across the lifespan whenever speech processing becomes challenging (e.g., when in a noisy environment and/or when faced with an unfamiliar dialect or language; Barenholtz, Mavica, & Lewkowicz, 2016; Buchan, Paré, & Munhall, 2008). TD children begin to show some sensitivity to the temporal synchrony of auditory and visual speech cues, looking preferentially to the visual cues that are more highly correlated with a fluent auditory speech stream in time by approximately their first birthday (i.e., 12-14 months of age; Lewkowicz, Minar, Tift, & Brandon, 2015) and display increasing temporal acuity for audiovisual speech throughout

childhood and adolescence (Hillock-Dunn, Grantham, & Wallace, 2016; Hillock, Powers, & Wallace, 2011; Lewkowicz & Flom, 2014).

Access to multisensory speech cues affords a number of functional benefits. Psychophysical studies have demonstrated that concurrent visual speech cues boost perceptual accuracy substantially for TD individuals, particularly in the presence of noise or otherwise difficult listening conditions (e.g., Fraser, Gagné, Alepins, & Dubois, 2010). Studies of electrophysiological responsiveness in TD children as well as adults have found that their processing of audiovisual speech is more efficient than their processing of auditory-only speech (e.g., Knowland, Mercure, Karmiloff-Smith, Dick, & Thomas, 2014; van Wassenhove, Grant, & Poeppel, 2005). This increase in speech processing efficiency is evident via faster latencies and reduced amplitudes for multiple EEG waveform components indexing the brain's response to speech, in particular in the negative-going deflection that occurs around 100 ms (i.e., N1 or N100) and the positive deflection that occurs around 200 ms (i.e., P2 or P200) following stimulus onset (Knowland et al., 2014; van Wassenhove et al., 2005).

Attention to and Integration of Audiovisual Speech in Individuals with ASD

A large and ever-growing literature utilizing diverse measures of multisensory function suggests that children with ASD display differences in their attention to and integration of audiovisual speech relative to their TD peers (see Feldman et al., 2018 for a review). For example, studies using eye tracking technology have shown that children with ASD display diminished attention to audiovisual speech (i.e., reduced looking to the mouth of a speaker) in comparison to TD controls (Grossman, Steinhart, Mitchell, & McIlvane, 2015; Riby & Hancock, 2009). Investigations using psychophysical approaches have additionally reported that children with ASD tend to show reduced multisensory integration (i.e., report fewer perceptual fusions) in

response to discrepant McGurk stimuli (Iarocci, Rombough, Yager, Weeks, & Chua, 2010; Irwin, Tornatore, Brancazio, & Whalen, 2011; Williams, Massaro, Peel, Bosseler, & Suddendorf, 2004). Furthermore, children with ASD have been observed to exhibit a lesser degree of audiovisual "gain" for speech-in-noise stimuli when compared to TD children (Foxe et al., 2015; Smith & Bennetto, 2007).

School age children with ASD are also less attuned to the typical temporal relations between auditory and visual speech cues. In the context of a simultaneity judgment task, responses to paired audiovisual stimuli presented at various temporal offsets can be used to estimate the window of time over which an individual tends to "bind" auditory and visual information together, or perceive such multisensory information as arising from a unitary event (i.e., the temporal binding window; TBW). When compared to TD controls, children with ASD present with significantly wider TBWs (Stevenson et al., 2014; Woynaroski et al., 2013). Findings from studies using EEG and event-related potentials (ERPs) are limited for this clinical population; however, those that are available suggest that neural responses to multisensory stimuli differ for school age children with ASD as a function of the severity of their core and related symptoms (Brandwein et al., 2013; 2015; Woynaroski et al., in prep).

A Need To Assess Stability of Variables Derived from Commonly Used Measures of Multisensory Speech Processing

The aforementioned findings of altered multisensory function have engendered interest in exploring whether measures of multisensory integration may be valid for predicting ASD and related symptomatology or may be sensitive to effects of sensory-based interventions geared towards children on the autism spectrum. However, we currently know little about the stability of commonly used estimates of audiovisual functioning across observations and contexts. In fact, to

date no study has comprehensively investigated the stability of the variables derived from measures routinely used to tap audiovisual speech processing and perception in children with ASD. Ascertaining the stability of indices of multisensory function in this clinical population is critical because the stability of any given variable limits its validity for detecting effects of interest (i.e., the validity cannot exceed the square root of the stability; DeVellis, 2006; Nunnally, 1978).

Generalizability (G) studies are a useful tool for measuring stability because they allow us to parse the variance in a given variable that is attributable to the construct of interest versus facets of measurement error. Decision (D) studies then draw upon the results of a generalizability study and extrapolate beyond observed data to predict the level of stability that would be achieved for a hypothetical number of observations (and/or levels of other facets of interest in the study; Mushquash & O'Connor, 2006). This study uses G&D studies to ascertain the degree to which variables derived from some of the most frequently used measures of multisensory function are stable and to determine how many observations are required to obtain acceptable stability for variables of interest (Yoder, Lloyd, & Symons, 2018). Our specific research questions were:

(a) How stable are variables derived from the various commonly used measures of multisensory function, in particular indices of selective attention to and integration of audiovisual speech, in children with ASD?

(b) How many observations are required to reach acceptable stability for each of the variables of interest?

Methods

Participants

Eleven children (7 male; 4 female) aged 7-16 years old participated in the study (see Table 1 for descriptive information). Eligibility criteria were as follows: (a) diagnosis of ASD as confirmed by research-reliable administrations of the Autism Diagnostic Observation Schedule, second edition (ADOS-2; Lord et al., 2012) and clinical judgment of a licensed clinician on the research team, (b) no history of seizure disorders, (c) no diagnosed genetic disorders, such as Fragile X, Down syndrome, or tuberous sclerosis, and (d) normal or corrected-to-normal vision and normal hearing, as confirmed by screening at entry to the study.

Procedures

This study was conducted at [WITHHELD FOR BLIND REVIEW]. Participants completed a series of psychophysical, EEG, and eye tracking measures once per day, on two different days, within a one week timeframe. For each participant, data collection was conducted in the same order, and each procedure took place at the same time of day across measurement days; however, procedure order was randomized across participants. All procedures were approved by the Institutional Review Board at [WITHHELD FOR BLIND REVIEW]. Parents provided written informed consent, and participants provided written or verbal assent prior to participation in the study. All participants were compensated for their participation.

Psychophysical measures. Participants completed all psychophysical measures in a sound and light attenuated booth (WhisperRoom Inc., Morristown, TN, USA). Stimulus presentation for all tasks was managed by E-Prime software. Visual stimuli were presented on a Samsung Syncmaster 2233RZ 22 inch PC monitor. Auditory stimuli were presented binaurally via Sennheiser HD550 series supra-aural headphones (simultaneity judgment task and McGurk tasks) or via an M-AUDIO BX8 D2 speaker (listening-in-noise task).

Simultaneity judgment task. Audiovisual stimuli for the simultaneity judgment task consisted of a neutral-faced adult female speaker saying the syllable "ba" against a white background. The auditory and visual components of the stimuli were separated in the video editing software Adobe Premiere. Stimuli were presented either synchronously or asynchronously at various stimulus onset asynchronies (SOAs; the difference in the presentation of the auditory and visual components of the stimuli, with negative values indicating auditory-first and positive values indicating visual-first). Asynchronous stimuli were presented at 14 SOAs: ±500 ms, ±400 ms, ±350 ms, ±300 ms, ±250 ms, ±150 ms, and ±100 ms.

Each participant was instructed to report whether s/he saw and heard the speech at the "same time" or at a "different time" by pressing the "1" and "2" keys on the keyboard, respectively. To ensure comprehension of the task, participants completed a practice round, consisting of two trials of stimuli presented synchronously and two trials of stimuli presented at an SOA of ±900 ms in a randomized order. Participants were required to correctly respond to all items of the practice round prior to starting the task. After this comprehension check, synchronous trials and asynchronous trials at each SOA were presented four times, in a random order (total of 60 trials per run). Participants completed five runs (300 total trials) of the task each day.

Data from E-Prime were exported into MATLAB. TBWs were derived for each child by fitting two psychometric functions to the data for his/her reported rate of perceived synchrony across SOAs (i.e., the number of times that the child answered "synchronous" over the total number of trials presented for each SOA) using the glmfit function in MATLAB (see Powers, Hillock, & Wallace, 2009; Stevenson et al., 2014 for a detailed description of this approach), one for auditory-leading (left) trials and another for visual-leading (right) trials, after normalizing the

data (i.e., setting the maximum value to 100%; see Figure 1). The point at which each psychometric function crossed 75% perceived synchrony was considered the left- and right-TBW. The TBW was then calculated as the difference between these values.

McGurk tasks. Audiovisual stimuli utilized in the McGurk task were derived from media files of the same adult female speaker described above saying the syllables "ba," "ga," "pa," and "ka" with a neutral facial expression. Adobe Premiere was used to create visual-only, auditory-only, matched audiovisual, and mismatched audiovisual (i.e., McGurk; auditory "ba" + visual "ga"; auditory "pa" + visual "ka") stimuli. Participant responses were recorded via a four-button response box labeled with four syllables for each task (i.e., "ba," "ga," "da," "tha" for the Ba/Ga task; "pa," "ka," "ta," "ha" for the Pa/Ka task).

Participants completed two runs each of the two different McGurk tasks, one task with the auditory "ba" and visual "ga" syllables (which frequently induce a fused percept of "da" or "tha") and one task with auditory "pa" and visual "ka" syllables (which frequently induces a fused percept of "ta" or "ha"). Prior to starting the task, participants were provided oral instructions to press the button that corresponded to the syllable they perceived during each trial (e.g., participants were told to "press 'ba' if you think she says 'ba', 'ga' if you think she says 'ga'," etc). Prior to each run of the task, the participants completed a comprehension check wherein they were prompted to press the designated button for each syllable in a random order. During each run, participants were presented with 10 trials of each syllable in the auditory-only, visual-only, and matched audiovisual conditions and 10 trials of the incongruent audiovisual (McGurk) stimuli in a randomized order (70 trials per run). After each trial, participants reported the syllable they perceived using the four-button response box.

Data from E-Prime were exported into MATLAB. Magnitude of multisensory integration in response to McGurk measures was operationalized as the proportion of trials in which participants reported perceiving the illusory percept in response to incongruent audiovisual stimuli in each task (i.e., “da” and “tha” for the Ba/Ga task; “ta” and “ha” for the Pa/Ka task).

Listening-in-noise task. Listening-in-noise stimuli were videos of an adult female speaker saying monosyllablic words with a neutral facial expression (described in Picou, Ricketts, & Hornsby, 2011). These words were arranged in eight lists of 25 words each (as in Picou, Charles, & Ricketts, 2017). Each list was balanced for audibility, and stimuli were presented via a single (mono) speaker positioned above a monitor at 0° azimuth and calibrated to a sound level of 50 dB SPL. Speech-shaped noise was created in MATLAB by generating gaussian white noise via the wgn function and shaped based on the long term average spectrum of the speech (LTASS; described in Donley, Ritz, & Kleijn, 2018). This noise was presented at 53 dB (for a -3dB speech-to-noise ratio [SNR]) and 56 dB (for a -6dB SNR). These SNRs were selected based on the largest group differences previously reported between individuals with ASD and individuals with typical developmental histories in this age range (Foxe et al., 2015). At each SNR, stimuli were presented in audiovisual and auditory-only conditions.

