Clinical Linguistics & Phonetics, 2014; Early Online: 1–8
© 2014 Informa UK Ltd.
ISSN: 0269-9206 print / 1464-5076 online
DOI: 10.3109/02699206.2014.966395

informa
healthcare

1769

Development of an audiovisual speech perception app for children with autism spectrum disorders

JULIA IRWIN[1,2], JONATHAN PRESTON[1,3], LAWRENCE BRANCAZIO[1,2], MICHAEL D'ANGELO[1], & JACQUELINE TURCIOS[1,2]

[1]*Haskins Laboratories, New Haven, CT, USA,* [2]*Department of Psychology, Southern Connecticut State University, New Haven, CT, USA, and* [3]*Department of Communications Sciences and Disorders, Syracuse University, Syracuse, NY, USA*

(*Received 11 September 2014; accepted 14 September 2014*)

Abstract
Perception of spoken language requires attention to acoustic as well as visible phonetic information. This article reviews the known differences in audiovisual speech perception in children with autism spectrum disorders (ASD) and specifies the need for interventions that address this construct. Elements of an audiovisual training program are described. This researcher-developed program delivered via an iPad app presents natural speech in the context of increasing noise, but supported with a speaking face. Children are cued to attend to visible articulatory information to assist in perception of the spoken words. Data from four children with ASD ages 8–10 are presented showing that the children improved their performance on an untrained auditory speech-in-noise task.

Keywords: Audiovisual app, autism spectrum disorder, speech perception

Audiovisual speech perception in typical and ASD listeners

Speech perception is a critical element to developing successful communication in childhood. Although developmental research and clinical practice tend to emphasize *auditory* perception, most children have extensive *visual* exposure to the mouth movements of speech in face-to-face communication. Indeed, visible phonetic information from mouth articulations contributes substantially to the comprehension of speech (McGurk & MacDonald, 1976; Reisberg, McLean, & Goldfield, 1987).

One population that may have attenuated experience with the speaking face is individuals with autism spectrum disorders (ASD). In addition to ASD's hallmarks of deficits in social

Correspondence: Dr Julia Irwin, Senior Scientist, Haskins Laboratories – Child Language Studies, 300 George Street, New Haven, CT 06511, USA. Associate Professor, Southern CT State University – Psychology, 501 Crescent Street, New Haven, CT 06515, USA. E-mail: Julia.Irwin@Haskins.Yale.edu

communication and social interactions and of repetitive behaviours and interests (American Psychiatric Association, 2013), a commonly reported feature of ASD is facial gaze avoidance and reduced eye contact with others in social situations (Hobson, Ouston, & Lee, 1988; Hutt & Ounstead, 1966; Phillips, Baron-Cohen, & Rutter, 1992; Volkmar & Mayes, 1990). Critically, Irwin and Brancazio (2014) recently reported that children with ASD look less overall at speaking faces and less to the mouth of the speaker compared to typically developing children in the context of auditory noise. Accordingly, children with ASD may have generally reduced exposure to the visual phonetic information that might aid in the development of robust neural representations of speech. Moreover, visual speech information plays an especially helpful role in the perception of auditory speech in noise (Sumby & Pollack, 1954), a condition that has been reported to be difficult for children with ASD (Alcántara, Weisblatt, Moore, & Bolton, 2004). Unfortunately, the available evidence indicates that children with ASD use the face *less* to identify speech in the presence of background noise than their typically developing peers (Irwin, Tornatore, Brancazio, & Whalen, 2011). Thus, interventions designed to increase children's attention to articulatory information on the face of a speaker could support access to the speech signal, especially in difficult listening environments.

In light of evidence that early intervention can positively influence language outcome in children with ASD (Tager-Flusberg et al., 2009), we posit that drawing attention to the face of a talker, and in particular the mouth, may be especially helpful at improving identification of the speech signal. Accordingly, we created an audiovisual speech perception program to improve perceptual sensitivity to speech in children with ASD. As the goal is for children to develop more robust perceptual representations of spoken language, we sought to determine whether audiovisual training can generalize to perception of auditory speech in noise. The purpose of this brief report is to describe the rationale and the structure of the program, and to present preliminary data on whether any changes in speech perception can be observed.

