**QUEEN'S
UNIVERSITY
BELFAST**

Measuring auditory discrimination thresholds in preschool children: An empirically based analysis

Gillen, N., Agus, T., & Fosker, T. (2019). Measuring auditory discrimination thresholds in preschool children: An empirically based analysis. *Auditory Perception and Cognition*, 1(3-4), 173-204.
<https://doi.org/10.1080/25742442.2019.1600970>

Published in:
Auditory Perception and Cognition

Document Version:
Peer reviewed version

Queen's University Belfast - Research Portal:
[Link to publication record in Queen's University Belfast Research Portal](#)

Publisher rights
© 2018 Informa UK Limited, trading as Taylor & Francis Group. This work is made available online in accordance with the publisher's policies. Please refer to any applicable terms of use of the publisher.

General rights
Copyright for the publications made accessible via the Queen's University Belfast Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The Research Portal is Queen's institutional repository that provides access to Queen's research output. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact openaccess@qub.ac.uk.

Measuring auditory discrimination thresholds in preschool children: An empirically based analysis

Nicola Gillen¹, Trevor Agus², Tim Fosker^{1, CA}

¹ School of Psychology, Queen's University Belfast

² SARC, School of Arts, English and Languages, Queen's University Belfast

Corresponding Author:

Dr Tim Fosker

School of Psychology

Queen's University Belfast

David Keir Building

18-30 Malone Road

Belfast, BT9 5BN

Northern Ireland, United Kingdom

Tel: +44 (0) 28 9097 5576

Email: t.fosker@qub.ac.uk

Word Count (excluding references): 13,413

Keywords: auditory discrimination; adaptive procedures; frequency discrimination; preschool auditory development

Abstract

Auditory discrimination is an important perceptual skill that seems to develop substantially during early childhood and is predictive of key developmental outcomes like language ability. However, the estimation of reliable auditory discrimination thresholds is impeded by non-sensory limitations in young children that impact task performance. Here we used computerized simulations of child-like and adult-like performance, as well as novel behavioural task modifications with 3 and 4 year old preschool children and adults, to investigate key parameters in the successful estimation of auditory discrimination thresholds in preschool-aged children. The results indicated the most suitable adaptive procedure that is not widely used with young children (75% weighted 1-up 1-down procedure), an appropriate number of trials (20) and step size (relatively small). In addition, using a novel manipulation of level of physical engagement, it was found that threshold estimates are more reliable when children were given a task that involved greater physical engagement (i.e. 3D objects, rather than 2D images on a screen). The overall results provide recommendations for designing procedures to estimate thresholds in preschool or developmentally delayed children.

Introduction

Perception of the acoustic differences between sounds (auditory discrimination) is important for the development of a number of skills used in everyday life. For example, the ability to recognise a person by their voice, locate the origin of a sound, and recognise fluent speech in a range of environments, are all somewhat dependant on the ability to reliably discriminate sounds on different acoustic parameters (Chermak, Somers, & Seikel, 1998). At a very minimum, the ability to discriminate relatively subtle temporal changes in pitch and amplitude are essential for the development of speech and language (Fitch, Miller, & Tallal, 1997). In fact, individual differences in auditory discrimination abilities are significant predictors of differences in the development of language and literacy during the primary (elementary) school years (Talcott et al., 2002). In addition, children identified with auditory discrimination deficits have difficulties remembering complex or multistep instructions, are easily distracted and have difficulty comprehending speech in the presence of background noise (Chermak et al., 1998; Ferguson, Hall, Riley, & Moore, 2011). Therefore, reliable quantification of auditory discrimination prior to formal education has the potential to form the foundation of future screening tests for assessing the risk of developing language and literacy difficulties. There is already some evidence that measuring the auditory discrimination abilities of children just prior to, or just beginning, formal education may be useful in identifying children at risk of later literacy difficulties (Boets et al., 2011; Corriveau, Goswami, & Thomson, 2010; Law, Vandermosten, Ghesquière, & Wouters, 2017). However, establishing robust auditory markers of risk for future language and literacy difficulties have been hindered by difficulties in reliably estimating the auditory discrimination capacities of preschool children. Arguably, these difficulties stem from limitations in the cognitive capacities of preschool children, relative to the adults that typical auditory discrimination tasks have been designed for.

