AJA

Research Article

Behavioral Assessment of Hearing in 2- to 7-Year-Old Children: Evaluation of a Two-Interval, Observer-Based Procedure Using Conditioned Play-Based Responses

Angela Yarnell Bonino,[a] Michael E. Ramsey,[b] Haley M. McTee,[a] and Eric A. Vance[b]

Purpose: It is challenging to collect reliable behavioral data from toddlers and preschoolers. Consequently, we have significant gaps in our understanding of how auditory development unfolds during this time period. One method that appears to be promising is an observer-based procedure that uses conditioned, play-based responses (Bonino & Leibold, 2017). In order to evaluate the quality of data obtained with this method, this study presented a suprathreshold signal to determine the number of trials 2- to 7-year-old children could complete, as well as the associated hit rate and observer confidence.

Method: Participants were 23 children (2–7 years old). Children were taught to perform a play-based motor response when they detected the 1000-Hz warble tone signal (at 30 dB SPL). An observer evaluated children's behavior using a 2-interval, 2-alternative testing paradigm. Testing was terminated after 100 trials or earlier, if signs of habituation were observed.

Results: Data were successfully collected from 22 of the 23 children. Of the 22 children, all but 1 child completed 100 trials. Overall hit rate was high (0.88–1.0; *M* = 0.94) and improved with listener age. Hit rate was stable across the test session. Strong agreement was seen between the correctness of the response and the observer's confidence in the judgment.

Conclusion: Results of this study confirm that the 2-interval, observer-based procedure described in this article is a powerful tool for measuring detection and discrimination abilities in young children. Future research will (a) evaluate coder reliability and (b) examine stability of performance across a test session when the signal intensity is manipulated.

Supplemental Material: https://doi.org/10.23641/asha.8309273

Substantial differences in behavioral data between infants and school-age children suggest that the auditory system undergoes significant development in early childhood (reviewed by Werner, 2017; Buss, Hall, & Grose, 2012). However, it is not clear how auditory development unfolds during the toddler and preschool years due to methodological constraints. In order to address the shortage of behavioral methods for testing young children, Bonino and Leibold (2017) developed an observer-based method in which a child's behavior is judged by an experimenter (called an *observer*) using a two-interval, two-alternative testing paradigm. Children's response to the stimulus is further shaped by training them to perform a conditioned, play-based response to the sound. This method is called the *Play Observer-Based, Two-Interval (PlayO2I) method*. While initial feasibility data for this method are promising (Bonino & Leibold, 2017), questions remain regarding the feasibility and reliability of this method. How many trials can children perform in a single test session? How frequently are lapses in attention experienced for the child–observer team? How confident are observers at judging children's responses in a two-interval task? In an effort to address these questions, this study presented a suprathreshold stimulus—a 1000-Hz warble tone at 30 dB SPL in quiet—throughout the testing session. Data were examined from 2- to 7-year-old children to determine hit

[a]Department of Speech, Language and Hearing Sciences, University of Colorado Boulder
[b]Laboratory for Interdisciplinary Statistical Analysis, Department of Applied Mathematics, University of Colorado Boulder

Correspondence to Angela Yarnell Bonino: angela.bonino@colorado.edu

Editor-in-Chief: Sumitrajit Dhar
Editor: Kathy Vander Werff

Received January 15, 2019
Revision received February 21, 2019
Accepted March 14, 2019
https://doi.org/10.1044/2019_AJA-19-0004

Disclosure: *The authors have declared that no competing interests existed at the time of publication.*

rate, observer confidence, and the number of trials that could be performed in a single test session.

The development of this method was motivated by the observation that there are few paradigms for testing hearing in toddlers and preschoolers. Researchers have had mixed success testing toddlers and preschoolers by modified existing paradigms designed for either infants or school-age children (e.g., Allen & Wightman, 1992, 1994; Eisenberg, Martinez, & Boothroyd, 2007; Garadat & Litovsky, 2007; Holt & Lalonde, 2012; Jensen & Neff, 1993). Based on infant paradigms, toddlers and preschoolers have been tested by measuring their looking time toward an object (e.g., Newman, 2011) or by training children to make a response, such as a head turn, when the signal is heard (e.g., Eisenberg et al., 2007; Schneider, Trehub, Morrongiello, & Thorpe, 1986). However, toddlers often quickly habituate on these paradigms (e.g., Eisenberg et al., 2007; Primus & Thompson, 1985). Another challenge with many of the infant paradigms is that it is difficult to interpret developmental trends over a wide age span of childhood. This is because the data are often vulnerable to a variety of listener or observer factors (e.g., motivation, habituation, lapses in attention, or response bias) that may change as a function of listener age (e.g., Jones, Kalwarowsky, Braddick, Atkinson, & Nardini, 2015).

The other strategy researchers have used is to employ forced-choice tasks that were designed for school-age children (e.g., Allen & Wightman, 1992, 1994; Garadat & Litovsky, 2007; Jensen & Neff, 1993). For these tasks, children typically indicate the interval that a tonal signal was played or select the picture that corresponds to the target speech signal. In addition to being time efficient, the other benefit of this approach is that forced-choice paradigms are expected to guard against response bias (e.g., Green & Swets, 1966; although data from Jones, Moore, and Shub, 2015, challenge this assumption). However, young children often lack the cognitive abilities required to complete forced-choice tasks. Thus, even with extensive training, some 3- and 4-year-olds and most 2-year-olds cannot perform these tasks (e.g., Allen & Wightman, 1992, 1994). There are two notable exceptions that have shown to be feasible for 2-year-olds in the laboratory. First, using a forced-choice, picture-pointing task (CRISP-Jr.; Garadat & Litovsky, 2007), Hess, Misurelli, and Litovsky (2018) recently reported masked word recognition thresholds for 2-year-old children. All of the 2-year-olds with normal hearing sensitivity appeared to be able to perform the task. However, the utility of this paradigm is limited to measuring speech recognition. Second, Holt and Lalonde (2012) evaluated 2-year-olds' speech-sound discrimination with a modified, change/no-change procedure. This task required children to move to one of two areas in the booth, one for the "no-change" stimulus array and the other for the "change" stimulus array. While most of the 2- and 3-year-olds were able to perform this task, four of the youngest children (< 2.5 years; 10% of the sample) could not. Furthermore, in a follow-up article, the authors recommended that this method not be used until at least 2.5 years of age, and perhaps later, depending on the cognitive and language abilities of the children in the sample (Lalonde & Holt, 2014).

Given the limited success researchers have had testing hearing in young children—especially 2-year-olds—our approach was to draw on the successes of conditioned play audiometry (CPA). CPA is routinely used by audiologists to obtain thresholds from 2- to 5-year-old children (e.g., Barr, 1955; M. Thompson & Thompson, 1972). Children are trained to provide a play-based, motor response (e.g., putting a peg into a board) that is time-locked to the presentation of an auditory signal. Using this method, a conventional clinical audiogram can be obtained for > 90% of 3-year-olds (e.g., Barr, 1955; G. Thompson & Weber, 1974) and 60%–70% of 2-year-olds (e.g., Nielsen & Olsen, 1997; M. Thompson, Thompson, & Vethivelu, 1989). CPA uses a single-interval adaptive procedure, which is designed to be efficient and has high patient acceptance. Unfortunately, CPA is susceptible to observer and listener response bias (Green & Swets, 1966). For example, the audiologist, who initiates the signal presentation, may be influenced by factors such as expectations regarding the child's hearing sensitivity. In addition, threshold estimates can be affected by the listener's decision strategy in a single-interval task. In other words, a conservative response criterion results in a higher threshold than a liberal response criterion. Moreover, response criterion is affected by listener age and task (e.g., Bonino, Leibold, & Buss, 2013; Leibold & Werner, 2006; Marshall & Jesteadt, 1986). Because the CPA procedure does not control for these forms of bias, it is difficult to compare threshold estimates across listener age groups and stimulus conditions.