Overview of listening to faces

Listening to Faces (L2F) is a theoretically driven, researcher-developed application designed for use with an iPad. The L2F program is an interactive, adaptive program that presents videos of speakers producing monosyllabic words in varying levels of auditory noise. The program adapts to the user's performance, increasing in difficulty with improved performance. The focus is on anterior consonants, which are easily confusable in perception. The speakers vary in age and gender, increasing variability in the speech signal, which may aid generalisability in perceptual training (Bradlow & Pisoni, 1999; Lively, Logan & Pisoni, 1993; Magnuson & Nusbaum, 2007; Rvachew, 1994). After each word is presented, images depicting the meaning of four words (including the word actually spoken and three similar words) appear on screen (Figure 1, panel A). The child responds by touching the image of the word that they heard. If the child responds correctly, positive feedback is provided (a smiling cartoon face with a chime) and he/she moves on to the next trial. If the child responds incorrectly, visual and auditory feedback is provided by displaying a red X over the incorrect choice and an auditory prompt that says ''sorry''. The correct choice is then identified with the smiling cartoon face (Figure 1, panel B). If children choose the incorrect word for two consecutive trials, a red arrow pointing to the mouth of the speaker will appear and an auditory prompt ''Look at the mouth'' is presented (Figure 1, panel B). After six (non-consecutive) correct trials, a brief reinforcer video plays (such as fireworks or animated animals dancing) to sustain interest in the task.