The auditory discrimination abilities of preschoolers above 2 years of age are typically measured in the same way as adults, by estimating discrimination thresholds. Auditory discrimination thresholds identify the minimum amount of difference required between two sounds to reliably tell them apart, lower threshold estimates indicating that an individual can discriminate smaller differences between sounds. Importantly, the auditory discrimination thresholds of preschool children (4 – 6 year olds) have been shown to be significantly larger and more variable than the thresholds of older children (> 10 years old) and adults (Banai & Yifat, 2011; Jensen & Neff, 1993; Moore, Ferguson, Halliday, & Riley, 2008). Some studies using common threshold estimation procedures have proposed that the differences in threshold estimates found for preschoolers compared with older children and adults, may be the result of underdeveloped sensory capacities (Jensen & Neff, 1993; Schneider, Trehub, Morrongiello, & Thorpe, 1989). However, studies using observation of habituation and novelty preference to sound differences (e.g. head-turn and preferential looking paradigms) have shown that by 6 months old, infants can discriminate smaller differences between sounds than many preschoolers (Litovsky, 2015; Olsho, Koch, & Halpin, 1987; Olsho, Schoon, Sakai, Turpin, & Sperduto, 1982). This has led to the assertion that the methods used to estimate the discrimination thresholds of school-aged children and adults may not be robust with preschoolers (Litovsky, 2015). Similarly, the habituation procedures used with infants may not be a suitable alternative for use with preschoolers, because preschoolers are less likely to make the same clear overt responses to expectancy violations that infants do.

Auditory discrimination thresholds are typically estimated using adaptive (staircase) procedures that require the presentation of only a subset of the possible differences between the sounds being discriminated. A staircase procedure adjusts the size of the difference between sounds based on responses to previous trials in the task. Staircase procedures are

typically described as making upward steps following incorrect responses (making the difference between the sounds larger and thus easier to discriminate), and downward steps following correct responses (making the difference between sounds smaller and thus harder to discriminate). For example, a 1-up 2-down staircase requires correct performance on two consecutive trials before the difference between sounds is made smaller and one incorrect response before the difference is made larger. Discrimination thresholds are then estimated as the average of the stimulus values at a given number of reversals closest to the end of the procedure (changes in staircase direction from up to down or down to up).

Staircase procedures differ in a number of characteristics that impact their ability to estimate the discrimination thresholds of individuals reliably. Specifically, staircase procedures differ in: (1) the target level of performance, i.e. the proportion of correct responses expected to be observed at the end of the staircase procedure; (2) the number of correct consecutive trials that must be attained before the staircase can step up or step down (step rule); (3) the relative size of adjustment that is made to the difference between sounds after a correct or incorrect trial (step size); (4) the absolute value of the step size typically derived from the starting discrimination. The target threshold (1) is closely linked to the rules governing when a change is made (2) and the relative sizes of those steps (3) through statistical assumptions, whereas the absolute size of the steps (4) is theoretically more independent but may affect the reliability with which the target threshold (1) is measured and even introduce systematic biases. To use these staircase procedures in test paradigms, decisions also have to be made as to the size of the discrimination on the first trial (starting discrimination); when to terminate the procedure (termination rule); the response paradigm to use; and how the response paradigm will actually be presented to participants.