In order to overcome the shortcomings of CPA, the PlayO2I method combines the response method and behavioral shaping strategies of CPA with the psychometric rigor of the observer-based psychoacoustic procedure (OPP; Olsho, Koch, Halpin, & Carter, 1987). OPP is designed to control for observer bias and is used to test infants. In this method, the experimenter (called an *observer*) watches the infant's behavior to determine whether a signal or no-signal trial has occurred. Although infants often provide a head-turn response, any consistent behavior that is time-locked with the presentation of the stimulus can be used. This method is designed to control for observer bias because the observer is unaware to whether the trial is a signal or no-signal trial. One limitation of OPP is that half of the trials are no-signal trials, resulting in extensive wait time between signal trials and a limited amount of data. In order to address this challenge, Browning, Buss, and Leibold (2014) modified OPP to be a two-interval, forced-choice procedure. In this adaptation, each trial contained two observation intervals, and the signal was randomly presented in one of them. At the end of each trial, the observer selected which interval contained the signal based on the infant's behavior. This procedure appears to be feasible and efficient based on 7- to 8-month-old infants' training data reported by Browning et al.

Drawing on the infant work discussed above, the PlayO2I method is an observer-based procedure that uses a two-interval, forced-choice paradigm. In this method, children are trained to perform a play-based, conditioned motor response when a signal is heard. However, the observer can use any behavioral response that is time-locked with the presentation of the signal (e.g., eye movements, change in activity level, and eyebrow furrow). On each trial, two observation windows are presented with the signal being randomly placed in one of the intervals. Based on the child's behavior, the observer determines if the signal occurred in Interval 1 or Interval 2. Using the PlayO2I method, we recently measured tone detection thresholds for thirty-three 2- to 4-year-old children with no known hearing problems (Bonino & Leibold, 2017). A valid threshold estimate was obtained for 82% of the children. Thus, the PlayO2I method appears to be a promising procedure for measuring thresholds, while guarding against observer bias, from an age group that has historically been difficult to test.

Building upon the promising feasibility data (Bonino & Leibold, 2017), the purpose of the current study was to evaluate the reliability of the data obtained with the PlayO2I method. Specifically, a suprathreshold stimulus was presented to 2- to 7-year-old children to determine the number of trials completed before habituation and the associated hit rate and observer confidence. The test session was terminated after 100 trials or earlier, if performance was consistent with the effects of habituation or fatigue. The termination criterion was achieved if the observer judged four out of five consecutive trials as "not confident" after the activation of the mechanical toys. Consistent with data from Bonino and Leibold (2017), we anticipated that > 80% of the children tested would be able to perform the task and that the proportion of correct responses (hit rate) would be > 0.80. Our hypothesis was that the number of trials children performed would be dependent on their chronological age. It was predicted that 3- to 4-year-old children would complete 60–90 trials and that 2-year-olds would complete fewer trials. M. Thompson et al. (1989) reported that 24- to 27-month-old toddlers completed a mean of 28 trials with the CPA task. However, in our work with this method, we have observed that 2-year-olds are typically able to complete at least a single adaptive threshold run per visit (30–45 trials; Bonino & Leibold, 2017).

Method

Participants

Participants were 23 children (14 girls and 9 boys): eight 2-year-olds, five 3-year-olds, seven 4-year-olds, and three 5- to 7-year-olds. Selection criteria include (a) no risk factors for hearing loss as assessed by parental report, (b) no history of pressure equalization tubes placement, (c) not under treatment for otitis media within the prior month, (d) no risk factors for developmental delays as assessed by parental report on the Comprehensive Parent/Caregiver Form of the Vineland-3 Adaptive Behavior Scales (Sparrow, Cicchetti, & Saulnier, 2016), (e) had not participated in more than one previous hearing experiment,[1] and (f) healthy on test day. All children passed a distortion product otoacoustic emission and a tympanometry screening on the day of testing. The pass screening criterion for otoacoustic emissions was a signal-to-noise ratio of ≥ 6 dB for three out of the four screening frequencies (2000, 3000, 4000, and 5000 Hz). The pass screening criterion for tympanometry was a peak admittance of ≥ 0.2 mmhos at a pressure between −200 and 50 daPa. All children were able to complete testing within a single 1-hr visit. This research was approved by the institutional review board at the University of Colorado Boulder.

Stimuli

The signal was a 500-ms, 1000-Hz warble tone (10-ms, $\cos^2$, rise/fall ramps) with a 20-Hz modulation rate and 5% modulation. For all stages of training and testing, the signal was presented in quiet at a fixed intensity of 30 dB SPL. This intensity level was expected to be well above threshold for the age range tested here (Bonino & Leibold, 2017; Schneider et al., 1986). Stimuli were generated by TDT RZ6 hardware that was controlled by a custom MATLAB script. Stimuli were presented to the left ear over a pair of Sennheiser HD 25 light-weight headphones.

Booth Setup and Experimenters

All testing was completed in a double-walled, sound-isolated booth (Industrial Acoustics). Testing was performed by two experimenters, referred to as a "test assistant" and an "observer." The test assistant sat with the child at a table inside the test booth. The role of the test assistant was to train the child, provide social reinforcement, redirect the child when needed, and change the "game" every 10–15 trials. The test assistant trained the child to perform the task with oral directions, modeling, and hand-over-hand assistance. Next to the test assistant (but out of sight of the child) was a variety of games that could be selected from during testing. Inside the booth were also two mechanical toys with lights in dark Plexiglas boxes that could be activated for additional reinforcement. The observer was located in the adjacent control room with the computer that controlled the experiment. The role of the observer was to run the experiment by initiating trials and judging the child's behavior. The observer was able to watch the child and assistant through an observation window. The observer and test assistant were able to communicate via a two-way communication system.

Six experimenters collected data. All experimenters were cross-trained and could serve as either the observer or the test assistant. Experimenters had a minimum of 1 year of experience testing young children's hearing and

[1] Of the 23 children recruited, 14 children had never participated in a hearing study (including seven of the eight 2-year-old children). The remaining nine children had completed a masked speech detection experiment 1–8 months earlier.

completed a ~3-month training period to learn the PlayO2I method.

Procedure

This study used the PlayO2I method—an observer-based procedure that used a two-interval, forced-choice paradigm (Bonino & Leibold, 2017). For each trial, there were two temporal observation windows. Observation intervals were 1,065 ms in duration separated by a 500-ms interstimulus interval (ISI).[2] Real-time visual and audio markers for each interval were provided to the observer. The real-time audio markers were also provided to the test assistant through headphones. In contrast to the experimenters, children were not provided any indicators that a trial had been initiated by the observer nor were they aware that there were two observation intervals. Trials were only initiated by the observer if the child was judged to be in a "ready state" (e.g., response toy was positioned on the belly or cheek; child was quiet). The observer then watched the child's behavior during the trial to determine which interval contained the signal. Recall that, although the child was trained to perform a motor-based response when the signal was heard, the observer was able to use any type of behavior provided by the child. The observer also indicated her confidence level for each judgment ("confident" vs. "not confident"). The observer was provided trial-by-trial feedback by the software.