Four wordlists (1-4) were used on the first observation day (i.e., two modalities x two SNRs), and four different wordlists (5-8) were used on the second observation day. The testing order was randomized for each participant on each day. Participants were instructed to listen and repeat the word they perceived to a research assistant who then typed the word and confirmed the participants’ responses orally and via the typed response on the monitor. To ensure comprehension of the task, participants were presented with five words without white noise. The participant was required to correctly identify each word before proceeding to the task. After this

comprehension check, participants were presented with the four lists, one word at a time. Identification accuracy for listening-in-noise measures was calculated as the percent of whole words correctly identified in each condition for each recording day.

ERP measure. Stimuli used in the ERP measure were consonant-vowel syllables (i.e., "ba") naturally spoken by an adult female speaker using a neutral facial expression (see simultaneity judgement and McGurk tasks) in two conditions. In the audiovisual (AV) condition, the corresponding auditory and visual stimuli were presented in synchrony (i.e., with the visual cues from the face and neck temporally preceding the auditory cues in onset as naturally produced by the speaker). In the auditory-only (AO) condition, the auditory stimulus was presented in conjunction with a static face (i.e., a still image of the speaker) in order to isolate the contribution of visual articulatory cues versus simply the presence/absence of a face on speech processing.

Stimuli were presented via E-Prime in conjunction with an Eyelink 1000 Plus eyetracker, which ensured that videos were presented only when participants were gazing at the screen (i.e., when each participant's gaze was focused on a fixation cross centered on the speaker's face for the 500 ms interval immediately preceding stimulus presentation). Data were collected using NetStation and a 128-channel Geodesic sensor net (Net Amps 400 amplifier, Hydrocel GSN 128 EEG cap, EGI Systems Inc.). The raw EEG signal was sampled at 1000 Hz and referenced to the vertex (Cz). Electrode impedances were kept at or below 40 kΩs.

Prior to the task, children's eye gaze was calibrated using a five-point calibration procedure. This procedure was performed twice to validate accuracy of calibration. Participants then viewed a video wherein a member of the research team briefly described the task, and a TD peer modeled the task (i.e., wore the EEG cap and attended to the screen during stimulus

presentation). Following calibration and presentation of the introductory video, the experimental task was initiated. The task employed an equiprobable paradigm, wherein 50 trials of each stimulus type (i.e., AV and AO, as described above) were presented in random order in two blocks for a total of 100 trials of each stimulus type across the two blocks. Trials were separated by an interstimulus interval (ISI) that was randomly jittered between 400 ms and 800 ms plus the 500 ms gaze contingency period (i.e., minimum ISI between 900 ms and 1300 ms). Between the two blocks, children took a scheduled break. During each block, images of cartoon aliens were presented periodically in between trials (i.e., after every fourth trial, there was a 50% chance of an alien image appearing) to maintain participant attention to the task. Participants were instructed to hit a BIGmack button (AbleNet Inc., Roseville, MN, USA) to "catch" the aliens each time one appeared on the screen.

EEG data were bandpass filtered from 0.5Hz to 50Hz, using the EEGLab firfiltnew.m function, which implements a bidirectional zero-phase finite impulse response filter, and artifacts and bad channels were manually removed in EEGLAB (Delorme & Makeig, 2004). An average of 73.2% of trials (146.4 trials) were retained across children. After data were cleaned, they were re-referenced to the average, and removed channels were interpolated. Trials were baseline corrected from 200 ms to 0 ms pre-stimulus onset. The amplitude and latency of the N1 (i.e., window defined a priori as occurring between 100 ms and 140 ms post-stimulus onset) and P2 (i.e., window defined a priori as occurring between 160 ms and 240 ms) as measured at a centrally located electrode site (Cz) were extracted from the grand average waveform of each participant for each EEG observation day and manually reviewed (see Table 2 for further detail re: ERP variables).

Eye tracking measure. Stimuli utilized in the eye tracking measure were 50 second video clips utilized in several past studies of attention to multisensory speech (e.g., Lewkowicz & Hansen-Tift, 2012; Pons, Bosch, & Lewkowicz, 2019). In each video clip, an adult female actor recited a prepared monologue in children's native language (i.e., English) or non-native language (i.e., Spanish) in a child-directed manner (i.e., with high pitch excursions, prosodically exaggerated speech and slow articulation, while smiling) or in an adult-directed manner (i.e., with minimal pitch variation, average speed of articulation, and neutral affect). Visual stimuli were presented on a 24 inch computer monitor positioned approximately 50 cm in front of the participant. Auditory stimuli were presented at 75 dB by an M-AUDIO BX8 D2 speaker placed in front of the participant just below the computer monitor. A Sensorimotorics Instrument (SMI) REDn Scientific Eye Tracking System (SMI, Teltow, Germany) was used to control stimulus presentation and randomization and to track eye gaze via pupil-centered corneal reflection.

Participants were seated in front of the eye tracking system, monitor, and speaker. Eye gaze was calibrated using a five-point calibration procedure during which participants were instructed to watch a looming star that moved from the center to each corner of the computer screen. Following calibration, participants were presented with the four video clips (i.e., English and Spanish clips presented in a child- and adult-directed manner) in random order, on each observation day. Prior to the presentation of each video clip, participants were instructed to "please watch the movie."

SMI's BeGaze software was utilized to automatically quantify the duration of looking to a priori specified regions of interest (ROIs; i.e., the mouth, eyes, and face) during stimulus presentation (see Figure 2). Attention to audiovisual speech was operationalized as the proportion of total looking time (PTLT) deployed to the mouth ROI (the source of multisensory

redundancy) and the eye ROI (a commonly used contrast region), respectively, out of the total time spent fixating any part of the face during stimulus presentation in each condition (i.e., English infant-directed, Spanish infant-directed, English adult-directed, Spanish adult-directed).

Analytic Plan

Generalizability (G) and Decision (D) studies were carried out using EduG (Swiss Society for Research in Education Working Group, 2012). EduG is freeware created specifically for generalizability analysis. G and D studies were conducted on the variables derived from psychophysics, EEG, and eye tracking measures (see Table 2 for a summary). For each of these variables, random effects models constituting a total of 22 observations (11 participants X 2 days) in a crossed design (Participant X Day) were run. Absolute *g* coefficients, which are preferred over relative *g* for their inclusion of all effects of measurement facets in the computation of the coefficient (Yoder et al., 2018), were derived to quantify the level of stability achieved for observed data (i.e., one and two observations). In the D studies, the *g* coefficient was projected beyond the number of observed sessions to determine how many observations would be needed to achieve acceptably stable scores. Our a priori threshold for acceptable stability was set at $g = .8$, a criterion commonly applied in previous stability studies (e.g., Bottema-Beutel et al., 2019; Sandbank & Yoder, 2014; Woynaroski et al., 2017; Yoder, Woynaroski, & Camarata, 2016).

Results

Variables Derived from Eye Tracking Measure

Variables derived from the eye tracking measure showed relatively high stability. For English adult-directed speech, stability was high for both proportion of time looking at the mouth (*g* for a single observation = 0.91) and proportion of time looking at the eyes (*g* for a single

observation = 0.87). These variables exceeded acceptable stability with one observation. During English infant-directed speech, stability was also high for proportion of time looking at the mouth (*g* for a single observation = 0.70) and proportion of time looking at the eyes (*g* for a single observation = 0.75). Both of these variables were acceptably stable after two observations.

For Spanish adult-directed speech, proportion of time looking at the mouth (*g* for a single observation = 0.95) and proportion of time looking at the eyes (*g* for a single observation = 0.96) were both highly stable. These variables both exceed our established threshold for acceptable stability with a single observation. For Spanish infant-directed speech, proportion of time looking at the mouth (*g* for a single observation = 0.74) and proportion of time looking at the eyes (*g* for a single observation = 0.93) were acceptably stable after two observations and one observation, respectively. Figure 3 depicts the results for variables derived from the eye tracking measure.

Variables Derived from Psychophysical Measures

The stability of variables derived from psychophysical measures was mixed. Proportion of reported McGurk illusions in response to auditory “pa” and visual “ka” and TBW for audiovisual speech showed the highest stability (*g* coefficients = 0.84 and 0.74 for a single observation). These variables were acceptably stable after one and two observations, respectively. The remaining variables, the proportion of reported McGurk illusions in response to auditory “ba” and visual “ga” stimuli (*g* for a single observation = 0.47) and whole-word recognition of audiovisual speech presented at –3 dB SNR (*g* for a single observation = 0.31) and –6 dB SNR (*g* for a single observation = 0.19), were less stable. The D studies indicated that proportion of reported Ba/Ga McGurk illusions would be acceptably stable after five observations. Whole word recognition of audiovisual speech presented at –3 dB SNR would be

acceptably stable after nine observations, and whole word recognition of audiovisual speech at –6 dB SNR would require more than ten observations to achieve acceptable stability. Figure 4 summarizes the results for variables derived from psychophysics tasks.

Variables Derived from ERP Measure

The stability of variables derived from the ERP measure was highly heterogeneous. Of these variables, P2 amplitudes for the auditory-only (g for a single observation = 0.74) and audiovisual (g = 0.90) conditions were the most stable, exceeding our criterion for acceptable stability after two observations and one observation, respectively. N1 amplitudes were relatively less stable (g = 0.00 for a single observation in the auditory-only condition, and g = 0.41 for a single observation in the audiovisual condition). D studies indicate that six observations would be required to achieve acceptable stability for N1 amplitude in the audiovisual condition and that it is likely not possible to obtain a stable estimate of N1 amplitude in the auditory-only condition in school age children with ASD even with repeated sampling (i.e., the model shows no sign of converging on acceptable stability even for estimated coefficients at ten or more observations).

In the auditory-only condition, latency variables for both N1 (g for a single observation = 0.05) and P2 (g for a single observation = 0.24) had low stability. In the audiovisual condition, stability for latency of N1 (g for a single observation = 0.11) and P2 (g for a single observation = 0.00) was also low. According to D studies, it would take more than ten observations to achieve acceptable stability for latencies of N1 and P2 across conditions. Refer to Figure 5 for a summary of findings for variables derived from the ERP task. Table 3 provides a detailed summary of estimated stability for each variable of interest according to varied numbers of observations, to facilitate planning for future studies.

Discussion

This study examined the stability of several variables derived from commonly used measures of multisensory function in children with ASD. Stability of variables is critical if measures of multisensory function are to be employed in studies aiming to predict heterogeneity in broader ASD and related symptomatology and/or to assess intervention efficacy in children on the autism spectrum, though such psychometric work has been limited to date (Basu Mallick, Magnotti, & Beauchamp, 2015; Powers et al., 2009). The present results indicate that the stability of variables derived from measures of multisensory function differs across (and in some cases within) measure type in school age children with ASD. Averaging scores across repeated observations will often be required to obtain acceptably stable estimates and increase the likelihood of detecting effects of interest, as it relates to multisensory function in this clinical population.