Three different staircase procedures, designed to estimate discrimination thresholds based on relatively different proportions of correct performance, account for around 65% of

the 84 studies published from 2013-2016 that measured auditory discrimination thresholds in children and adults (Scopus using search term “auditory discrimination threshold” and selecting all experimental studies which estimated auditory discrimination thresholds with non-speech stimuli published between 2013-2016). Two of these staircases are transformed up-down procedures described by Levitt (1971) targeting 71% correct performance (1-up 2-down) or 79% correct performance (1-up 3-down), and the final one is the weighted up-down (1-up 1-down) procedure described by Kaernbach (1991). All of these staircase procedures (step rules) have been used in paradigms estimating auditory thresholds with children. The transformed up-down procedure targeting 71% correct performance has been used widely with children as young as 4 years old (Jensen & Neff, 1993; Killan, Royle, Totten, Raine, & Lovett, 2015), with one study including children as young as 3 years old using this procedure (Porter, Grantham, Ashmead, & Tharpe, 2014). The transformed procedure targeting the 79% correct point has also been used with children as young as 4 years old (Banai & Yifat, 2011; Wang, Chen, Chiang, Lai, & Tsao, 2016). However, the weighted up-down procedure, using the 1-up 1-down step rule, has only been used with children 7 years old and older (Rickard, Heidtke, & O’Beirne, 2013). It is worth noting that the reliability of the thresholds estimated in these studies was not assessed.

The transformed up-down procedures identified above (1-up 2-down, 1-up 3-down) use the same step size for up and down movements, resulting in threshold convergence at 70.7% correct performance and 79.4% correct performance respectively. In contrast, the weighted up-down procedure is adjusted on every trial (1-up 1-down) with the size of the asymmetry in the up and down movements being adjusted to target different levels of correct performance. Typically studies using the weighted up-down procedure have chosen to focus on asymmetric step sizes for up and down movements which are designed to target 75% correct performance. In fact, of the 74 studies published before 2017, detailing their use of

the weighted up-down procedure, 88% opted to estimate thresholds based on the 75% correct performance level (Scopus search for all experimental studies which cited Kaernbach 1991 as the procedure used to estimate thresholds).

Staircase procedures are typically conducted under forced-choice response paradigms. These consist of the presentation of a number of stimulus intervals (sounds, termed n Interval) and a number of alternative choices of these intervals (termed n Alternative Forced Choice), from which participants are to identify the interval which contains the target stimulus. Commonly, the number of alternative choices implemented is two (3 interval- 2 alternative forced choice; 3I-2AFC) or three (3 interval- 3 alternative forced choice; 3I-3AFC), in which participants must select the stimulus that is the “odd-one-out”.

Based on the characteristics of auditory discrimination tasks outlined above, attention is perhaps the most important cognitive skill that is sensitive to the staircase procedures used to reliably estimate auditory discrimination thresholds. Attention is considered to involve a number of overlapping processes of orientation, vigilance, sustained attention and selective attention. However, the temporal nature of auditory discrimination tasks, relying on many successive trials, is considered to load most heavily on sustained attention (Gomes, Molholm, Christodoulou, Ritter, & Cowan, 2000). Sustained attention is the specific ability to maintain attentional focus on a specific task for a prolonged period of time. Sustained attention is known to still be developing during the preschool years, not reaching adult-like levels until around 10 to 13 years of age (Betts, McKay, Maruff, & Anderson, 2006; Lin, Hsiao, & Chen, 1999). In fact, young children (6 years old and younger) have been found to struggle to sustain attention as the duration and difficulty of a task increases (Betts et al., 2006; Mahone & Schneider, 2012). Typical threshold estimation tasks require one or both of these elements, i.e. usually being relatively long in duration, requiring quite a large number of trials of

repeating stimuli, and as participants near their threshold region (the point where they can just discriminate a difference between the sound stimuli), the difficulty of the task increases.