Testing of the participants consisted of three stages: two stages of training followed by the collection of experimental data. Stage 1 was a conditioning phase in which the child was taught to perform a play-based, motor response (e.g., put a block in a bucket) when the target signal was heard. In Stage 1, the signal was always presented in the second interval to allow the assistant to pair the stimulus and the response. This stage continued until the observer judged that the child was able to independently produce a response that was time-locked to the signal. Children were required to perform at least three trials in this stage. On average, children completed 7.09 conditioning trials (SD = 5.48, max = 23). In Stage 2, the criterion phase, the child–observer team had to correctly identify the interval that contained the signal for four out of five consecutive trials. This criterion was met in ≤ 6 trials for all participants (M = 4.14, SD = 0.47). Upon the successful completion of training (Stages 1 and 2), experimental data were collected in Stage 3. The signal was presented in Stage 3 for a maximum of 100 trials. In Stages 2 and 3, the a priori probability of the signal occurring in an interval was 0.50. Experimenters were also blinded to which interval contained the signal in Stages 2 and 3.

Throughout the test session, care was taken to prevent habituation to the task and/or reinforcers (reviewed by McSweeney & Murphy, 2014). The primary strategy used to reduce the risk of habituation was that the test assistant changed the game every 10–15 trials. The test assistant selected games based on the child's developmental abilities and interests. These games included activities such as placing blocks in a bucket, building towers, racing cars down a ramp, and constructing puppets with Velcro attachments. Moreover, the complexity of the games often increased with time (e.g., elements of pretend play were introduced). Another strategy used to address potential habituation was to provide additional reinforcement with the activation of the mechanical toys. Mechanical toys were activated if an observer judged four out of five consecutive trials as "not confident" in Stage 3 of testing.[3] Once activated, mechanical toys were used for all remaining trials. In order for the software to activate the mechanical toys on a given trial, the interval selection had to be both correct and judged as confident by the observer. Only three children (2.51, 3.17, and 4.30 years old) were tested with the mechanical toys. Testing was terminated if the child continued to demonstrate performance consistent with habituation after the activation of the mechanical toys. The criterion for terminating testing was that the observer had to score 4 out of 5 consecutive responses as "not confident" after the mechanical toys had been activated.

Results

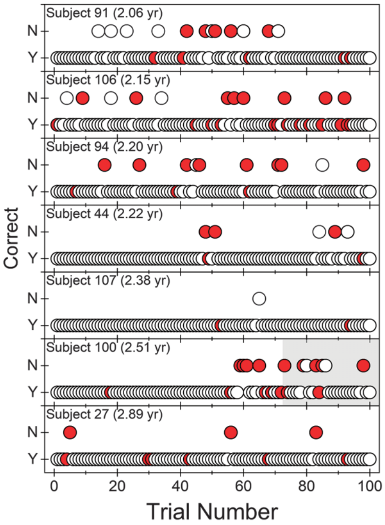
Useable data were obtained from 22 out of 23 children recruited for this study. The child who was not able to successfully complete training was 2.47 years old. Among the 22 children who could perform the task, 21 completed 100 trials during Stage 3 of testing. Only one child (Subject 6, 4.30 years old) met the termination criterion. For Subject 6, the mechanical toys were activated after 65 trials, and testing was terminated after 73 trials. Because this was the only child who demonstrated signs of habituation based on our criterion, his data were excluded from all statistical analyses below.

In Figure 1, individual trial-by-trial plots are provided for the seven 2-year-olds who were able to perform the task. Each panel provides trial-by-trial (circle) data for an individual child over the course of Stage 3 testing. The location of the symbol on the ordinate indicates if the observer was accurate in identifying the interval that contained the signal based on the child's behavior. If the correct interval was selected, the data point was plotted at "Yes (Y)." However, if the observer selected the wrong

[2] In Bonino and Leibold (2017), a 1,065-ms observation interval and a 300-ms ISI were employed. A 1,065-ms interval was programmed to allow for longer .wav files (e.g., a disyllabic target word) to also be presented in the same custom MATLAB software. The ISI was increased to 500 ms in the current study to account for potentially slower responses in 2-year-old children compared to older children. Pilot data from a young 2-year-old child suggested improved observer confidence for an ISI of 500 ms compared to 300 ms.

[3] For Subject 100 (2.51 years old), the observer deviated from this rule. Instead, the mechanical toys were activated because the child was inconsistently responding. The average observer confidence rate was 0.50 for the 10 trials preceding the activation of the mechanical toys at Trial 74.

Figure 1. Trial-by-trial data (Stage 3 only) are provided for the seven 2-year-old children. Recall that one child (2.47 years old) was unable to successfully complete training in this age group. Each panel represents a single child. Response accuracy is indicated by placement of the data point on the ordinate, "Yes (Y)" for a correct response and "No (N)" for an incorrect response. Observer confidence is indicated by fill color: white for "confident" and red for "not confident." Trials in which the mechanical toys were activated have a gray background. yr = years.



interval, the data point was plotted at "No (N)." Symbol color reflects the observer's confidence in her judgment for that trial: White is "confident," and red is "not confident." Trial-by-trial plots for all other children are provided in the Supplemental Materials: 3-year-olds (n = 5) in Supplemental Material S1, 4-year-olds (n = 7) in Supplemental Material S2, and 5- to 7-year-olds (n = 3) in Supplemental Material S3.

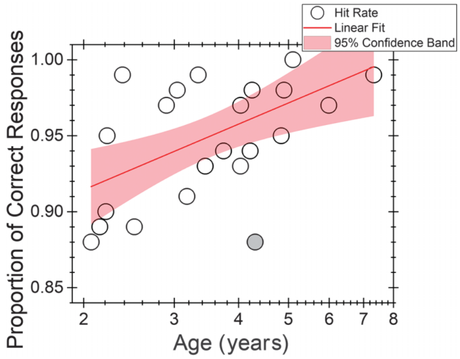
Hit Rate

The hit rate was calculated by determining the proportion of trials where the correct interval was selected by the child–observer team. Recall that, in the PlayO2I method, the observer's ability to make an accurate judgment is dependent on the child providing a reliable response to the signal. Across the test session, hit rate ranged from 0.88 to 1.0 (M = 0.95, SD = 0.04) for the test sessions in which 100 trials were completed (n = 21). For the remaining session, Subject 6 (4.30 years old), the hit rate was 0.88 prior to the child demonstrating signs of habituation (61 trials).