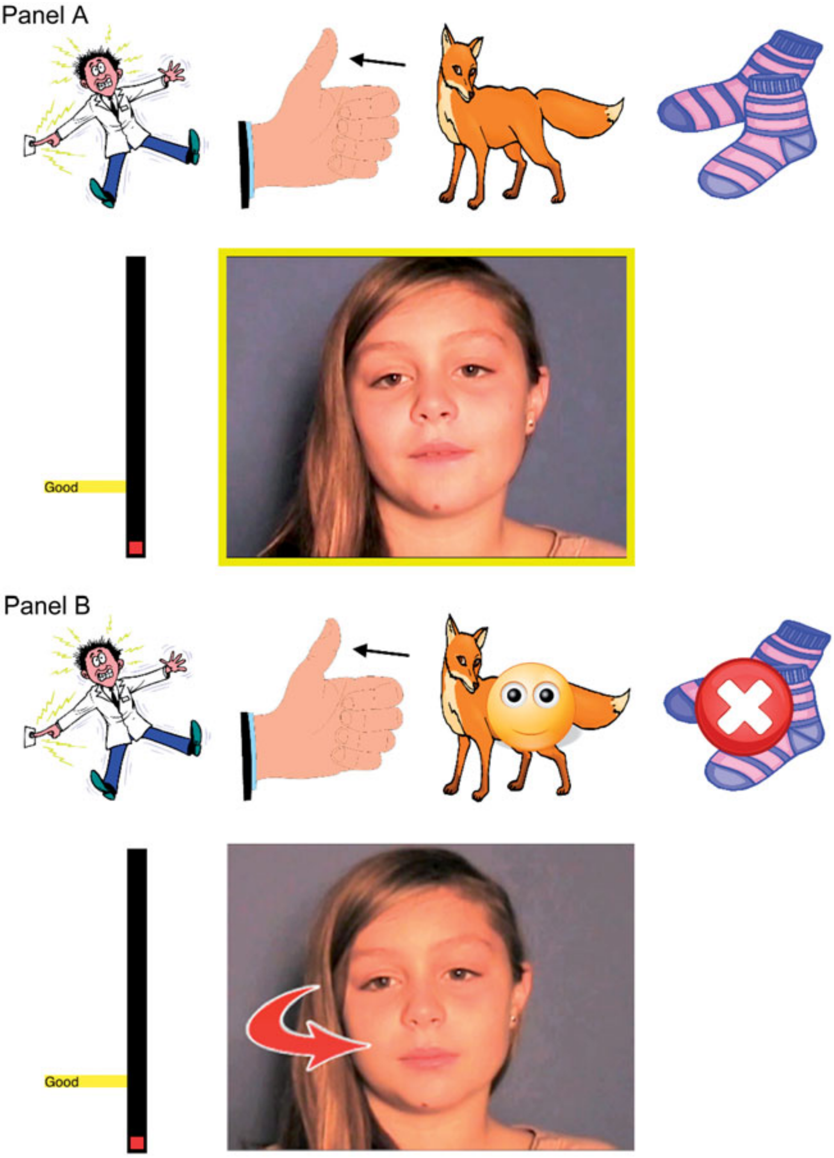

Figure 1. Panel A: Image of the iPad display as seen by the child participant. Along the top panel are the images for the monosyllable words: *shock, thumb, fox and socks*. Below the imaged words is video of the speaker who produced the target word *fox*. The frame presented depicts the labiodental contact for the /f/ which helps to visually discriminate the word ‘‘*fox*’’ from the other words. To the left of the panel is a progress bar to show the child how well he or she is progressing through the task. Panel B: Image of the iPad display after the child participant’s response. Along the top panel is the same set of monosyllabic words: shock, thumb, fox, socks. In this case, the child incorrectly chose the imaged word socks (seen in the final image from the left with an X overlaid across the image of socks). Next to the incorrect choice is the target word fox (overlaid with a smiling face). Due to the error, participants receive feedback to ‘‘Look at the mouth,’’ drawing attention to the mouth of the speaker. To the left of the panel is a progress bar to show the child how well he or she is progressing through the task.

Stimuli and protocol

Training stimuli

The four words presented during any given trial include the target (correct response) and three foils. One foil is a word that rhymes with the target, and the remaining two foils are selected to begin with consonants that differ in place of articulation (they may or may not rhyme with the target). All stimuli are monosyllabic; items include nouns, verbs or adjectives that can be represented by a colour picture. Initial consonants include bilabial /b/, labiodental /f/, interdental /θ/, alveolar /s/ and palatal/rounded-lip /ʃ/ (e.g. bat, fox, think, sick, shield). Each of the four words in a trial begins with a different initial consonant with visible articulatory information (and visibly distinct from the other consonants used in the training set), ensuring that, if the child attends to the speaking face, the items are visually discriminable.

Recorded stimuli were produced by child, adolescent and adult monolingual native speakers of American English ($N = 10$). Adult speakers consisted of three females (age range of speakers 21–24 years). Adolescent speakers consisted of two males (ages 15 and 17 years). Child speakers consisted of two girls and three boys (age range 8–12 years). Speakers were video recorded in a sound attenuated recording studio using a digital video recorder. High quality audio was simultaneously recorded with a separate microphone at 44 kHz sampling rate. Audio and video files were edited and ranged from 1 to 2 s in duration. The audio segment for each word was extracted, normalised at 70 dB using Praat (Boersma & Weenink, 2014) and re-dubbed with the appropriate video.

Background noise files were created for each stimulus and were signal-correlated, meaning that the amplitude envelope of the noise was temporally matched to that of the word. In Praat, Gaussian (white) noise was multiplied by the intensity envelope for each token and saved as a separate file. Multiple versions of each noise file at different amplitude levels were created for different signal-to-noise (SNR) ratios by normalising the noise. The stimuli and the noise are mixed online in the L2F program by presenting both simultaneously; the noise level in each trial is specified by the programmed training schedule, described below.

Auditory noise assessment stimuli

To assess transfer from audiovisual training to auditory perception (generalisation), an assessment module was created. This auditory noise assessment (ANA) consists of words spoken by a single adult female native speaker of American English. This speaker's voice is never included in the training and thus reflects across-speaker generalisation. The ANA consists of 50 words from the training set, with 10 words from each phoneme category /b, f, ʃ, θ, s/. Using Praat, each word was normalised at 65 dB, and multiple files of signal-correlated noise at different amplitude levels were created for each word, using the technique described earlier. Individual word-plus-noise files were created by mixing the word with the signal-correlated noise at a given noise level. Within each phoneme list of 10 words, one word was randomly chosen to be presented with each of the following signal-correlated noise SNR's: +40, +20, +10, +5, 0, −5, −8, −10, −15 dB. Each time the ANA is presented, the same words are used; however, the order of the 50 words is randomised.

Programmed training schedule

Training procedures are designed to adaptively increase in difficulty based on performance. The audiovisual stimuli are presented in blocks of 24 trials at a time. Each block of 24 trials takes approximately 5 min to complete, and the program is designed to be practiced in two blocks per day (approximately 10 min for a single session). During a block, the program selects five /b, f, ʃ, θ/ and four /s/ words, then it selects a speaker while ensuring that all 10 speakers are represented in each block. As there are only 10 speakers and 24 trials per block, each speaker appears for multiple words. The 24 speaker-word combinations are then randomised to be presented in the block.

Stimuli are presented at 70 dB and the noise is adaptively varied between no noise and SNRs ranging from +40 to −10 dB. The adaptive procedure was designed to converge upon a noise level that is challenging, but not impossible, for each participant. When a participant meets certain thresholds of accuracy, the noise level increases for the next day's training blocks; the size of the increase in noise level is tied to the participant's accuracy level. When a participant correctly identifies at least 75% of the words (18–24 correct responses), there is a large noise level increase during the next day (increases of 30, 20, 10 and then 5 dB on successive blocks with at least 75% correct). When accuracy is above 45% but below 75% (11–17 correct responses), the noise level for the next day is increased by 2 dB. When accuracy is below 45% (0–10 correct responses), the noise level for the next day is reduced by 5 dB.