Fundamentally related to the ability to maintain attention on a task is the ability to retain and manipulate the attended to information for a brief period of time (working memory). Similar to attention, there is evidence of significant age-related improvements in working memory skills during the preschool years (Carlson, Moses, & Breton, 2002), with adult-like levels being reached somewhere between the ages of 13 and 15 years old (Luciana, Conklin, Hooper, & Yarger, 2005). A fundamental level of working memory capacity has been identified as important for reliable performance in auditory discrimination tasks (France et al., 2002; Moore, Ferguson, Edmondson-Jones, Ratib, & Riley, 2010), and working memory performance itself correlates somewhat with auditory discrimination threshold estimates in adulthood (18.5 % of shared variance; Zhang et al., 2016).

Overall, the variability that is observed in the auditory discrimination thresholds of preschool children has been suggested to be a consequence of the varying demands imposed by auditory discrimination task parameters coupled with the restricted and developing cognitive capacities (including attentional and working memory capacities) of preschoolers. Here we aimed to establish parameters for auditory discrimination testing that are most robust to the ‘non-sensory’ or ‘non-auditory’ factors that are known to be highly variable in young children and negatively impact accurate threshold estimation (Dawes et al., 2009; Wightman & Allen, 1992; Wightman, Dolan, Allen, Jamieson, & Kistler, 1989).

The present study used computational simulations and behavioural testing to investigate the impact of several parameters (including age) on the reliability of auditory discrimination threshold estimates. Experiment 1 simulated the performance of 60,000 participants to establish which of three selected staircase procedures led to threshold estimates that were most robust to varying levels of inattention. Firstly, the robustness of each

staircase procedure was evaluated in terms of varying numbers of presented trials (Experiment 1). Secondly, the threshold estimates produced by the selected staircase (based on findings from Experiment 1) were evaluated in terms of the step size (Experiment 2). Finally, the findings of the simulations were applied to a group of preschoolers (3 and 4 year olds) and a control group of adults in the context of a frequency discrimination task. Thresholds, threshold variability, threshold validity and re-test reliabilities were investigated based on age, response paradigm (3I-2AFC and 3I-3AFC), and level of physical engagement with the task (Experiment 3).

Experiment 1: Selecting a staircase procedure based on successful threshold estimation with the smallest number of trials

The reliability and accuracy of threshold estimates are fundamentally impacted by the number of trials presented during a staircase procedure (Rose, Teller, & Rendleman, 1970). Provided attention is maintained throughout the procedure, more trials will typically lead to a greater number of reversals and a commensurate improvement in the reliability and accuracy of threshold estimates (Witton, Talcott, & Henning, 2017). However, procedures containing relatively large numbers of trials are also time-consuming and increase the likelihood that younger participants like preschoolers may struggle to sustain attention throughout the task. Typical threshold estimation procedures with adults, and many with preschool children, terminate based on a fixed number of reversals to ensure a high level of reliability and accuracy. However, the exact number of trials required for participants to reach a fixed number of reversals will vary somewhat from participant to participant. For adult participants, who are likely to have relatively similar sustained attention spans, this termination rule will result in them completing procedures in similar numbers of trials.

However, this is unlikely to be the case for preschool-aged participants. Considerable individual differences in the sustained attention spans of preschoolers are likely to result in threshold procedures of substantially different durations and consequently marked variation in the resulting threshold estimates. In contrast, termination once a fixed number of trials has elapsed allows for control over the duration of the task and potentially the reliability across participants of varying sustained attention capacities.

Studies that have identified very large inter-individual variability in the threshold estimates of preschoolers may be considered to be the most demanding for sustaining attention, by having a considerably large number of trials (e.g. Banai & Yifat, 2011 having 100 trials). Somewhat counterintuitively, measures with the largest number of trials may also be the least reliable with young children, because the increased duration of these procedures means that they are more likely to contain lapses in attention. Although the sustained attention capacities of preschoolers are highly variable and difficult to estimate within the context of any one task, Mahone and Schneider (2012) after reviewing a wide range of tasks requiring sustained attention, suggest that tasks lasting less than 7 minutes are most appropriate for children under 5 years old. Certainly, most current auditory discrimination tasks that successfully estimate accurate and reliable thresholds last substantially longer than 7 minutes.