Variables Derived from Eye Tracking are Highly Stable

Variables derived from the eye tracking measure were highly stable. This result is somewhat surprising, given the limited sampling (i.e., less than one minute of data collection per condition) and relative lack of structure (i.e., passive viewing with little instruction beyond a request to “watch the movie”) associated with this sampling context. Though such eye tracking tasks do not necessarily tap “integration” of multisensory stimuli, the present findings suggest that measures of eye gaze patterns yield variables that are highly stable and thus have potential construct validity for indexing attention to multisensory stimuli, which has been theoretically and empirically linked to social, communication, and language development in children with or at risk for ASD (Klin, Jones, Schultz, Volkmar, & Cohen, 2002; Santapuram et al., in prep; Tenenbaum, Amso, Abar, & Sheinkopf, 2014). It is also notable that these brief and low demand measures have high potential to translate to clinical practice if sufficient support is obtained for

their validity in predicting symptomatology and/or detecting effects of interventions targeting looking behavior and more distal ASD symptoms.

Stability of Variables Derived from Psychophysics Measures is More Variable

In contrast, the stability of variables derived from psychophysics tasks was more heterogeneous. Although two indices exceeded the a priori criterion we established for acceptable stability with only one or two observations (e.g., TBW for audiovisual speech and one variable indexing magnitude of integration in response to incongruent audiovisual McGurk stimuli), other variables would necessitate much more extensive sampling to achieve acceptable stability. We considered the possibility that restricted variance among participants could explain the relatively low stability of some variables derived from psychophysics tasks[1]. However, there was substantial variability among participants in variables derived from all measures of multisensory function employed in the present work, in accord with the extant literature. For example, on both McGurk tasks, participants showed a high degree of heterogeneity in their responses, reporting rates of perceived fusion ranging from 0% - 100%, which is consistent with prior work reporting on individual differences in integration on McGurk tasks (e.g., Basu Mallick, Magnotti, & Beauchamp, 2015). Thus, it is unlikely that a truncated range of responses could explain the relatively low stability observed for variables derived from McGurk tasks in the present report.

There are some other possible explanations for the variability observed for stability of scores derived from psychophysical measures. First, it is notable that the listening-in-noise task represents the only measure for which the exact same stimuli were not employed across observation days. The use of different wordlists was necessary to control for the possibility of

[1] We are grateful to R2 for prompting us to attend to this possible alternative explanation to our results.

children simply "learning" the words with repeated exposure. The relative instability observed for speech-in-noise variables across sessions may, however, reflect less than optimal balancing of stimuli, though wordlists were reportedly designed to be equally audible/intelligible in prior work (Picou et al., 2017).

The contrast in stability for variables derived from different McGurk tasks is somewhat more challenging to explain. These tasks technically did involve different stimuli, but the stimuli utilized in each task were highly similar consonant-vowel syllables that were spoken by the same speaker in the same manner and that have previously been shown to induce a perceptual fusion (e.g., Iarocci et al., 2010; Irwin et al., 2011). The two versions of the McGurk task differed only in the specific audiovisual stop consonant-vowel combinations employed. The acoustic features that facilitate fusion for these two incongruent canonical syllable pairs does differ. Specifically, the audio "pa" plus visual "ka" combination (commonly inducing a percept of "ta" or "ha") reflects an instance of ambiguity in voice onset time, whereas the audio "ba" plus visual "ga" combination (frequently leading to a fused percept of "da" or "tha") reflects an instance of ambiguity in the frequency of the second formant. It is notable that this acoustic distinction of consonants (e.g., second formant frequency of "ba" vs "ga") is less consistent, especially for natural speech stimuli (Ohde & Sharf, 1992). It is unclear whether this explanation could account for less reliable fusion of the "ba" and "ga" stimuli. It is clear, however, that the increased measurement error present for the auditory "ba" plus visual "ga" stimulus combination could account for the failure to replicate findings across the large extant literature that has employed these particular stimuli to investigate the magnitude of multisensory integration for audiovisual speech in children with ASD (Basu Mallick et al., 2015; Woynaroski et al., 2013).

Variables Derived from the ERP Measure are Relatively Unstable

The variables derived from our ERP task were the least stable of all indices we explored. Nevertheless, differences in relative stability across indices were apparent. Consistent with prior literature, latency variables were, on the whole, less stable than amplitude variables (Cassidy, Robertson, & O'Connell, 2012; Huffmeijer, Bakermans-Kranenburg, Alink, & van IJzendoorn, 2014) in our sample. Additionally, N1 indices were less stable than P2 indices, likely because the N1 component is still emerging and not fully consolidated in early childhood, for children with or without ASD (Espy, Molfese, Molfese, & Modglin, 2004). These findings collectively point toward a focus on one ERP variable – P2 amplitude – for future research into the neural response to multisensory speech in school age children with autism.

Limitations

The present study has clear implications for future research focused on multisensory function in children with ASD, but it is not without limitations. First, it is notable that our sample size was small. The small *n* was necessary, given the extensive and repetitive nature of measurement and was within the range of sample sizes frequently considered sufficient for estimating the amount of variance attributable to construct/s of interest versus other facets of the measurement context (i.e., 10-20 participants; Bottema-Beutel et al., 2019; Woynaroski et al., 2017; Yoder et al., 2016). However, a known ramification of small sample size, reflected in wide confidence intervals, is reduced confidence that results will represent variable stability in similar participants. Furthermore, our participant sample was limited to children who were relatively older and higher functioning (i.e., cognitively and linguistically able), specifically children who were age 7 and up and capable of completing the broad range of tasks utilized in the present study, including tasks that necessitated attending and actively responding for an extended period of time. Our participant sample is comparable to those of previous studies employing similar

measures of multisensory functioning (e.g., Basu Mallick et al., 2015; Grossman et al., 2015; Hillock et al., 2011; Iarocci et al., 2010). Future studies are needed, though, to explore the stability of variables of multisensory function in children with ASD who represent the broader range of chronological ages, developmental stages, and functioning levels. It is expected that stability of variables derived from these frequently used measures of multisensory function would likely be lower than observed here in children who are chronologically or developmentally younger (Sandbank & Yoder, 2014). Additionally, our findings may not generalize beyond the methods used in this study; for example, it is possible that a different number of trials in the psychophysical and EEG tasks or different stimuli may yield different results.

Conclusion

To our knowledge, this is the first study to comprehensively examine the stability of variables derived from commonly used measures of multisensory function in children with ASD. Results suggest that the stability of such variables is highly heterogeneous. Though a number of indices demonstrated relatively high stability, suggesting they hold some promise for detecting effects of interest as they relate to multisensory function in future research (and perhaps ultimately in clinical practice), other variables were much less stable. Thus, obtaining representative estimates of constructs of interest related to multisensory processing may require averaging scores across repeated observations or, in some cases, may not be feasible. Collectively, our findings highlight the importance of considering psychometrics in planning future studies focused on children with autism spectrum disorder and other neurodevelopmental disorders.

References

American Psychological Association.(2013). *Diagnostic and statistcal manual of mental disorders-5*. Washington.

Barenholtz, E., Mavica, L., & Lewkowicz, D. J. (2016). Language familiarity modulates relative attention to the eyes and mouth of a talker. *Cognition*, *147*, 100–105. https://doi.org/10.1016/J.COGNITION.2015.11.013

Basu Mallick, D., Magnotti, J. F., & Beauchamp, M. S. (2015). Variability and stability in the McGurk effect: contributions of participants, stimuli, time, and response type. *Psychonomic Bulletin & Review*, *22*, 1299–1307. https://doi.org/10.3758/s13423-015-0817-4

Baum, S. H., Stevenson, R. A., & Wallace, M. T. (2015). Behavioral, perceptual, and neural alterations in sensory and multisensory function in autism spectrum disorder. *Progress in Neurobiology*, *134*, 140–160. https://doi.org/10.1016/J.PNEUROBIO.2015.09.007

Bottema-Beutel, K., Kim, S. Y., Crowley, S., Augustine, A., Kecili-Kaysili, B., Feldman, J., & Woynaroski, T. (2019). The stability of joint engagement states in infant siblings of children with and without ASD: Implications for measurement practices. *Autism Research*. https://doi.org/10.1002/aur.2068

Brandwein, A. B., Foxe, J. J., Butler, J. S., Russo, N. N., Altschuler, T. S., Gomes, H., & Molholm, S. (2013). The development of multisensory integration in high-functioning autism: High-density electrical mapping and psychophysical measures reveal impairments in the processing of audiovisual inputs. *Cerebral Cortex*, *23*, 1329–1341. https://doi.org/10.1093/cercor/bhs109

Brandwein, Alice B., Foxe, J. J., Butler, J. S., Frey, H.-P., Bates, J. C., Shulman, L. H., & Molholm, S. (2015). Neurophysiological indices of atypical auditory processing and

multisensory integration are associated with symptom severity in autism. *Journal of Autism and Developmental Disorders*, *45*, 230–244. https://doi.org/10.1007/s10803-014-2212-9

Buchan, J. N., Paré, M., & Munhall, K. G. (2008). The effect of varying talker identity and listening conditions on gaze behavior during audiovisual speech perception. *Brain Research*, *1242*, 162–171. https://doi.org/10.1016/J.BRAINRES.2008.06.083

Calvert, G. A., & Campbell, R. (2003). Reading speech from still and moving faces: The neural substrates of visible speech. *Journal of Cognitive Neuroscience*, *15*, 57–70. https://doi.org/10.1162/089892903321107828

Cardinet, J., Johnson, S., & Pini, G. (2010). *Applying Generalizability Theory Using EduG*. New York: Taylor & Francis Group.

Cassidy, S. M., Robertson, I. H., & O'Connell, R. G. (2012). Retest reliability of event-related potentials: Evidence from a variety of paradigms. *Psychophysiology*, *49*, 659–664. https://doi.org/10.1111/j.1469-8986.2011.01349.x

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Computational Biology*, *5*, e1000436. https://doi.org/10.1371/journal.pcbi.1000436

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*, 9–21. https://doi.org/10.1016/J.JNEUMETH.2003.10.009

DeVellis, R. F. (2006). Classical test theory. *Medical Care*, *44*, S50–S59.