Developmental Trends

Individual hit rate values (open circles) are plotted in Figure 2 as a function of child age. For visual reference, data from the child excluded from the model (Subject 6, 4.30 years old) are represented by the filled circle. Age (in years) was transformed to $\log_{10}$ to allow for the observed decelerating effects of development with increasing age (e.g., Buss, Leibold, Porter, & Grose, 2017; Mayer & Dobson, 1982; Moller & Rollins, 2002). Improvement in hit rate with child age was fitted using a linear function of the form $y = a + b \times x$, where y is hit rate and x is child age in $\log_{10}$ years. In Figure 2, the solid line represents the mean trend line[4] and the shaded band represents 95% confidence bands for the regression line. Age (in $\log_{10}$ years) accounted for 35% of the variability in hit rate (R^2 = .35). A significant coefficient on $\log_{10}$ years (p = .004) was obtained. The predicted hit rates for 2-, 3-, and 4-year-olds were 0.91, 0.94, and 0.96, respectively.

Figure 2. Individual hit rates as a function of child age in $\log_{10}$ years scale. A linear function (line) was fitted to the data. The shading represents the 95% confidence bands for the regression line. Subject 6 (4.3 years old) was excluded from the model but is plotted here as a filled circle.



Stability of Performance Over the Test Session

One concern with this method is that performance for the child–observer team may change over the course of a test session. In order to evaluate this concern, two analyses were performed using the data from the 21 children who completed 100 trials in Stage 3. The first analysis examined hit rate at two different time points (within Stage 3): the first 30 trials and the last 30 trials. Results from this analysis are shown in Figure 3. Boxes represent ± 1 SD of the mean (dashed red line). For reference, overall hit rate (100 trials) is also reported. Symbols represent data from individual children, with fill color and symbol type indicating child age. Mean hit rate at the beginning of the test session was 0.97 (SD = 0.04), and that at the end of the session was 0.94 (SD = 0.06). This difference was not significantly different as tested by the Wilcoxon signed-ranks test, a nonparametric paired test (p = .09). Consistent with this group trend, hit rate for individual test sessions appeared to be similar for these two time points. Three test sessions had a deterioration in hit rate of > 0.1, meaning that ≥ 3 trials were incorrectly identified by the child–observer team in the last 30 trials compared to the first 30 trials. However, overall hit rate for these three test sessions was high: 0.95 (Subject 44, 2.20 years old), 0.89 (Subject 100, 2.51 years old), and 0.93 (Subject 89, 3.44 years old).

[4] The fitted regression line is $y = 0.87117 + 0.14363(x)$.

Figure 3. Hit rate was calculated for three time points (Stage 3): first 30 trials, last 30 trials, and all (100) trials. For each time point, mean (dashed line) and individual hit rates (symbols) are provided. The box indicates ± 1 *SD* of the mean. Symbol type and fill reflect child age. yr = years.

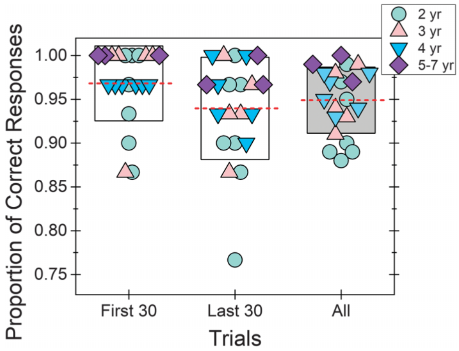
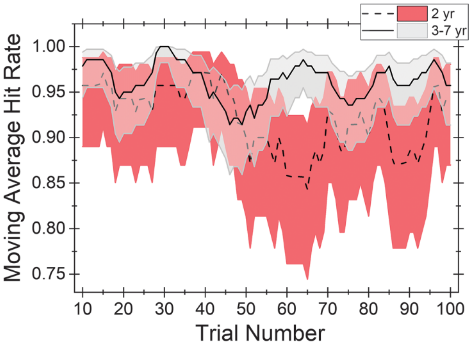


Figure 4. Simple moving averages for hit rate are provided for two age groups of children: 2-year-olds (dashed line) and 3- to 7-year-olds (solid line). Shading represents the 95% credible interval around the mean: red for 2-year-olds and gray for 3- to 7-year-olds. Trial number corresponds to the proportion of trials the child–observer team were correct, calculated over the current trial and the nine preceding trials for each group. For example, Trial 15 is the proportion correct among Session Trials 6–15. yr = years.

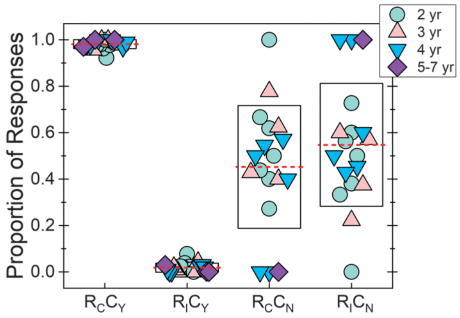


The second analysis calculated a simple moving average based on the hit rate for every 10 trials. To better understand potential age-related trends, two separate averages were calculated: one average for test sessions with 2-year-old children (n = 7) and another for test sessions with 3- to 7-year-old children (n = 14). In Figure 4, the average hit rates for test sessions with 2-year-olds and 3- to 7-year-olds are represented by the dashed and solid lines, respectively. Shading represents the 95% credible interval around the average hit rate.[5] Over the course of the testing session, the simple moving average did not fall below 0.84 and 0.91 for 2-year-olds and 3- to 7-year-olds, respectively. Based on visual examination of the moving average data, hit rate for the child–observer teams appears to be similar for the two child age groups for the first ~50 trials. However, after 50 trials, hit rate of the child–observer team appears to slightly worsen for 2-year-old children, but not for 3- to 7-year-old children. The worst performance for test sessions of 2-year-olds occurred at moving average Trial 65 (average of Trials 56–65). However, by Session Trial 75, hit rate for 2-year-olds' test sessions seemed to recover to a level comparable to that of the child–observer teams evaluating 3- to 7-year-old children.

Observer Confidence

The observer confidence was calculated by determining the proportion of trials that the observer indicated that she was "confident" about her judgment of which interval contained the signal. Across children, the average observer confidence was 0.93 (*SD* = 0.058; min = 0.79, max = 1.0). Observer confidence was highly correlated with overall hit rate (r = .82, p < .001). To further assess this, we created boxplots for all four of the correct/confident pairs for the observer in Figure 5. Boxes represent ± 1 *SD* of the mean (dashed red line). Each symbol represents data from an individual child, with fill color and symbol type indicating the child's age. On the x-axis, "R" stands for response, and the subscript "C" or "I" indicates whether the trial interval judgment was made correctly or incorrectly. Similarly, "C" stands for confident, and the subscript "Y" or "N" stands for "yes" or "no," respectively. For example, R_CC_Y stands for the proportion of correct responses, given that the observer responded that she was confident in her judgment of the child's behavior. In Figure 5, we see that this proportion was high, indicating that if a child provided a response that the observer could confidently judge,

Figure 5. The proportion of correct versus incorrect responses was calculated for each of the two possible observer confidence levels. Along the x-axis are the four accuracy–confidence pairs: R_CC_Y (correct/yes-confident), R_IC_Y (incorrect/confident), R_CC_N (correct/not-confident), and R_IC_N (incorrect/not-confident). For each correct–confidence pair, mean (dashed line) and individual (symbols) data are provided. The box indicates ± 1 *SD* of the mean. Symbol type and fill reflect child age. Three children were not included in the R_CC_N or R_CC_N calculations because all trials were coded as confident. yr = years.