The number of trials to include in a staircase procedure cannot be considered in isolation, but instead they should be considered in conjunction with several other staircase parameters including the starting discrimination level, step size and step rule. The step rule of a staircase will influence the number of trials required for a procedure to converge on a threshold estimate. For instance, transformed procedures require that a number of consecutive trials are answered correctly before the staircase steps down (in most studies, two or three consecutive correct responses are required), whereas the weighted 75% procedure consists of

steps occurring after every participant response. In both behavioural and computational studies, the weighted procedure has been found to converge on thresholds in approximately 10% less trials than the transformed procedure (Kaernbach, 1991; Rammsayer, 1992). However, these studies did not consider which staircase procedure is most efficient when the attention capacities of participants are variable.

An ideal staircase procedure should present enough trials to maximise the chance that the estimated threshold for participants is representative of their underlying sensory abilities. However, a procedure should not contain too many trials that may potentially lead to young participants losing attentional focus and thus increase the chance that they fail to converge on a threshold estimate that represents their underlying sensory abilities. While it is not possible to discern the underlying sensory threshold of a participant in a behavioural task, known underlying thresholds and measured thresholds can be compared via computational methods. A ‘true’, underlying threshold can be modelled and staircase procedures can be evaluated in terms of accuracy (the magnitude of difference between the threshold estimate and true threshold) across a sample of simulated participants.

This experiment aimed to investigate the most appropriate staircase procedure to use with preschool children by examining the differences in threshold estimates across different lengths of procedures. The transformed 71%, weighted 75% and transformed 79% staircases were simulated for procedures with numbers of trials ranging from 10 to 100. Simulations were conducted to reflect child-like and adult-like levels of attentiveness. The success of the simulated threshold estimates was evaluated in terms of their accuracy of representing the true threshold, their variability across the sample and the percentage of simulated participants that failed to meet the criterion of convergence on a valid threshold estimate. Consistent with previous findings for adults and adult-like simulations (Kaernbach, 1991; Rammsayer, 1992), we hypothesized that the weighted 75% staircase procedure would be

more successful in estimating the true threshold for shorter procedures than the transformed 71% and 79% procedures for child-like simulations.

Method

Modelling Parameters

A total of sixty thousand staircase runs (equivalent to participants) were simulated for each of the three staircase procedures (transformed 71%, weighted 75% and transformed 79%). Ten procedure lengths were tested, ranging from 10 to 100 trials in linear steps. The variability in attention of a group of preschool children was simulated by counterbalancing equally across six levels of attention: fully attentive and 10%, 20%, 30%, 40% and 50% inattentive (child-like sample). For the fully attentive level, individual responses were generated independently, based on a model psychometric function with a cumulative-normal shape between asymptotes of chance performance (50% or 33.3% depending on the response paradigm) and perfect performance (100%). The five levels of inattention were simulated by designating each of the trial responses a 10%, 20%, 30%, 40% or 50% probability of being random, independently. Five thousand staircases were simulated for each level of attention at each trial number for each staircase procedure. Thirty-thousand staircases at the fully attentive level were simulated to act as a control (adult-like sample).

This design was modelled according to two response paradigms, a 2 alternative forced choice (2AFC) and a 3 alternative forced choice (3AFC) version, which were collapsed together for analysis in a counterbalancing approach. Each staircase began with an initial phase which consisted of a step size of 20%¹. The step size was then reduced to 10% after the first response error occurred.

Simulated thresholds were estimated as the average of the maximum even number of reversals. On this basis, thresholds were only estimated for simulated staircase runs which resulted in at least 2 reversals. ‘True’ thresholds were computed assuming normal psychometric function gradients and upper asymptotes of unity (i.e. equal to 100% correct) for the targeted percentage of correct performance of each procedure (71%, 75% and 79%).