Donley, J., Ritz, C., & Kleijn, W. B. (2018). Multizone soundfield reproduction with privacy- and quality-based speech masking filters. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(6), 1041–1055. https://doi.org/10.1109/TASLP.2018.2798804

Espy, K. A., Molfese, D. L., Molfese, V. J., & Modglin, A. (2004). Development of auditory event-related potentials in young children and relations to word-level reading abilities at age 8 years. *Annals of Dyslexia*, *54*, 9–38. https://doi.org/10.1007/s11881-004-0002-3

Feldman, J. I., Dunham, K., Cassidy, M., Wallace, M. T., Liu, Y., & Woynaroski, T. G. (2018). Audiovisual multisensory integration in individuals with autism spectrum disorder: A systematic review and meta-analysis. *Neuroscience and Biobehavioral Reviews*, *95*, 220–234. https://doi.org/10.1016/j.neubiorev.2018.09.020

Foxe, J. J., Molholm, S., Del Bene, V. A., Frey, H.-P., Russo, N. N., Blanco, D., … Ross, L. A. (2015). Severe multisensory speech integration deficits in high-functioning school-aged children with autism spectrum disorder (ASD) and their resolution during early adolescence. *Cerebral Cortex*, *25*, 298–312. https://doi.org/10.1093/cercor/bht213

Fraser, S., Gagné, J.-P., Alepins, M., & Dubois, P. (2010). Evaluating the effort expended to understand speech in noise using a dual-task paradigm: The effects of providing visual speech cues. *Journal of Speech, Language, and Hearing Research*, *53*(1), 18–33. https://doi.org/10.1044/1092-4388(2009/08-0140)

Grossman, R. B., Steinhart, E., Mitchell, T., & McIlvane, W. (2015). “Look who’s talking!” gaze patterns for implicit and explicit audio-visual speech synchrony detection in children with high-functioning autism. *Autism Research*, *8*, 307–316. https://doi.org/10.1002/aur.1447

Hillock-Dunn, A., Grantham, D. W., & Wallace, M. T. (2016). The temporal binding window for audiovisual speech: Children are like little adults. *Neuropsychologia*, *88*, 74–82. https://doi.org/10.1016/J.NEUROPSYCHOLOGIA.2016.02.017

Hillock, A. R., Powers, A. R., & Wallace, M. T. (2011). Binding of sights and sounds: Age-related changes in multisensory temporal processing. *Neuropsychologia*, *49*, 461–467.

https://doi.org/10.1016/J.NEUROPSYCHOLOGIA.2010.11.041

Huffmeijer, R., Bakermans-Kranenburg, M. J., Alink, L. R. A., & van IJzendoorn, M. H. (2014). Reliability of event-related potentials: The influence of number of trials and electrodes. *Physiology & Behavior*, *130*, 13–22. https://doi.org/10.1016/J.PHYSBEH.2014.03.008

Iarocci, G., Rombough, A., Yager, J., Weeks, D. J., & Chua, R. (2010). Visual influences on speech perception in children with autism. *Autism*, *14*, 1362–3613. https://doi.org/10.1177/1362361309353615

Irwin, J., & DiBlasi, L. (2017). Audiovisual speech perception: A new approach and implications for clinical populations. *Language and Linguistics Compass*, *11*, 77–91. https://doi.org/10.1111/lnc3.12237

Irwin, J. R., Tornatore, L. A., Brancazio, L., & Whalen, D. H. (2011). Can children with autism spectrum disorders "hear" a speaking face? *Child Development*, *82*, 1397–1403. https://doi.org/10.1111/j.1467-8624.2011.01619.x

Klin, A., Jones, W., Schultz, R., Volkmar, F., & Cohen, D. (2002). Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. *Archives of General Psychiatry*, *59*, 809–816. https://doi.org/10.1001/archpsyc.59.9.809

Knowland, V. C. P., Mercure, E., Karmiloff-Smith, A., Dick, F., & Thomas, M. S. C. (2014). Audio-visual speech perception: A developmental ERP investigation. *Developmental Science*, *17*, 110–124. https://doi.org/10.1111/desc.12098

Lewkowicz, D. J., & Flom, R. (2014). The audiovisual temporal binding window narrows in early childhood. *Child Development*, *85*(2), 685–694. https://doi.org/10.1111/cdev.12142

Lewkowicz, D. J., & Hansen-Tift, A. M. (2012). Infants deploy selective attention to the mouth

of a talking face when learning speech. *Proceedings of the National Academy of Sciences of the United States of America*, *109*, 1431–1436. https://doi.org/10.1073/pnas.1114783109

Lewkowicz, D. J., Minar, N. J., Tift, A. H., & Brandon, M. (2015). Perception of the multisensory coherence of fluent audiovisual speech in infancy: Its emergence and the role of experience. *Journal of Experimental Child Psychology*, *130*, 147–162. https://doi.org/10.1016/J.JECP.2014.10.006

Lord, C., Rutter, M., DiLavore, P., Risi, S., Gotham, K., & Bishop, S. L. (2012). *Autism Diagnostic Observation Schedule, second edition*. Los Angeles, CA: Western Psychological Services.

Magnotti, J. F., & Beauchamp, M. S. (2018). Published estimates of group differences in multisensory integration are inflated. *PLOS ONE*, *13*(9), e0202908. https://doi.org/10.1371/journal.pone.0202908

Massaro, D. W., & Palmer, S. E. (1998). *Perceiving talking faces: From speech perception to a behavioral principle* (Vol. 1). Mit Press.

Mccrae, R. R., Kurtz, J. E., Yamagata, S., & Terracciano, A. (2011). Internal consistency, retest reliability, and their implications for personality scale validity. *Personality and Social Psychology Review*, *15*, 28–50. https://doi.org/10.1177/1088868310366253

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746–748. https://doi.org/10.1038/264746a0

Mushquash, C., & O'Connor, B. P. (2006). SPSS and SAS programs for generalizability analysis. *Behavior Research Methods*, *38*, 542–547. Retrieved from www.education.uiowa.edu/casma/GenovaPrograms.htm.

Nunnally, J. (1978). *Psychometric Theory* (2nd ed.). New York: McGraw-Hill.

Ohde, R. N., & Sharf, D. J. (1992). *Phonetic Analysis of Normal and Abnormal Speech*. Ann Arbor, Michigan: Allyn and Bacon.

Picou, E. M., Charles, L. M., & Ricketts, T. A. (2017). Child–adult differences in using dual-task paradigms to measure listening effort. *American Journal of Audiology, 26*, 143–154. https://doi.org/10.1044/2016_AJA-16-0059

Picou, E. M., Ricketts, T. A., & Hornsby, B. W. Y. (2011). Visual cues and listening effort: Individual variability. *Journal of Speech, Language, and Hearing Research, 54*, 1416–1430. https://doi.org/10.1044/1092-4388(2011/10-0154)

Pons, F., Bosch, L., & Lewkowicz, D. J. (2019). Twelve-month-old infants' attention to the eyes of a talking face is associated with communication and social skills. *Infant Behavior and Development, 54*, 80–84. https://doi.org/10.1016/J.INFBEH.2018.12.003

Powers, A. R., Hillock, A. R., & Wallace, M. T. (2009). Perceptual training narrows the temporal window of multisensory binding. *Journal of Neuroscience, 29*, 12265–12274. https://doi.org/10.1523/JNEUROSCI.3501-09.2009

Riby, D., & Hancock, P. J. B. (2009). Looking at movies and cartoons: eye-tracking evidence from Williams syndrome and autism. *Journal of Intellectual Disability Research*, *53*, 169–181. https://doi.org/10.1111/j.1365-2788.2008.01142.x

Roid, G. H., Miller, L. J., Pomplun, M., & Koch, C. (2013). *Leiter International Performance Scale* (3rd ed.). Torrance, CA: Western Psychological Services.

Sandbank, M., & Yoder, P. (2014). Measuring representative communication in young children with developmental delay. *Topics in Early Childhood Special Education, 34*, 133–141. https://doi.org/10.1177/0271121414528052

Smith, E. G., & Bennetto, L. (2007). Audiovisual speech integration and lipreading in autism.

Journal of Child Psychology and Psychiatry, *48*, 813–821. https://doi.org/10.1111/j.1469-7610.2007.01766.x

Smith, E., Zhang, S., & Bennetto, L. (2017). Temporal synchrony and audiovisual integration of speech and object stimuli in autism. *Research in Autism Spectrum Disorders*, *39*, 11–19. https://doi.org/10.1016/J.RASD.2017.04.001

Soto-Faraco, S., Calabresi, M., Navarra, J., Werker, J., & Lewkowicz, D. J. (2012). Multisensory development. In *Multisensory Development* (pp. 207–228).

Stevenson, R. A., Siemann, J. K., Schneider, B. C., Eberly, H. E., Woynaroski, T. G., Camarata, S. M., & Wallace, M. T. (2014). Multisensory temporal integration in autism spectrum disorders. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *34*, 691–697. https://doi.org/10.1523/JNEUROSCI.3615-13.2014

Swiss Society for Research in Education Working Group. (2012). *EduG*. Retrieved from https://www.irdp.ch/institut/english-program-1968.html

Tenenbaum, E. J., Amso, D., Abar, B., & Sheinkopf, S. J. (2014). Attention and word learning in autistic, language delayed and typically developing children. *Frontiers in Psychology*, *5*, 490. https://doi.org/10.3389/fpsyg.2014.00490

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 1181–1186. https://doi.org/10.1073/pnas.0408949102

Williams, J. H. G., Massaro, D. W., Peel, N. J., Bosseler, A., & Suddendorf, T. (2004). Visual–auditory integration during speech imitation in autism. *Research in Developmental Disabilities*, *25*, 559–575. https://doi.org/10.1016/J.RIDD.2004.01.008

Woynaroski, T. G., Kwakye, L. D., Foss-Feig, J. H., Stevenson, R. A., Stone, W. L., & Wallace,

M. T. (2013). Multisensory speech perception in children with autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *43*(12), 2891–2902. https://doi.org/10.1007/s10803-013-1836-5

Woynaroski, T., Oller, D. K., Keceli-Kaysili, B., Xu, D., Richards, J. A., Gilkerson, J., … Yoder, P. (2017). The stability and validity of automated vocal analysis in preverbal preschoolers with autism spectrum disorder. *Autism Research*, *10*, 508–519. https://doi.org/10.1002/aur.1667

Yoder, P. J., Loyd, B. P., & Symons, F. J. (2018). *Observational Measurement of Behavior* (2nd ed.). Baltimore: Paul H. Brookes Publishing Co.

Yoder, P. J., Woynaroski, T., & Camarata, S. (2016). Measuring speech comprehensibility in students with down syndrome. *Journal of Speech, Language, and Hearing Research*, *59*, 460–467. https://doi.org/10.1044/2015_JSLHR-S-15-0149

Table 1

Description of Participant Characteristics

	M (*SD*)	Range
Age (Years)	10.68 (2.81)	7.52 – 16.00
Sex	7 male, 4 female	
Nonverbal IQ	110.5 (10.0)	90 – 126
Expressive Language	108.9 (12.8)	88 – 124
Receptive Language	112.5 (17.9)	87 - 136

Note. Nonverbal IQ, receptive language, and expressive language are indexed by standard scores. Nonverbal IQ was measured by the Leiter International Performance Scale, 3rd edition (Leiter-3; Roid et al., 2013). Expressive language was measured by the Expressive One Word Picture Vocabulary Test. Receptive language was measured by the Receptive One Word Picture Vocabulary Test.