A familiarisation module was developed to ensure users of L2F know the target vocabulary items prior to perceptual training. For example, items such as ''thaw'', ''shack'' and ''sack'' are imageable but might require familiarisation. The 62 items are presented auditorily without noise, and the child selects a picture from a field of four to demonstrate comprehension. Feedback is provided for incorrect responses, this familiarisation takes approximately 15 min.

The training protocol begins with the familiarisation module, followed by a series of initial pre-test ANA sessions, after which training can begin. The app also allows for ANA sessions at variable points between training blocks. The protocol finishes with post-test ANA sessions.

Preliminary data

Four monolingual American English speaking children with ASD participated in an initial trial of L2F. All were Caucasian males, and ranged in age from 8;2 (years; months) to 10;3 (mean age 9;5). Children met criteria for ASD according to three criteria: (1) had an existing diagnosis of an ASD from a licensed clinician familiar with autism, (2) met or exceeded cut-off scores for autism spectrum or autism proper on the ADOS; Autism Diagnostic Observation Schedule – Generic (ADOS-G; Lord, Rutter, DiLavore, & Risi, 2002; Lord et al., 2000) and (3) met or exceeded cut-off criteria on the language/communication, reciprocal social interactions and repetitive behaviour/interest domains on the Autism Diagnostic Interview-Revised (ADI-R; Lord, Rutter, & LeCouteur, 1994).

Standardised language and cognitive assessments were administered for descriptive purposes and are presented in Table 1. These included the Block Design and Matrix Reasoning subtests of the Weschler Abbreviated Scale of Intelligence (WASI; Wechsler, 1999), the recalling sentences and formulating sentences of the Clinical Evaluation of Language Fundamentals-4 (CELF-4; Semel, Wiig, & Secord, 2003), the Peabody Picture Vocabulary Test (PPVT-IV, Dunn & Dunn, 2007), the Expressive Vocabulary Test (EVT-2, Williams, 2007) and the Goldman–Fristoe Test of Articulation-2 (Goldman & Fristoe, 2000).

In addition, all four children with ASD scored below 85% on 100 trials of Speech Assessment and Interactive Learning System (SAILS), with 20 trials assessing /f, θ, s, ʃ, r/ (Rvachew, 1994). A normative sample from our laboratory of 15 children ages 8;2–10;10 with typical speech and language and no diagnosis of ASD achieved average SAILS scores of 85% (SD 7%, median 88%). Each child with ASD in this study scored below the median score of their typically developing peer group. Scores for the measures listed above for all ASD participants are summarised in Table 1.

Before children were introduced to the training, they completed the familiarisation module and the first ANA in the presence of a researcher at the laboratory. The children then took the iPad home. Families were instructed to have the child complete the training in a quiet environment (e.g. with no television or radio noise); they were also instructed not to modify the volume of the iPad, and the volume switch was covered so the output level could not be changed. Children were asked to use the training 3 days/ week (approximately 10 min each session) for 12 weeks. The app was pre-set to only run two blocks per day, three times per week. A staff member checked in weekly by email to ensure that the child was training, and one at-home visit was conducted mid-way through the program to observe the child, ensure compliance and answer questions. Participants received gift cards for their participation.

The planned design was for children to first complete a baseline assessment of at least three administrations of the ANA on separate days prior to training (Time 1). Although the planned design was to collect data from two participants who had undergone immediate training and two who had training after a 3-week delay, due to errors with the internal clock initiating start date of training with the iPad and non-compliance with the procedures at home, the amount of training between assessments was not uniform. Therefore, we present data as the program was executed and report on pre- and post-treatment comparisons.

These preliminary results show a pattern of improvement over time. Figure 2 depicts participants' performance on the ANA, which was administered between training blocks. As can be seen, all participants showed an increase in accuracy on this auditory-only task. All four participants began the training with ANA scores below 73% and ended with scores above 88%, reflecting increased accuracy at identifying (auditory-only) words spoken in noise. Across the participants, the mean pre-treatment scores (Time 1) was 66.7% (SD 7%). The mean post-treatment scores (Time 3 or 4) was 89.8% (SD 2%). A paired *t*-test indicated that mean post-treatment scores on the ANA were significantly higher than pre-treatment scores ($t[3] = 6.87$, $p = 0.006$).