[5] 95% Cedible intervals were computed with Bayesian methods, using a beta-binomial model and assuming a noninformative Jeffreys prior for the hit rate (Gelman et al., 2013).

the child–observer team performed well in the experiment. Complementary to that result, there was a low proportion of trials in which the observer was falsely confident of the judgment.

For the limited number of trials in which the observer was not confident in the interval selection, two different patterns were possible in the data (R_CC_N vs. R_IC_N). One possibility was that, in these trials, the observer was essentially guessing, because either the child did not provide any behavior or the response was weak and was not detected by the observer. In this case, we should see the same amount of correct and incorrect responses ($R_CC_N = R_IC_N$). Alternatively, if the child reliably provided subtle responses (e.g., eye movement, looking to test assistant) when a signal was heard, we should see a higher proportion for R_CC_N than for R_IC_N. This pattern would suggest that the observer was accurately judging the child's behavior despite limited confidence in the judgment. Only those children who had at least one trial that was coded as not confident were included in this analysis. Thus, three children (3.34, 5.10, and 5.99 years old) were excluded in the R_CC_N and R_IC_N boxplots. In Figure 5, boxplots for R_CC_N and R_IC_N are similarly distributed, with similar means. A Wilcoxon signed-ranks test failed to detect a significant difference in means between R_CC_N and R_IC_N (p = .62). Although there were a limited number of trials in which observers coded their interval section as not confident, these results suggest that, for these trials, observers were not using reliable, subtle auditory behavior to make their judgments.

Discussion

The goal of this study was to determine if the PlayO2I method is feasible and reliable for measuring hearing in 2- to 7-year-old children. The PlayO2I method (Bonino & Leibold, 2017) is a two-interval, forced-choice observer-based procedure that requires the observer to determine if the signal was presented in Interval 1 or Interval 2 based on the child's behavior. Children are taught to perform a time-locked, play-based response when the signal is detected. The approach used in the current study evaluated performance over a test session in which the 1000-Hz warble tone signal was presented at a fixed, suprathreshold intensity level (30 dB SPL). Results from this study indicate that almost all children can perform at least 100 trials and that performance is stable across the test session. The PlayO2I method appears to be a powerful research tool for measuring hearing in 2- to 4-year-old children. Data collected with the PlayO2I method will allow us to map out how auditory development unfolds during the toddler and preschooler years—a time period where we know little about development. Having a comprehensive model of auditory development is clinically significant because it will pave the way for developing clinical methods for monitoring auditory functioning in young children with hearing loss.

Feasibility of the PlayO2I Method

Data collected with the PlayO2I method indicate that most 2- to 4-year-old children can be trained to perform this task. Of the twenty 2- to 4-year-old children recruited in the current study, all but one child (2.47 years old) was able to successfully complete training. Moreover, seven of the eight 2-year-olds—six of which were < 2.5 years old—could perform the task. The yield rate of 2- to 4-year-olds in the current study (95%) is slightly higher than our previous work that evaluated thirty-three 2- to 4-year-old children with the same method (Bonino & Leibold, 2017). That study resulted in an 88% yield rate, with data being successfully collected from five of the eight 2-year-olds. This difference is likely due to a change in transducer type (insert earphone vs. light-weight headphones), as Bonino and Leibold (2017) reported that the most common reason for unusable data was due to intolerance of the insert earphone. Results from the PlayO2I method indicate that the yield rate is high and it is at least comparable to, if not better than, the clinical CPA procedure. While studies evaluating CPA have consistently reported high yield rates (≥ 95%) for 3- and 4-year-old children, yield rates for 2-year-old children are variable (e.g., Barr, 1955; Kemaloğlu, Gündüz, Gökmen, & Yilmaz, 2005; Nielsen & Olsen, 1997; G. Thompson & Weber, 1974; M. Thompson et al., 1989). Across these studies, yield rate for 2-year-old children ranged from 20% to 70% for children < 2.5 years old and from 80% to 90% for children > 2.5 years old. Differences across these studies are likely a result of methodology differences (i.e., inserts vs. sound field; one vs. two experimenters).

In addition to determining the number of children who could perform the task, of particular interest was to determine how age affected the number of trials children completed prior to demonstrating signs of habituation or fatigue. Previous research indicates that 2-year-olds provide a limited number of responses with the clinical CPA procedure (e.g., Nielsen & Olsen, 1997; M. Thompson et al., 1989; G. Thompson & Weber, 1974). For example, M. Thompson et al. (1989) measured the number of trials 24- to 27-month-old toddlers could complete with CPA prior to habituation. The habituation criterion was met when four out of five consecutive stimulus trials were coded as a "no response" for the 2-s noise signal presented at 60 dB SPL. Of the 2-year-olds who could perform the task (n = 15, 68%), the mean number of trials performed was 28.33 (SD = 18.97) prior to habituation. No child was able to complete more than 50 trials. Similar findings to M. Thompson et al. (1989) have also been reported for CPA in clinical settings (e.g., Kemaloğlu et al., 2005; Nielsen & Olsen, 1997). Nielsen and Olsen (1997) reported that audiologists were able to obtain ≥ 6 clinical thresholds in one testing session on 10% of 24- to 29-month-olds and 40% of 30- to 35-month-olds. In contrast, nearly 75% of 3-year-olds tested completed ≥ 6 thresholds. Based on this previous work, we predicted that the number of trials children could complete in a single test session would increase as a function of age. However, we did not detect a developmental trend; nearly all children tested were able to complete 100 test trials. Of particular surprise was that, of the 2-year-olds who could perform the task, all seven were able to complete 100 trials in Stage 3.

There are several possible explanations for why 2-year-old children tested with the PlayO2I method had better outcomes—in terms of the number of trials completed—than what has been previously reported for the CPA procedure. One potential explanation is that the families and children who have volunteered to participate in this research study may not be representative of the general population, as has been observed for other auditory investigations (e.g., Lalonde & Holt, 2014). While we did not directly measure children's language or cognitive abilities, parents did complete a developmental questionnaire that probed communication, motor, socialization, and daily living skills. Based on their Adaptive Behavior Composite scores from the Vineland-3 (Sparrow et al., 2016), all 2-year-olds tested in the current study had an overall adaptive functioning score that was within 1 *SD* of the normative mean. Thus, results from the Vineland suggest that our sample of children have adaptive behaviors that are consistent with the general population. Furthermore, our sample of children is likely representative of the children who are recruited by university laboratories through large child databases of families who are willing to participate in research. Another potential explanation is that children are less likely to habituate to the warble tone at 30 dB SPL used here than they are for the stimuli used in other studies. For example, limited data may have been obtained by Nielsen and Olsen (1997) because they were varying signal intensity in order to measure thresholds. However, M. Thompson et al. (1989) reported that 2-year-olds habituated after 28 trials for a noise sample presented at a fixed level of 60 dB SPL. Furthermore, 2-year-olds' mean response rate appears to be independent of signal level, assuming the levels are not near threshold (30 vs. 45 dB HL; M. Thompson & Thompson, 1972). Thus, differences across the studies cannot be accounted for by stimulus differences.