Table 2

Summary of Variables Tested in Generalizability (G) and Decision (D) Analyses

Variable Label	Precise Operational Definition of Variable
Variables Derived from Psychophysical Measures	
Temporal Binding Window for Audiovisual Speech	The difference (in ms) between the points where two psychometric curves fit to data for the proportion of reported synchrony across SOAs cross 0.75 (see Figure 1)
Proportion of Fusions in Response to McGurk Ba/Ga Stimuli	Proportion of “da” and “tha” responses to the number of mismatched audiovisual (i.e., auditory “ba” + visual “ga”) trials
Proportion of Fusions in Response to McGurk Pa/Ka Stimuli	Proportion of “ta” and “ha” responses to the number of mismatched audiovisual (i.e., auditory “pa” + visual “ka”) trials
Audiovisual Word Recognition Identification Accuracy -3dB SNR	Number of whole words correctly identified during the speech-in-noise task in the audiovisual condition with a -3dB SNR
Audiovisual Word Recognition Identification Accuracy -6dB SNR	Number of whole words correctly identified during the speech-in-noise task in the audiovisual condition with a -6dB SNR
Variables Derived from Event Related Potential (ERP) Measure	

N1 Amplitude	The average amplitude of the grand-average waveform in response to AO and AV stimuli between 100 ms and 140 ms post-stimulus onset
N1 Latency	Length of time, in ms, for the grand-average waveform in response to AO and AV stimuli to reach the maximum amplitude between 100 ms and 140 ms post-stimulus onset
P2 Amplitude	The average amplitude of the grand-average waveform in response to AO and AV stimuli between 160 ms and 240 ms post-stimulus onset
P2 Latency	Length of time, in ms, for the grand-average waveform in response to AO and AV stimuli to reach the maximum amplitude between 160 ms and 240 ms post-stimulus onset
Variables Derived from Eye Tracking Measure	
Proportion of Total Looking Time to the Eyes	Proportion of time looking to the eyes AOI of total time looking to the face AOI in each condition (i.e., English ID, Spanish ID, English AD, Spanish AD)
Proportion of Total Looking Time to the Mouth	Proportion of time looking to the mouth AOI of total time looking to the face AOI during each condition (i.e., English ID, Spanish ID, English AD, Spanish AD)

Note. AD = adult-directed; AO = auditory-only; AV = audiovisual; AOI = area of interest; ID = infant-directed; SNR = signal-to-noise ratio; SOA = stimulus onset asynchrony.

Table 3

Absolute G Coefficients by Variable and Number of Samples Across Which Scores are Averaged

	# Samples Across Which Scores are Averaged					# Samples Required for $g \geq .8$
Variable	1	2	3	4	5	
Spanish AD: Look to Eyes	**.959**	**.979**	**.986**	**.990**	**.992**	1
Spanish AD: Look to Mouth	**.949**	**.974**	**.982**	**.987**	**.989**	1
Spanish ID: Look to Eyes	**.931**	**.964**	**.976**	**.982**	**.985**	1
English AD: Look to Mouth	**.905**	**.950**	**.966**	**.974**	**.979**	1
AV P2 Amplitude	**.896**	**.945**	**.963**	**.972**	**.977**	1
English AD: Look to Eyes	**.869**	**.930**	**.952**	**.964**	**.971**	1
McGurk Pa/Ka	**.844**	**.916**	**.942**	**.956**	**.964**	1
English ID: Look to Eyes	.754	**.860**	**.902**	**.925**	**.939**	2
AO P2 Amplitude	.744	**.853**	**.897**	**.921**	**.936**	2
TBW for Audiovisual Speech	.741	**.852**	**.896**	**.920**	**.935**	2
Spanish ID: Look to Mouth	.740	**.850**	**.895**	**.919**	**.934**	2
English ID: Look to Mouth	.697	**.822**	**.874**	**.902**	**.920**	2
McGurk Ba/Ga	.471	.640	.727	.780	**.816**	5
AV N1 Amplitude	.412	.584	.678	.737	.778	6
AV Word Recognition -3dB SNR	.310	.474	.575	.643	.692	9
AO P2 Latency	.243	.391	.490	.562	.616	NA
AV Word Recognition -6dB SNR	.187	.316	.409	.480	.535	NA
AV N1 Latency	.107	.194	.265	.325	.376	NA
AO N1 Latency	.047	.089	.128	.164	.197	NA
AO N1 Amplitude	.000	.000	.000	.000	.000	NA
AV P2 Latency	.000	.000	.000	.000	.000	NA

Note. TBW = Temporal binding window, AV Word Recognition = Whole word identification accuracy in the audiovisual condition of the speech-in-noise task at -3 or -6 dB signal-to-noise ratio (SNR), AO = Auditory-only condition of ERP task, AV = Audiovisual condition of ERP task, N1 = timeframe between 100 ms and 140 ms post-stimulus onset, P2 = timeframe between 160 ms and 240 ms, ID = Infant-directed speech, AD = Adult-directed speech, NA = Not applicable - *g* coefficients do not converge on acceptable stability even for estimated coefficients at 10 or more observations. See Table 2 for precise operational definitions of all variables. Bolded values are those that exceed our a priori stability criterion of $g = 0.8$. Underlined and bolded values reflect values that exceed the criterion at the lowest number of observations for that variable.

Figures

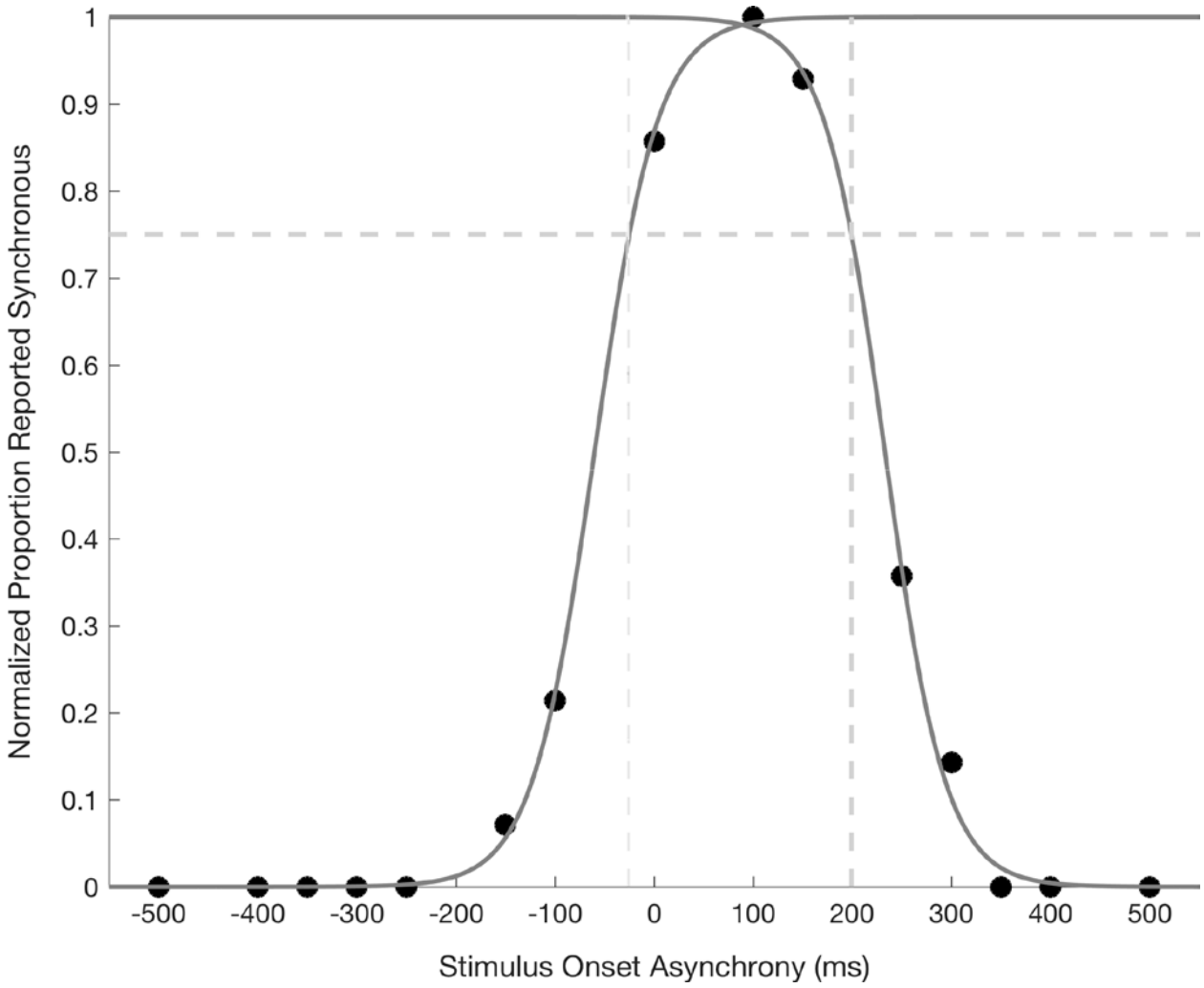

Figure 1. Representative temporal binding window (TBW) derived for a participant in the study from a single sample. The proportions of perceived synchrony are normalized, such that the maximum value is set to 1, and are fit to two psychometric functions, one for trials wherein visual stimuli precede the auditory stimuli (right line) and one for trials wherein auditory stimuli precede the visual stimuli (left line). The vertical dotted lines represent the point at which each line reaches the .75 threshold for perceived synchrony (the horizontal dotted line; i.e., –25.0 ms and 200.0 ms for the depicted example). The TBW is the distance between these two values (the distance between the two vertical dotted lines; i.e., 225.0 ms).

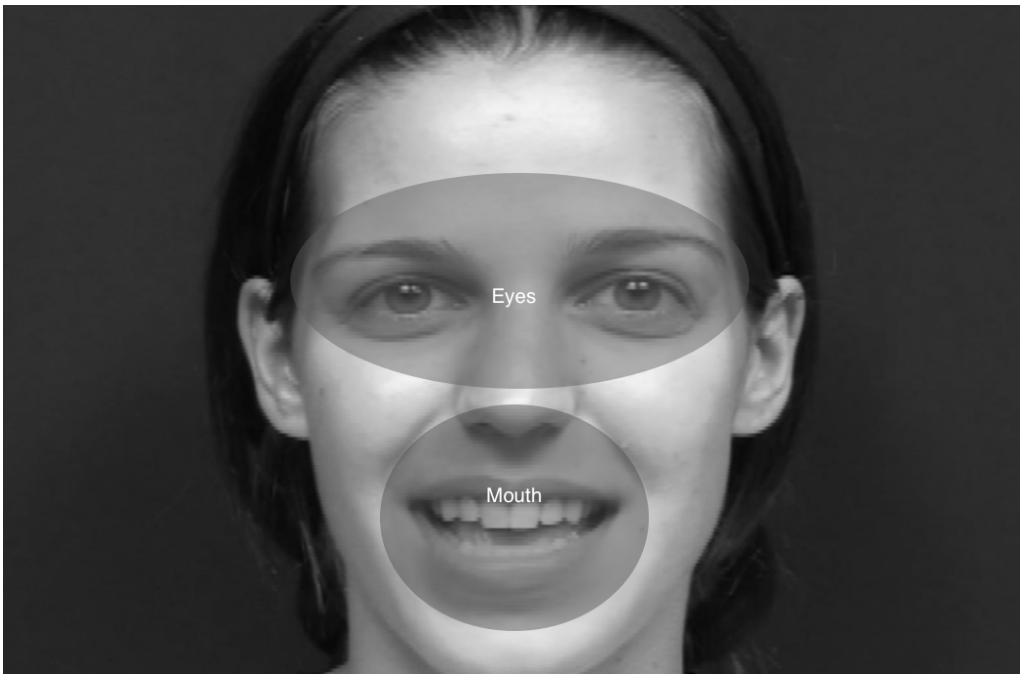

Figure 2. Example of the areas of interest (AOIs) used in the eye tracking measure (i.e., in the English infant-directed multisensory speech condition). Proportion of time looking to each area of interest was calculated as the time spent looking at the area (mouth or eyes)/time spent looking at the broader face during stimulus presentation.