In addition to improved performance in the ANA, all participants improved in performance in the training sessions: By the end of the training, all participants had progressed to the most difficult SNR of −15 dB (as described earlier, the noise level in each training block was set adaptively based on previous performance).

Table 1. Sociodemographics, language and cognitive profiles of participants.

	Diagnosis	Age	WASI.PRI	CELF.RS	CELF.FS	PPVT	EVT	GFTA	SAILS
P1	autism	9;10	100	4	9	78	88	107	83
P2	PDD-NOS	10;3	106	8	9	91	93	103	75
P3	autism	9;6	92	8	12	92	94	100	68
P4	autism	8;2	88	11	12	113	110	106	72

Scores for all of the measures are standard scores, with the exception of scaled scores for the CELF and the SAILS, which is scored out of a possible 100. WASI.PRI, Wechsler Abbreviated Scales of Intelligence Perceptual Reasoning Index; CELF.RS, Clinical Evaluation of Language Fundamentals-4 Recalling sentences; CELF.FS, Clinical Evaluation of Language Fundamentals Formulated Sentences; PPVT, Peabody Picture Vocabulary Test-4; EVT, Expressive Vocabulary Test-2; GFTA, Goldman–Fristoe Test of Articulation-2.

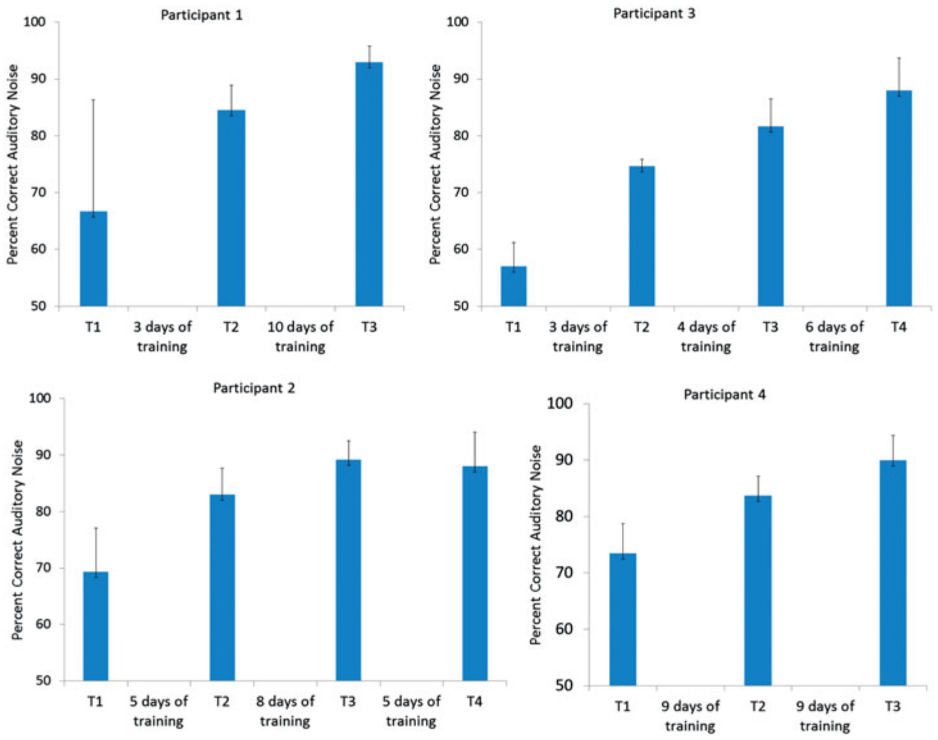

Figure 2. Mean scores on the Auditory Noise Assessment (ANA) at various time points (Time 1, Time 2, Time 3 and Time 4). Each mean consists of at least three consecutive administrations of the ANA. Error bars represent standard deviations.

Interpretations and future directions

The data presented here suggest that training speech-in-noise, with explicit focus on cueing visible articulation may facilitate perception of auditory-only speech in noise in children with ASD. These data must be interpreted with caution, given the small sample size. In addition, because this exploratory study did not include an untrained control group, it is not possible to rule out other competing hypotheses about mechanisms that may be responsible for change in performance on the ANA (such as a practice effect for the words in noise not due to the training). With these caveats in mind, initial acceptability and outcome data suggest that the L2F app can be tolerated by children. Training led to an increase in accuracy on the untrained ANA, indicating generalisation to perception of auditory only words in noise. These preliminary findings suggest that L2F may be helpful for children with ASD in perceiving speech in noise.