An alternative, and more likely, explanation is that the task reinforcement strategy used in the current study reduced the risk of habituation for young children. Following the standard procedure for CPA, children were taught to perform a motor-based response each time they heard the signal in the context of a game. The test assistant was required to change the game frequently and often increase the complexity of the game over time. On average, children played each game for 11.6 trials (*SD* = 2.65). All children were provided frequent social reinforcement. In addition, three children had the mechanical toys activated. In contrast to this approach, it appears that other researchers trained children to perform a single game for the duration of the testing session (e.g., putting a block in a box or a ring on a spindle) and provided frequent social reinforcement (Nielsen & Olsen, 1997; M. Thompson et al., 1989). The unique parameters of our task reinforcement strategy—frequently changing the game and manipulating the complexity of the game—appear to have successfully guarded against the risk of habituation (e.g., Lloyd, 1966; Rankin et al., 2009). It is not clear from our data why these two parameters guarded against habituation. One potential explanation is that manipulating these two parameters may have increased children's motivation, resulting in strengthening of the behavioral response (e.g., McSweeney & Murphy, 2014). A second possible explanation is that frequently changing the game and/or manipulating the complexity of the game served as an additional source of reinforcement. Although the game is traditionally viewed as part of the response in CPA, a more limited definition of the response is to include only the initial motor movement of the hand from the "ready state" (e.g., cheek). Under this interpretation, interacting with the game (e.g., putting the peg in the board) could be considered a form of reinforcement, and thus, changing the game introduces novel reinforcement. The third possible explanation is that the use of multiple reinforcers—games, social interaction, and mechanical toys—on different reinforcement schedules may have contributed to the number of trials obtained in the current study (e.g., Lloyd, 1966; Moore, Thompson, & Thompson, 1975; Primus & Thompson, 1985). Regardless of the underlying mechanism(s), it appears that the task reinforcement strategy used in the current study resulted in the completion of more trials than has been previously reported in the CPA literature (e.g., Nielsen & Olsen, 1997; M. Thompson et al., 1989). This finding is clinically significant because the primary elements of our task reinforcement strategy (i.e., switching the game, manipulating the complexity of the game) can be incorporated into a CPA test session to increase the number of thresholds obtained during a single clinical visit. Future research is needed to understand which elements of our task reinforcement strategy are the most effective at reducing the risk of habituation and how and when they should be combined to result in maximizing the number of trials in research and clinical settings.

Quality of Data Obtained With the PlayO2I Method

Results from this study indicate that overall hit rate was high across all child–observer teams ($M = 0.94$). For the group of 2-year-olds, hit rate ranged from 0.88 to 0.99 ($M = 0.92$, $SD = 0.04$). These findings are consistent with the 0.88 ($SD = 0.12$) hit rate reported for 24- to 27-month-olds with CPA (M. Thompson et al., 1989). Although there was limited variability in hit rate (0.88–1.0) across child–observer teams, 35% of the variability was accounted for by listener age ($\log_{10}$ years). Based on the linear model, the predicted hit rates for 2-, 3-, and 4-year-olds were 0.91, 0.94, and 0.96, respectively.

Because the signal was presented at a suprathreshold intensity level, the measured hit rate should correspond to the upper asymptote of the listener's underlying psychometric function. Thus, our results suggest that upper asymptotic performance improves across the age span tested here. This finding is consistent with the observation that upper asymptote for psychoacoustic tasks is different at two time points in development: ~0.85 for infants (e.g., Bargones, Werner, & Marean, 1995) and ~0.95 for school-age children (e.g., Buss, Hall, & Grose, 2009). In contrast, adults' mean upper asymptote was ≥ 0.97 in these studies (Bargones et al., 1995; Buss et al., 2009). Developmental effects for upper asymptotic performance can be interpreted to indicate that, with increased age, there are fewer trials in which the listener experiences inattention. Models of general inattention indicate that lapses in attention also result in elevated threshold and shallower slope estimates of the psychometric function (e.g., Viemeister & Schlauch, 1992; Wightman & Allen, 1992). Thus, our data highlight the importance of the recommendation to allow upper asymptote to be a free parameter during the fitting of children's psychometric functions in order to account for developmental differences in upper asymptotic performance (e.g., Buss et al., 2009; Manning, Jones, Dekker, & Pellicano, 2018). Moreover, Manning et al. (2018) recently suggested that developmental data collected with adaptive paradigms can be reprocessed with psychometric fitting software (psignifit toolbox; Wichmann & Hill, 2001) to address this concern. However, this effort may not be warranted for studies examining threshold differences with the PlayO2I method because the extent of inattentiveness reported here likely has a minimal effect on threshold estimates (Viemeister & Schlauch, 1992).

In addition to developmental differences in overall hit rate, we evaluated our data for possible shifts in performance over the course of the test session for child–observer teams. In general, hit rate appears to be relatively stable. Specifically, hit rate was similar for the first 30 trials and the last 30 trials. Moving average plots provide further support that performance was stable for test sessions of 3- to 7-year-old children across the 100 trials in Stage 3 of testing. In contrast, moving average data suggest that there may be a slight deterioration in data collected from 2-year-old children after 50 trials of testing. However, hit rate appears to recover later in the track. It is not clear why child–observer teams evaluating 2-year-olds had a temporary shift in performance, but potential explanations include the child or observer having experienced increased inattentiveness, increased rate of off-task child behavior during trials (e.g., talking), and changes in the child's response (e.g., slower to respond, less robust response) that reduced the observer's ability to make a correct judgment.

It is not clear the extent to which this deterioration in performance would affect the quality of the data collected from 2-year-old children. For studies using an adaptive paradigm, a single threshold estimate can be completed in fewer than 50 trials. Thus, performance is likely stable over the entire run, although it is possible that a deterioration in performance may occur prior to 50 trials when presentation levels are near threshold. Shifts in performance can be monitored by presenting probe trials throughout the testing block. Probe trials are the signal presented at a clearly audible level (typically the same intensity level as training) throughout data collection. As recommended by Bonino and Leibold (2017), two probe trials should be presented for every block of 12 trials and test sessions are required to have a probe hit rate of ≥ 0.8 in order for the data to be considered "useable." In future studies, this recommendation could be expanded upon by increasing the number of probe trials and monitoring for changes in probe hit rate across the test session. Data sets with evidence of nonstationary inattentiveness could be reprocessed by fitting psychometric functions to account for upper asymptote being < 1.0 (e.g., Manning et al., 2018). Another approach for monitoring of shifts in performance would be to test children with multiple interleaved adaptive tracks (e.g., Leek, 2001). If the tracks diverge, it is assumed that the underlying psychometric function is unstable. Further research is currently being conducted in our laboratory to understand the utility of these approaches.

The Advantages of the PlayO2I Method

There are several advantages associated with this method. The first advantage is that children as young as 2.0 years old can perform this task. Combining results from the current study and from our previous work (Bonino & Leibold, 2017), 12 out of the sixteen 2-year-olds recruited were able to perform the task. Moreover, seven out of the eight 2-year-olds tested in the current study were able to complete training and 100 experimental trials (Stage 3 of testing) in a single 1-hr visit. Thus, it appears that the response task and reinforcers employed are developmentally appropriate and engaging for this age group. The PlayO2I method likely also has less cognitive demand than other forced-choice tasks that have been successfully used with 2-year-olds in research settings. For example, the change/no-change speech-sound discrimination task requires children to select between two responses (Holt & Lalonde, 2012), and the CRISP-Jr. picture pointing task requires children to select a target word from an array of four choices (Garadat & Litovsky, 2007; Hess et al., 2018). In contrast to these procedures, the PlayO2I method places the burden of selecting the target interval on the observer, not the child. Moreover, the PlayO2I method allows observers to base their judgment on subtle auditory behavior provided by the child (e.g., eye movement, eyebrow furrow), in addition to the targeted motor response. These design features make the PlayO2I method feasible for young children who do not yet have the cognitive abilities to perform a traditional forced-choice paradigm. Furthermore, it may also be possible to modify this procedure to test children with motor or cognitive impairments who may have inconsistent or delayed responses (Browning, Buss, Porter, McLean, & Leibold, 2017; Porter, Buss, Browning, & Leibold, 2018).