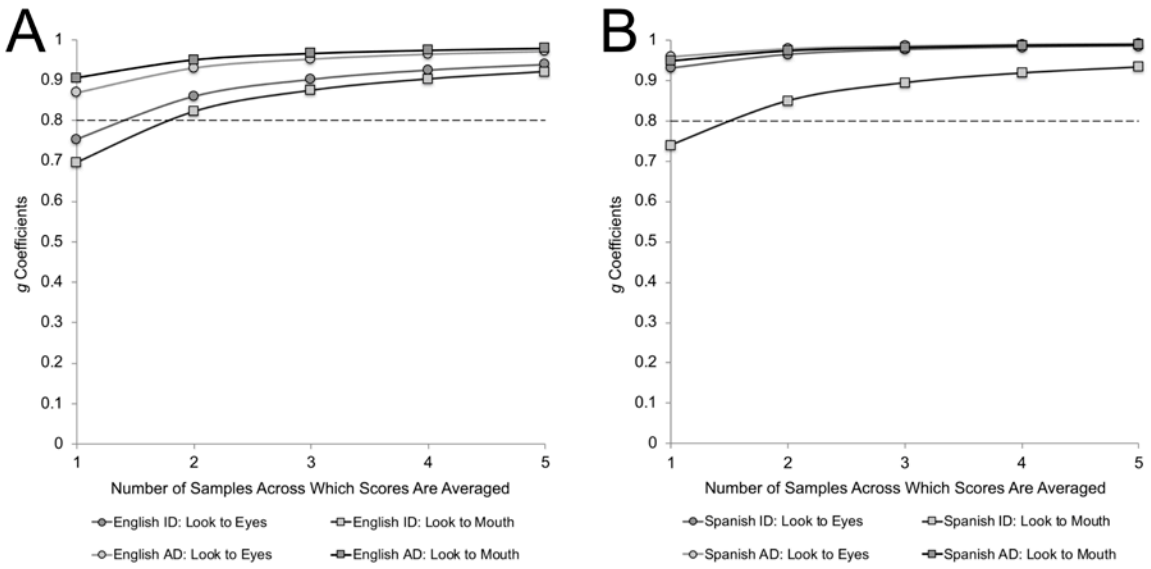

Figure 3. Stability of variables derived from eye tracking measure of attention to multisensory speech presented in (A) English and (B) Spanish. Variables derived from eye tracking have high stability ($g > 0.8$) with one to two observations. ID = Infant-directed speech; AD = Adult-directed speech. Generalizability coefficients are observed for two samples (i.e., day 1 and 2) and projected beyond two samples based on variance estimates derived via ANOVA carried out on observed data. The a priori threshold for acceptable stability was set at $g = .8$ (dashed line).

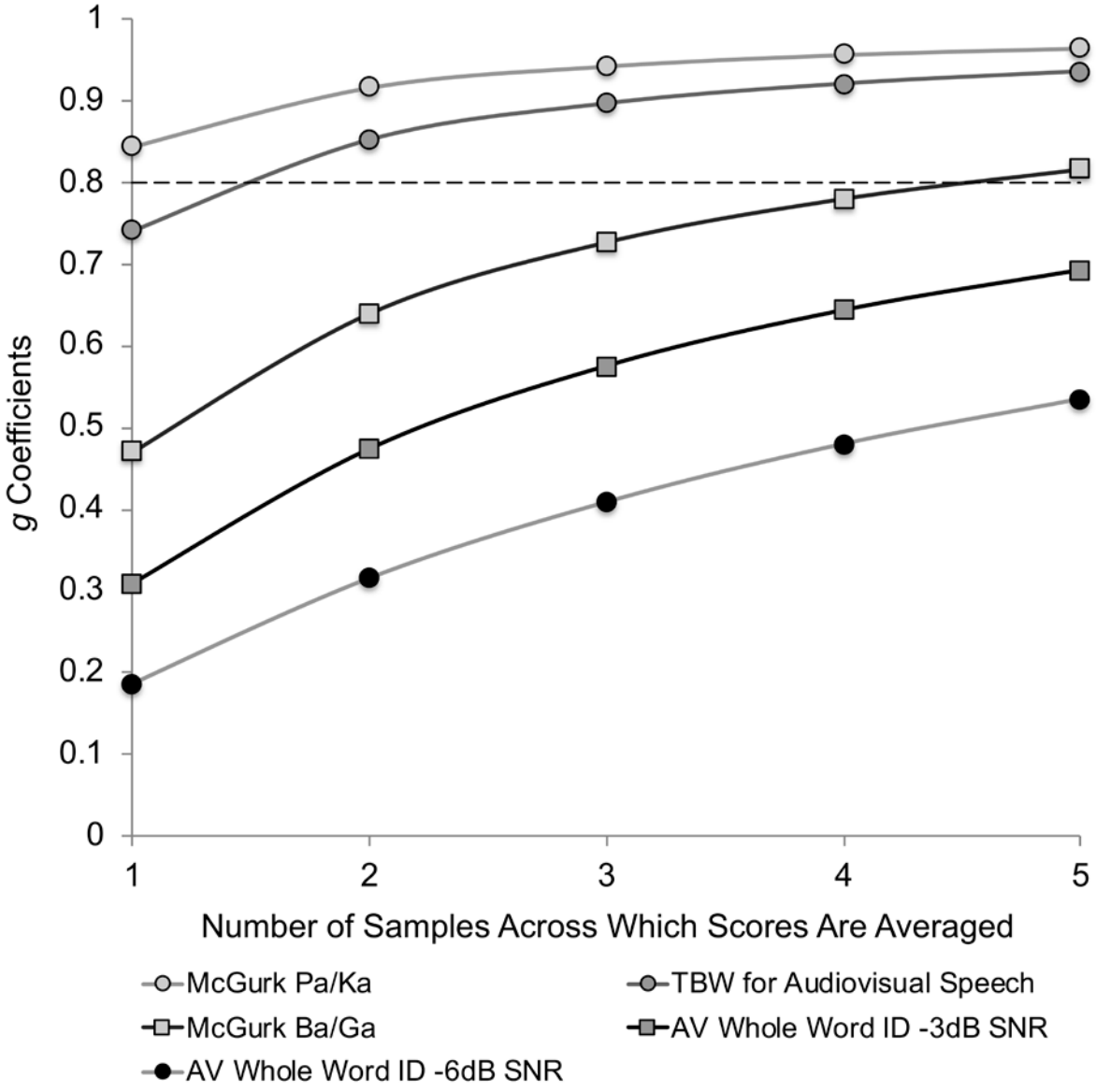

Figure 4. Stability of variables derived from psychophysics measures. Two variables derived from psychophysical measures met our a priori threshold for stability of $g = 0.8$ within two observations. AV = audiovisual; ID = identification accuracy; TBW = temporal binding window. Generalizability coefficients are observed for two samples (i.e., day 1 and 2) and projected beyond two samples based on variance estimates derived via ANOVA on observed data. The a priori threshold for acceptable stability was set at $g = .8$ (dashed line).

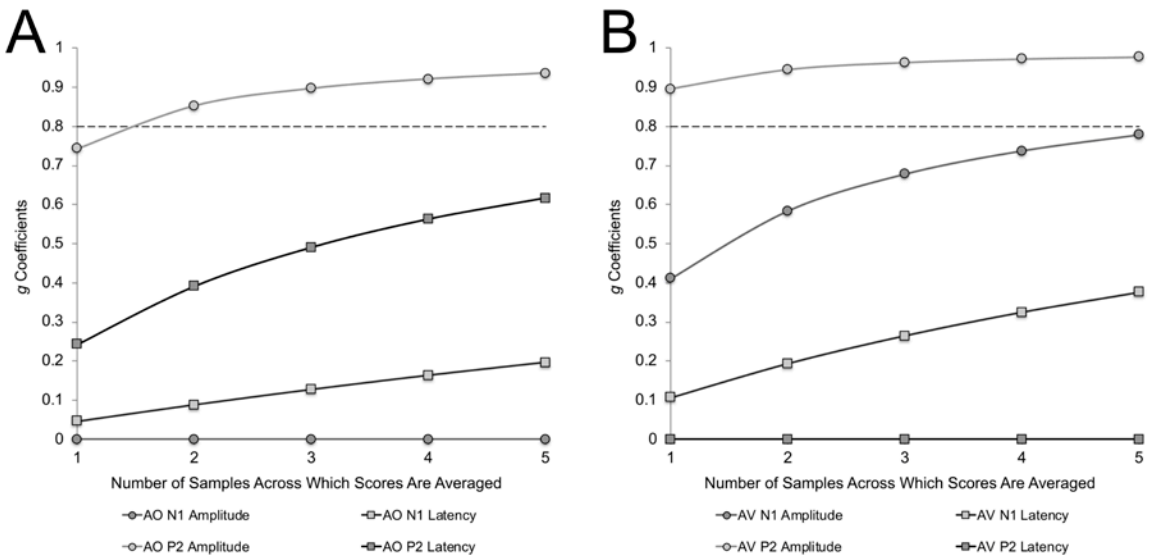

Figure 5. Stability of variables derived from ERP measure. P2 amplitude is the most stable of the ERP variables in both the (A) auditory-only (AO) condition and (B) audiovisual (AV) conditions. N1 window = 100 ms - 140 ms post-stimulus onset; P2 window = 160 ms - 240 ms post-stimulus onset. Generalizability coefficients are observed for two samples (i.e., day 1 and 2) and projected beyond two samples based on variance estimates derived via ANOVA carried out on observed data. The a priori threshold for acceptable stability was set at g = .8 (dashed line). Note that the consistent g coefficients of 0 for AO N1 amplitude and AV P2 latency suggest that it is not possible to obtain stable estimates for these variables in school age children with ASD even with repeated sampling.