Acknowledgements

The L2F app is in the development phase and is not currently available for commercial use.

Declaration of interest

The authors report no conflicts of interest. This work was supported by U.S. Department of Health and Human Services-National Institutes of Health (P01 HD-01994, R03 DC-007339, R15 DC01386401, R21 DC-011342) and Language Evaluation and Research Network (LEARN) Center at Haskins Laboratories.

References

Alcántara, J. I., Weisblatt, E. J. L., Moore, B. C. J., & Bolton, P. F. (2004). Speech-in-noise perception in high-functioning individuals with autism or Asperger's syndrome. *The Journal of Child Psychology and Psychiatry*, *45*, 1107–1114.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders (5th ed.)*. Arlington, VA: American Psychiatric Publishing.

Boersma, P., & Weenink, D. (2014). Praat: Doing phonetics by computer [Computer program]. Version 5.3.77. Retrieved May 18, 2014, from http://www.praat.org.

Bradlow, A. R., & Pisoni, D. B. (1999). Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors. *The Journal of the Acoustical Society of America*, *106*, 2074–2085.

Dunn, D., & Dunn, L. (2007). *Peabody picture vocabulary test, Fourth Edition (PPVT-4)*. San Antonio, TX: PsychCorp.

Goldman, R., & Fristoe, M. (2000). *Goldman Fristoe test of articulation (2nd ed.)*. Circle Pines, MN: AGS.

Hobson, R. P., Ouston, J., & Lee, A. (1988). What's in a face? The case of autism. *British Journal of Psychology*, *79*, 441–453.

Hutt, C., & Ounstead, C. (1966). The biological significance of gaze aversion with particular reference to the syndrome of infantile autism. *Behavioral Science*, *11*, 346–356.

Irwin, J. R. & Brancazio, L. (2014). Seeing to hear? Patterns of gaze to speaking faces in children with autism spectrum disorders. *Frontiers, Language Sciences*, *5*, 397.

Irwin, J. R., Tornatore, L., Brancazio, L., & Whalen, D. H. (2011). Can children with autism spectrum disorders ''hear'' a speaking face? *Child Development*, *82*, 1397–1403

Lively, S. E., Logan, J. S. & Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/ and /l/. II The role of phonetic environment and talker variability in learning new perceptual categories. *Journal of the Acoustical Society of America*, *94*, 1242–1255.

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

Lord, C., Rutter, M., DiLavore, P. C., & Risi, S. (2002). *Autism diagnostic observation schedule: Manual*. Los Angeles, CA: Western Psychological Services.

Lord, C., Rutter, M., & LeCouteur, A. (1994). Autism diagnostic interview revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of Autism and Developmental Disorders*, *24*, 659–685.

Magnuson, J. S., & Nusbaum, H. C. (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *Journal of Experimental Psychology: Human Perception and Performance*, *33*, 391–409.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746–748.

Phillips, W., Baron-Cohen, S., & Rutter, M. (1992). The role of eye contact in goal detection: Evidence from normal infants and children with autism or mental handicap. *Developmental Psychopathology*, *4*, 375–383.

Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli (Chapter). In B. Dodd, & R. Campbell, (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 97–113). London, UK: Lawrence Erlbaum Associates, Ltd.

Rvachew, S. (1994). Speech perception training can facilitate sound production learning. *Journal of Speech, Language, and Hearing Research*, *37*, 347–357.

Semel, E., Wiig, E. H., & Secord, W. A. (2003). *Clinical evaluation of language fundamentals (4th ed.)*. New York: The Psychological Corporation.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America*, *26*, 212–215.

Tager-Flusberg, H., Rogers, S., Cooper, J., Landa, R., Lord, C., Paul, R., Rice, M., Stoel-Gammon, C., Wetherby, A., & Yoder, P. (2009). Defining spoken language benchmarks and selecting measures of expressive language development for young children with autism spectrum disorders. *Journal of Speech, Language and Hearing Research*, *52*, 643.

Volkmar, F. R., & Mayes, L. C. (1990). Gaze behavior in autism. *Development and Psychopathology*, *2*, 61–69.

Wechsler, D. (1999). *Wechsler abbreviated scale of intelligence*. San Antonio, TX: The Psychological Corporation.

Williams, K. T. (2007). *EVT-2: Expressive vocabulary test*. Austin, TX: Pearson Assessments.