The second advantage is that the PlayO2I method allows researchers to obtain a substantial number of trials from young children. A common dilemma faced by developmental researchers is that the number of trials and/or conditions has to be reduced in order to ensure that the study is feasible for young children. For example, in our previous work, we limited data collection to one adaptive track (30–45 trials) per visit for 2- and 3-year-old children (Bonino & Leibold, 2017). However, the current data suggest that young children may be able to perform upward of 100 trials during a single test session. Further research is needed to verify that this recommendation continues to be appropriate for testing near threshold or for complex stimuli. Researchers are also able to capture more data with the utilization of a two-alternative, forced-choice paradigm than with other implementations of OPP. The traditional implementation of OPP requires the researcher to present a substantial number of no-signal trials (i.e., a 1:1 signal to no-signal ratio). In contrast, the two-interval design presents a signal (or probe signal) every trial. This approach may also reduce the risk of habituation because children experience a minimal delay between trials and reinforcement opportunities.

The third advantage of the PlayO2I method is that it guards against some forms of bias. The observer (and test assistant) is blind to which interval contained the signal in Stages 2 and 3 of testing. This design feature controls for the observer making a selection based on his or her expectations of the child's sensitivity, thus controlling for observer bias. It is unclear if the PlayO2I method also guards against response bias. Historically, it has been thought that a two-interval, forced-choice procedure is free from response bias (e.g., Green & Swets, 1966). Based on this perspective, it was proposed by Browning et al. (2014) that their two-interval adaptation of OPP guarded against response bias. However, that assumption has come under question for two reasons. The first reason is that, in the two-interval, observer-based procedure, the task is not a two-interval task from the perspective of the listener. In previous OPP studies where the listening interval was not defined to the listener, adults and school-age children tended to adopt a conservative decision-making strategy (e.g., Bonino et al., 2013; Leibold & Werner, 2006). In contrast, infants showed no response bias (e.g., Leibold & Werner, 2006). It is not known what the decision strategy tends to be of toddlers and preschoolers. Thus, potential developmental differences in response bias may interfere with the interpretation of data collected over a wide age span of childhood. The second reason is that forced-choice tasks may not be free from response bias. For example, Jones, Moore, et al. (2015) reported that naive adult listeners showed evidence of a bias toward selecting an interval based on which interval had been correct on previous trials. However, with subsequent practice, this bias was eliminated. One interpretation of the data from Jones et al. is that observers testing with the PlayO2I method may be bias-free in their selection of an interval, assuming that they have extensive practice with the method. Future research is needed to better understand how response bias affects data collected from the child–observer team within the context of the PlayO2I method.

The fourth advantage of this method is that it may be easier to train new observers and to establish reliability between observers compared to OPP. Results from this study confirm that children are able to make a fast, time-locked response to the signal that observers can accurately and confidently judge across the two temporal observation windows. Moreover, high observer confidence throughout the test session implies that children continued to provide a strong response. These results also confirmed that a 1,065-ms observation interval and a 500-ms ISI were appropriate. However, temporal constraints may need to be modified for other stimuli or for testing children who have developmental delays as their response time may be delayed and/or variable (Browning et al., 2017; Porter et al., 2018). One benefit of a two-interval task is that there is limited memory load on the observer because his or her judgment is based on the child's behavior across the two intervals within a given trial. In contrast, OPP requires the observer to make a yes/no decision based on his or her observation of the child's behavior over multiple trials. This reduction in memory load may reduce the training time needed for new observers to obtain reliable data with the PlayO2I method compared to OPP. Furthermore, reduction in memory load may also reduce variability across observers. Current work in our laboratory is examining the interobserver reliability of the PlayO2I method. We are also developing video training modules to ensure that all new observers are reliable with the method prior to testing children. This work will also guide us in better understanding how much training observers need with testing young children and with the PlayO2I method before reliable data can be collected in the laboratory.

Conclusions

The PlayO2I method provides a powerful tool for researchers that can be used to measure detection and discrimination abilities in toddlers and preschoolers. This method is feasible starting at a developmental age of 2.0 years. Results from the current study indicate that children were able to consistently provide a fast, time-locked response that was accurately and confidently judged by observers. Moreover, performance was stable across a long test session (100 trials). Future studies are planned to examine interobserver reliability and stability of the underlying psychometric function during adaptive testing. Training materials are currently being developed to promote the training of new testers across multiple laboratories.

Acknowledgments

This work was supported by the American Academy of Audiology/American Academy of Audiology Foundation Research Grants in Hearing and Balance Program (to A. Y. B.). The authors are grateful to the members of the Children's Auditory Perception Laboratory at the University of Colorado Boulder for their assistance with data collection.

References

Allen, P., & Wightman, F. (1992). Spectral pattern discrimination by children. *Journal of Speech and Hearing Research, 35,* 222–233.

Allen, P., & Wightman, F. (1994). Psychometric functions for children's detection of tones in noise. *Journal of Speech and Hearing Research, 37,* 205–215.

Bargones, J. Y., Werner, L. A., & Marean, G. C. (1995). Infant psychometric functions for detection: Mechanisms of immature sensitivity. *The Journal of the Acoustical Society of America, 98*(1), 99–111.

Barr, B. (1955). Pure tone audiometry for preschool children; a clinical study with particular reference to children with severely impaired hearing. *Acta Oto-Laryngologica: Supplementum, 121,* 1–84.

Bonino, A. Y., & Leibold, L. J. (2017). Behavioral assessment of hearing in 2 to 4 year-old children: A two-interval, observer-based procedure using conditioned play-based responses. *Journal of Visualized Experiments, 119,* e54788.

Bonino, A. Y., Leibold, L. J., & Buss, E. (2013). Effect of signal-temporal uncertainty in children and adults: Tone detection in noise or a random-frequency masker. *The Journal of the Acoustical Society of America, 134,* 4446–4457.

Browning, J., Buss, E., & Leibold, L. J. (2014). Preliminary evaluation of a two-interval, two-alternative infant behavioral testing procedure. *The Journal of the Acoustical Society of America, 136*(3), EL236–EL241.

Browning, J. M., Buss, E., Porter, H., McLean, H., & Leibold, L. J. (2017, April). *A two-interval observer-based procedure to test hearing in children with developmental disabilities*. Poster presentation at the American Academy of Audiology's AudiologyNow! Conference, Indianapolis, IN.

Buss, E., Hall, J. W., & Grose, J. H. (2009). Psychometric functions for pure tone intensity discrimination: Slope differences in school-aged children and adults. *The Journal of the Acoustical Society of America, 125,* 1050–1058.

Buss, E., Hall, J. W., & Grose, J. H. (2012). Development of auditory coding as reflected in psychophysical performance. In L. A. Werner, R. R. Fay, & A. N. Popper (Eds.), *Human auditory development, Springer handbook of auditory research* (Vol. 42, pp. 107–136). New York, NY: Springer.