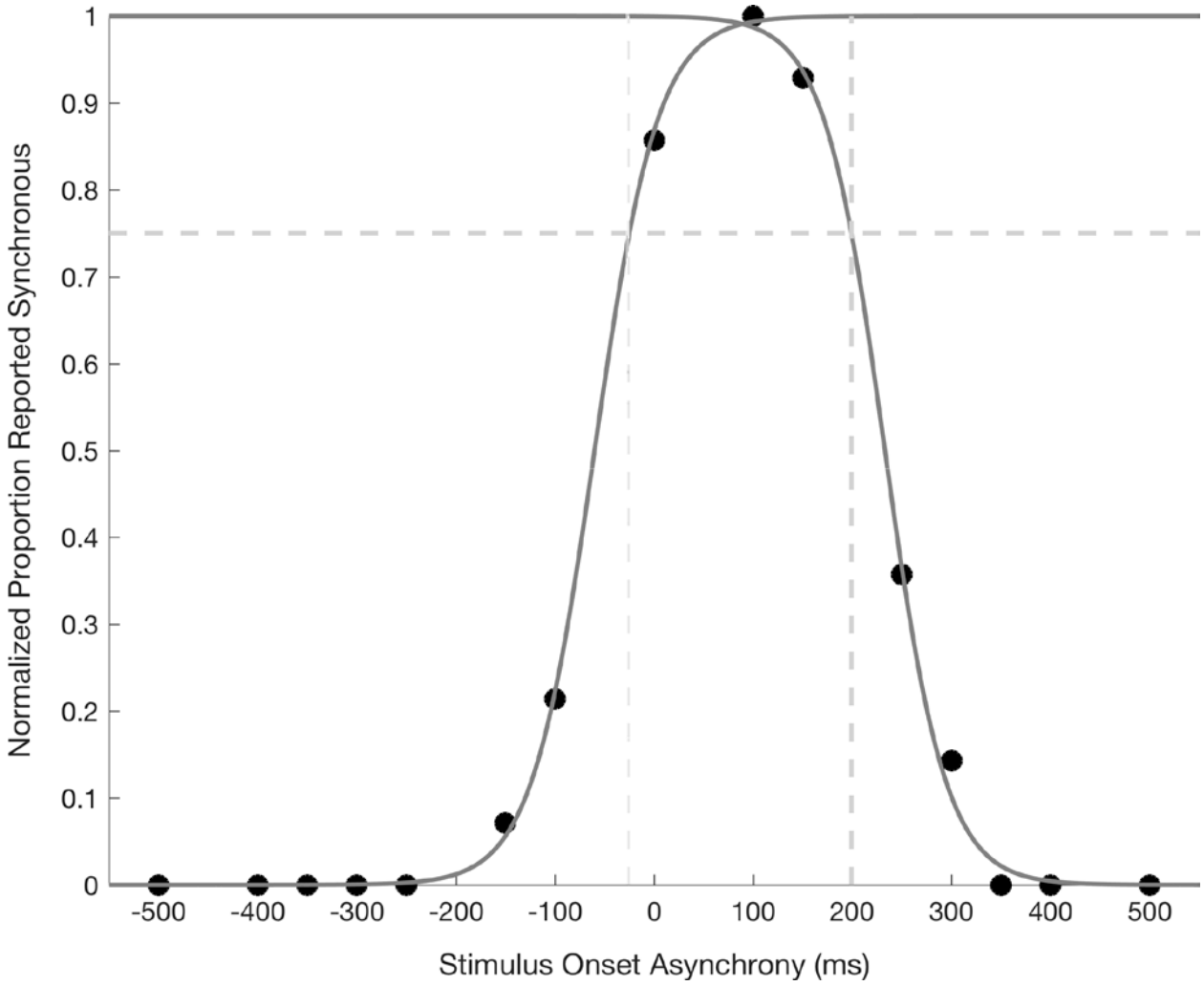

Figure 1. Representative temporal binding window (TBW) derived for a participant in the study from a single sample. The proportions of perceived synchrony are normalized, such that the maximum value is set to 1, and are fit to two psychometric functions, one for trials wherein visual stimuli precede the auditory stimuli (right line) and one for trials wherein auditory stimuli precede the visual stimuli (left line). The vertical dotted lines represent the point at which each line reaches the .75 threshold for perceived synchrony (the horizontal dotted line; i.e., –25.0 ms and 200.0 ms for the depicted example). The TBW is the distance between these two values (the distance between the two vertical dotted lines; i.e., 225.0 ms).

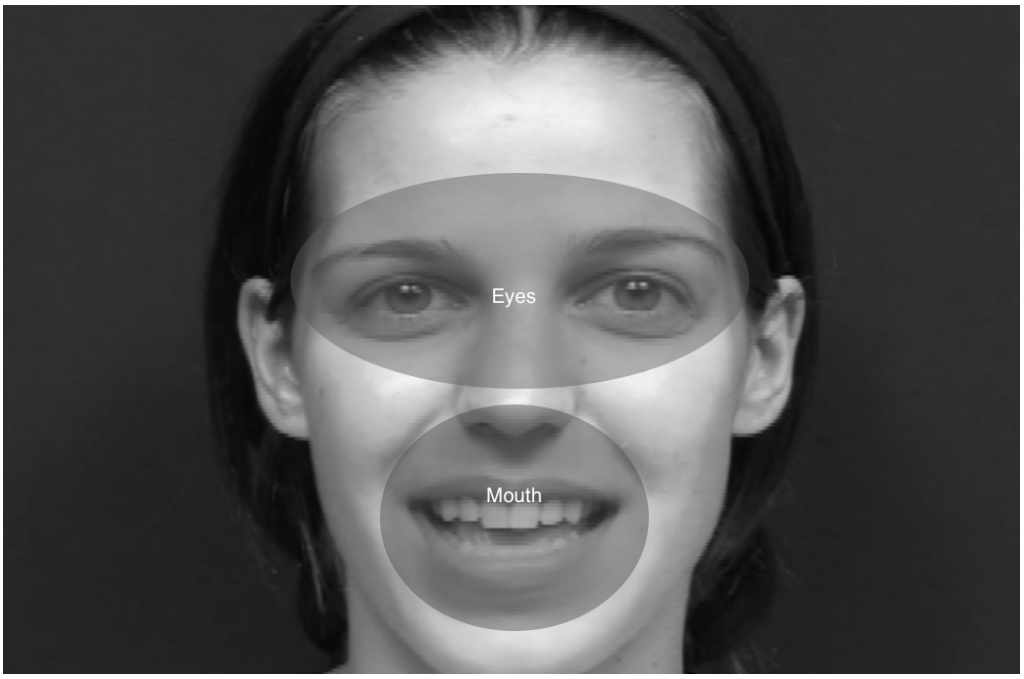

Figure 2. Example of the areas of interest (AOIs) used in the eye tracking measure (i.e., in the English infant-directed multisensory speech condition). Proportion of time looking to each area of interest was calculated as the time spent looking at the area (mouth or eyes)/time spent looking at the broader face during stimulus presentation.

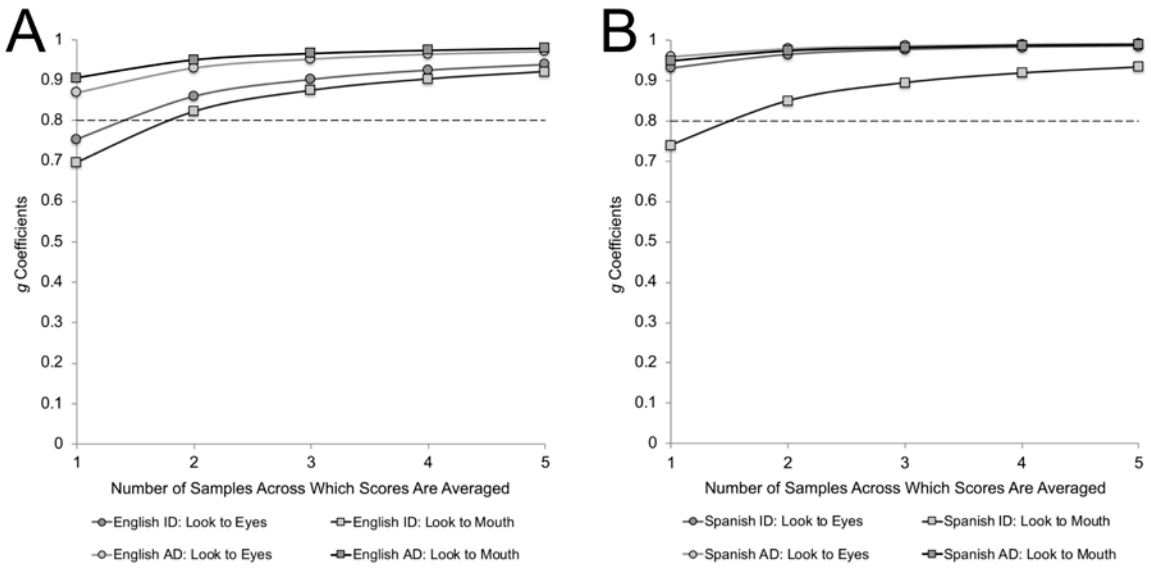

Figure 3. Stability of variables derived from eye tracking measure of attention to multisensory speech presented in (A) English and (B) Spanish. Variables derived from eye tracking have high stability ($g > 0.8$) with one to two observations. ID = Infant-directed speech; AD = Adult-directed speech. Generalizability coefficients are observed for two samples (i.e., day 1 and 2) and projected beyond two samples based on variance estimates derived via ANOVA carried out on observed data. The a priori threshold for acceptable stability was set at $g = .8$ (dashed line).