Buss, E., Leibold, L. J., Porter, H. L., & Grose, J. H. (2017). Speech recognition in one- and two-talker maskers in school-aged children and adults: Development of perceptual masking and glimpsing. *The Journal of the Acoustical Society of America, 141,* 2650–2660.

Eisenberg, L. S., Martinez, A. S., & Boothroyd, A. (2007). Assessing auditory capabilities in young children. *International Journal of Pediatric Otorhinolaryngology, 71,* 1339–1350.

Garadat, S. N., & Litovsky, R. Y. (2007). Speech intelligibility in free field: Spatial unmasking in preschool children. *The Journal of the Acoustical Society of America, 121*(2), 1047–1055.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.) New York, NY: Chapman and Hall/CRC.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Los Altos Hills, CA: Peninsula Publishing.

Hess, C. L., Misurelli, S. M., & Litovsky, R. Y. (2018). Spatial release from masking in 2-year-olds with normal hearing and with bilateral cochlear implants. *Trends in Hearing, 22,* 1–13.

Holt, R. F., & Lalonde, K. (2012). Assessing toddlers' speech-sound discrimination. *International Journal of Pediatric Otorhinolaryngology, 76,* 680–692.

Jensen, J. K., & Neff, D. L. (1993). Development of basic auditory discrimination in preschool children. *Psychological Science, 4*(2), 104–107.

Jones, P. R., Kalwarowsky, S., Braddick, O. J., Atkinson, J., & Nardini, M. (2015). Optimizing the rapid measurement of detection thresholds in infants. *Journal of Vision, 15*(11), 1–17.

Jones, P. R., Moore, D. R., & Shub, D. E. (2015). The role of response bias in perceptual learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(5), 1456–1470.

Kemaloğlu, Y. K., Gündüz, B., Gökmen, S., & Yilmaz, M. (2005). Pure tone audiometry in children. *International Journal of Pediatric Otorhinolaryngology, 69*(2), 209–214.

Lalonde, K., & Holt, R. F. (2014). Cognitive and linguistic sources of variance in 2-year-olds' speech-sound discrimination: A preliminary investigation. *Journal of Speech, Language, and Hearing Research, 57*(1), 308–326.

Leek, M. R. (2001). Adaptive procedures in psychophysical research. *Perception & Psychophysics, 63*(8), 1279–1292.

Leibold, L. J., & Werner, L. A. (2006). Effect of masker-frequency variability on the detection performance of infants and adults. *The Journal of the Acoustical Society of America, 119*(6), 3960–3970.

Lloyd, L. (1966). Behavioral audiometry viewed as an operant procedure. In E. Holzhauer, K. Hoff, & E. Cherow (Eds.), *Hearing impaired developmentally disabled children and adolescents: An interdisciplinary look at a special population* (Sec. II, pp. 75–84). Rockville, MD: American Speech-Language-Hearing Association.

Manning, C., Jones, P. R., Dekker, T. M., & Pellicano, E. (2018). Psychophysics with children: Investigating the effects of attentional lapses on threshold estimates. *Attention, Perception, & Psychophysics, 80,* 1311–1324.

Marshall, L., & Jesteadt, W. (1986). Comparison of pure-tone audibility thresholds obtained with audiological and two-interval forced-choice procedures. *Journal of Speech and Hearing Research, 29,* 82–91.

Mayer, D. L., & Dobson, V. (1982). Visual acuity development in infants and young children, as assessed by operant prefential looking. *Vision Research, 22,* 1141–1151.

McSweeney, F. K., & Murphy, E. S. (2014). Characteristics, theories, and implications of dynamic changes in reinforcer effectiveness. In F. K. McSweeney & E. S. Murphy (Eds.), *The Wiley-Blackwell handbook of operant and classical conditioning* (pp. 339–368). Malden, MA: Wiley Blackwell.

Moller, A. R., & Rollins, P. R. (2002). The non-classical auditory pathways are involved in hearing in children but not in adults. *Neuroscience Letters, 319,* 41–44.

Moore, J. M., Thompson, G., & Thompson, M. (1975). Auditory localization of infants as a function of reinforcement conditions. *Journal of Speech and Hearing Disorders, 40*(1), 29–34.

Newman, R. S. (2011). 2-year-olds' speech understanding in multitalker environments. *Infancy, 16*(5), 447–470.

Nielsen, S. E., & Olsen, S. (1997). Validation of play-conditioned audiometry in a clinical setting. *Scandinavian Audiology, 26*(3), 187–191.

Olsho, L. W., Koch, E. G., Halpin, C. F., & Carter, E. A. (1987). An observer-based psychoacoustic procedure for use with young infants. *Developmental Psychology, 23*(5), 627–640.

Porter, H., Buss, E., Browning, J., & Leibold, L. J. (2018, May). *A new method to test hearing in children with motor or developmental impairment*. Poster session presented at the Third Annual Human Movement Variability Conference, Omaha, NE.

Primus, M. A., & Thompson, G. (1985). Response strength of young children in operant audiometry. *Journal of Speech and Hearing Research, 28*(4), 539–547.

Rankin, C. H., Abrams, T., Barry, R. J., Bhatnagar, S., Clayton, D., Colombo, J., . . . Thompson, R. F. (2009). Habituation revisited: An updated and revised description of the behavioral characteristics of habituation. *Neurobiology of Learning and Memory, 92*(2), 135–138.

Schneider, B. A., Trehub, S. E., Morrongiello, B. A., & Thorpe, L. A. (1986). Auditory sensitivity in preschool children. *The Journal of the Acoustical Society of America, 79*(2), 447–452.

Sparrow, S. S., Cicchetti, D. V., & Saulnier, C. A. (2016). *Vineland Adaptive Behavior Scales* (3rd ed.). Bloomington, MN: NCS Pearson.

Thompson, G., & Weber, B. A. (1974). Responses of infants and young children to behavior observation audiometry (BOA). *Journal of Speech and Hearing Disorders, 39*(2), 140–147.

Thompson, M., & Thompson, G. (1972). Response of infants and young children as a function of auditory stimuli and test methods. *Journal of Speech and Hearing Research, 15*(4), 699–707.

Thompson, M., Thompson, G., & Vethivelu, S. (1989). A comparison of audiometric test methods for 2-year-old children. *Journal of Speech and Hearing Disorders, 54,* 174–179.

Viemeister, N. F., & Schlauch, R. S. (1992). Issues in infant psychoacoustics. In L. Werner & E. W. Rubel (Eds.), *Developmental psychoacoustics* (pp. 191–209). Washington, DC: American Psychological Association.

Werner, L. A. (2017). Infants and children at the cocktail party. In J. C. Middlebrooks, J. Z. Simon, A. N. Popper, & R. R. Fay (Eds.), *The auditory system at the cocktail party, Springer handbook of auditory research* (Vol. 60, pp. 199–226). New York, NY: Springer International Publishing.

Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics, 63*(8), 1293–1313.

Wightman, F., & Allen, P. (1992). Individual differences in auditory capability among preschool children. In L. A. Werner & E. W. Rubel (Eds.), *Developmental psychoacoustics* (pp. 113–133). Washington, DC: American Psychological Association.