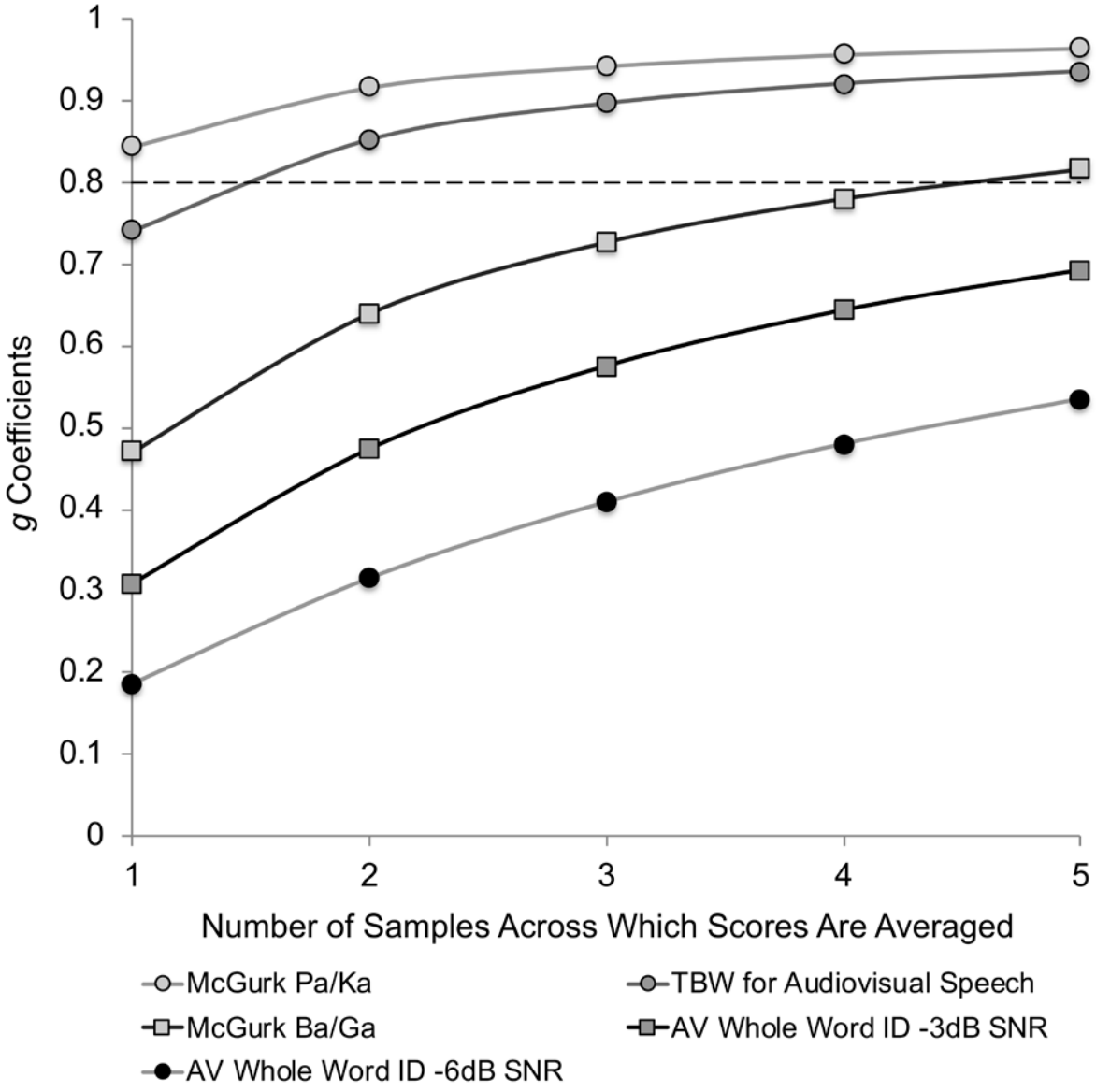

Figure 4. Stability of variables derived from psychophysics measures. Two variables derived from psychophysical measures met our a priori threshold for stability of $g = 0.8$ within two observations. AV = audiovisual; ID = identification accuracy; TBW = temporal binding window. Generalizability coefficients are observed for two samples (i.e., day 1 and 2) and projected beyond two samples based on variance estimates derived via ANOVA on observed data. The a priori threshold for acceptable stability was set at $g = .8$ (dashed line).

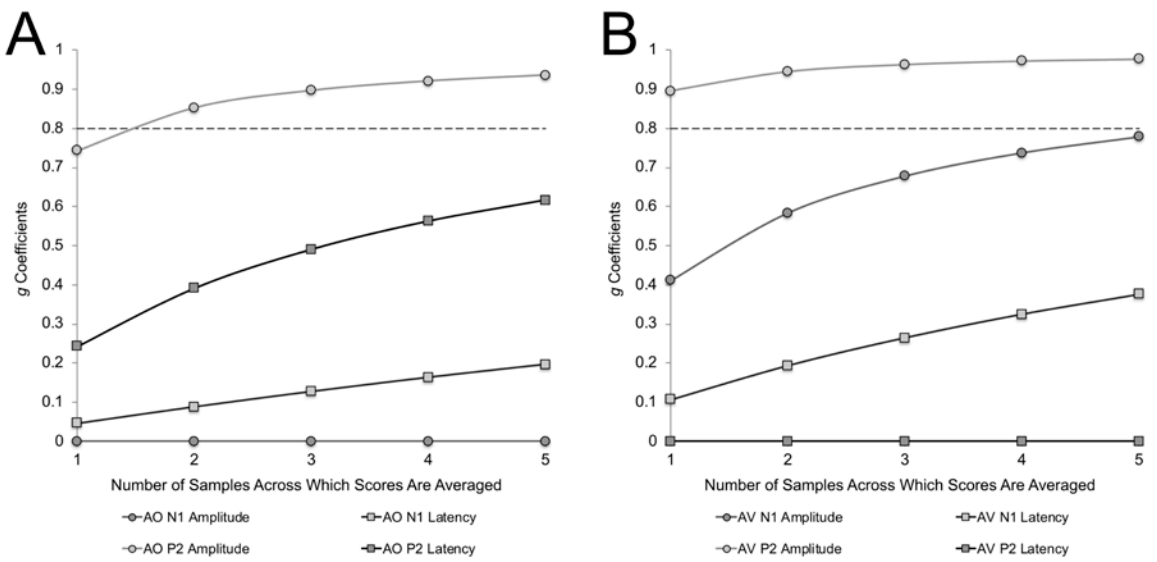

Figure 5. Stability of variables derived from ERP measure. P2 amplitude is the most stable of the ERP variables in both the (A) auditory-only (AO) condition and (B) audiovisual (AV) conditions. N1 window = 100 ms - 140 ms post-stimulus onset; P2 window = 160 ms - 240 ms post-stimulus onset. Generalizability coefficients are observed for two samples (i.e., day 1 and 2) and projected beyond two samples based on variance estimates derived via ANOVA carried out on observed data. The a priori threshold for acceptable stability was set at g = .8 (dashed line). Note that the consistent g coefficients of 0 for AO N1 amplitude and AV P2 latency suggest that it is not possible to obtain stable estimates for these variables in school age children with ASD even with repeated sampling.

Table 1

Description of Participant Characteristics

	M (*SD*)	Range
Age (Years)	10.68 (2.81)	7.52 – 16.00
Sex	7 male, 4 female	
Nonverbal IQ	110.5 (10.0)	90 – 126
Expressive Language	108.9 (12.8)	88 – 124
Receptive Language	112.5 (17.9)	87 - 136

Note. Nonverbal IQ, receptive language, and expressive language are indexed by standard scores. Nonverbal IQ was measured by the Leiter International Performance Scale, 3rd edition (Leiter-3; Roid et al., 2013). Expressive language was measured by the Expressive One Word Picture Vocabulary Test. Receptive language was measured by the Receptive One Word Picture Vocabulary Test.

Table 2

Summary of Variables Tested in Generalizability (G) and Decision (D) Analyses

Variable Label	Precise Operational Definition of Variable
Variables Derived from Psychophysical Measures	
Temporal Binding Window for Audiovisual Speech	The difference (in ms) between the points where two psychometric curves fit to data for the proportion of reported synchrony across SOAs cross 0.75 (see Figure 1)
Proportion of Fusions in Response to McGurk Ba/Ga Stimuli	Proportion of “da” and “tha” responses to the number of mismatched audiovisual (i.e., auditory “ba” + visual “ga”) trials
Proportion of Fusions in Response to McGurk Pa/Ka Stimuli	Proportion of “ta” and “ha” responses to the number of mismatched audiovisual (i.e., auditory “pa” + visual “ka”) trials
Audiovisual Word Recognition Identification Accuracy -3dB SNR	Number of whole words correctly identified during the speech-in-noise task in the audiovisual condition with a -3dB SNR
Audiovisual Word Recognition Identification Accuracy -6dB SNR	Number of whole words correctly identified during the speech-in-noise task in the audiovisual condition with a -6dB SNR
Variables Derived from Event Related Potential (ERP) Measure	

N1 Amplitude	The average amplitude of the grand-average waveform in response to AO and AV stimuli between 100 ms and 140 ms post-stimulus onset
N1 Latency	Length of time, in ms, for the grand-average waveform in response to AO and AV stimuli to reach the maximum amplitude between 100 ms and 140 ms post-stimulus onset
P2 Amplitude	The average amplitude of the grand-average waveform in response to AO and AV stimuli between 160 ms and 240 ms post-stimulus onset
P2 Latency	Length of time, in ms, for the grand-average waveform in response to AO and AV stimuli to reach the maximum amplitude between 160 ms and 240 ms post-stimulus onset
Variables Derived from Eye Tracking Measure	
Proportion of Total Looking Time to the Eyes	Proportion of time looking to the eyes AOI of total time looking to the face AOI in each condition (i.e., English ID, Spanish ID, English AD, Spanish AD)
Proportion of Total Looking Time to the Mouth	Proportion of time looking to the mouth AOI of total time looking to the face AOI during each condition (i.e., English ID, Spanish ID, English AD, Spanish AD)

Note. AD = adult-directed; AO = auditory-only; AV = audiovisual; AOI = area of interest; ID = infant-directed; SNR = signal-to-noise ratio; SOA = stimulus onset asynchrony.

Table 3

Absolute G Coefficients by Variable and Number of Samples Across Which Scores are Averaged

	# Samples Across Which Scores are Averaged					# Samples Required for $g \geq .8$
Variable	1	2	3	4	5	
Spanish AD: Look to Eyes	**.959**	**.979**	**.986**	**.990**	**.992**	1
Spanish AD: Look to Mouth	**.949**	**.974**	**.982**	**.987**	**.989**	1
Spanish ID: Look to Eyes	**.931**	**.964**	**.976**	**.982**	**.985**	1
English AD: Look to Mouth	**.905**	**.950**	**.966**	**.974**	**.979**	1
AV P2 Amplitude	**.896**	**.945**	**.963**	**.972**	**.977**	1
English AD: Look to Eyes	**.869**	**.930**	**.952**	**.964**	**.971**	1
McGurk Pa/Ka	**.844**	**.916**	**.942**	**.956**	**.964**	1
English ID: Look to Eyes	.754	**.860**	**.902**	**.925**	**.939**	2
AO P2 Amplitude	.744	**.853**	**.897**	**.921**	**.936**	2
TBW for Audiovisual Speech	.741	**.852**	**.896**	**.920**	**.935**	2
Spanish ID: Look to Mouth	.740	**.850**	**.895**	**.919**	**.934**	2
English ID: Look to Mouth	.697	**.822**	**.874**	**.902**	**.920**	2
McGurk Ba/Ga	.471	.640	.727	.780	**.816**	5
AV N1 Amplitude	.412	.584	.678	.737	.778	6
AV Word Recognition -3dB SNR	.310	.474	.575	.643	.692	9
AO P2 Latency	.243	.391	.490	.562	.616	NA
AV Word Recognition -6dB SNR	.187	.316	.409	.480	.535	NA
AV N1 Latency	.107	.194	.265	.325	.376	NA
AO N1 Latency	.047	.089	.128	.164	.197	NA
AO N1 Amplitude	.000	.000	.000	.000	.000	NA
AV P2 Latency	.000	.000	.000	.000	.000	NA

Note. TBW = Temporal binding window, AV Word Recognition = Whole word identification accuracy in the audiovisual condition of the speech-in-noise task at -3 or -6 dB signal-to-noise ratio (SNR), AO = Auditory-only condition of ERP task, AV = Audiovisual condition of ERP task, N1 = timeframe between 100 ms and 140 ms post-stimulus onset, P2 = timeframe between 160 ms and 240 ms, ID = Infant-directed speech, AD = Adult-directed speech, NA = Not applicable - *g* coefficients do not converge on acceptable stability even for estimated coefficients at 10 or more observations. See Table 2 for precise operational definitions of all variables. Bolded values are those that exceed our a priori stability criterion of $g = 0.8$. Underlined and bolded values reflect values that exceed the criterion at the lowest number of observations for that variable.