A maximum stimulus value was not applied to the staircase simulations. However, after the simulations were completed, but before thresholds were estimated, any staircase stimulus values that were greater than the initial stimulus level were Winsorized to the initial stimulus level value.

Analysis of Simulated Data

Since the simulated sample size was extremely large, all comparisons were significantly different and therefore inferential statistics are not reported. The staircase procedures were evaluated at each trial number for both child-like and adult-like simulations in terms of the success of the estimation procedure quantified by: (1) the accuracy of the threshold estimates. The accuracy of each staircase procedure was computed as the absolute difference between the simulated threshold estimate and the true threshold estimate; (2) the variability of the threshold estimates across the simulated sample (measured as the standard deviation, SD); and (3) threshold validity, the percentage of staircases that failed to meet the threshold estimation criterion (i.e. failed to make at least two reversals).

A sum of the three parameters described above was computed for the three staircase procedures at each trial number. In computing the summed score for threshold success, the values of threshold accuracy, variability, and validity were normalized to be on the same scale (0 to 100). Staircases were ranked as more reliable when the sum of the parameters was

lower. Lower values were attained when there was a lower difference between the obtained threshold estimate and the true threshold, less variability among thresholds in the sample, and/or fewer staircases which failed to converge on a threshold. The staircase procedure which was evaluated to have the lowest sum of the three parameters for the smallest trial numbers with the child-like simulations was considered to be the most appropriate staircase to select for use with preschoolers.

Results

Threshold Accuracy

For child-like simulated participants, the threshold accuracy improved using all procedures when the length was increased from 10 to 20 trials. However, for the adult-like participants, threshold accuracy decreased when the length was increased from 10 to 20 trials. When attentiveness was variable across simulated participants (child-like sample) threshold estimates were closest to the ‘true’ threshold (most accurate) for very short procedures (10 trials) using the weighted 75% procedure (see Figure 1a). However, for longer procedures (20 to 50 trials) better estimates were attained with the transformed 79% procedure, and for the longest procedures (60 to 100 trials) the transformed 71% resulted in the most accurate threshold estimates. This was in almost complete contrast to the adult-like sample where the best estimates were consistently generated using the weighted 75% procedure for procedures greater than 10 trials. The smallest procedure with only 10 trials was best estimated by the transformed 79% procedure. Overall, the threshold estimates of the child-like simulations were less accurate compared to the adult-like simulations across procedures and trials (mean

child-like = 1.2 , SD = 0.1; mean adult-like = 0.4, SD = 0.1). Importantly, as the length of the procedure increased beyond 20 trials, the child-like thresholds became less accurate while the thresholds of the adult-like simulations became more accurate. This different pattern of accuracy for the child-like and adult-like simulations was present for all three procedures.

Threshold Variability

The results of the variability of thresholds were very similar for both the child-like and adult-like samples (see Figure 1b). The weighted 75% staircase resulted in the lowest levels of variability for all procedures greater than 10 trials. For procedures consisting of 10 trials, the transformed 79% staircase resulted in the least variability in threshold estimates for both child-like and adult-like samples.

Threshold Validity (measured as the percentage of invalid staircase simulations)

Regardless of the number of trials, none of the procedures resulted in 100% valid threshold estimates for the child-like sample (see Figure 1c). For procedures with less than 30 trials, the weighted 75% led to the fewest number of simulations failing to obtain a valid threshold estimate (10 trials = 20.5% failing and 20 trials = 7.8% failing). For procedures longer than 30 trials, the transformed 71% had the fewest simulated participants failing to reach valid threshold estimates (4.37% - 1.936% failing across trial lengths 30 to 100; 95% CIs = 4.3677-4.3723% and 1.9352- 1.9382%, respectively).

For the adult-like sample, all procedures resulted in 100% valid thresholds when 60 or more trials were presented. The weighted 75% procedure led to the fewest number of invalid threshold estimates for all trial lengths less than 60 trials (10 trials = 10% failing, 20 trials =

0.06% failing and all remaining trial lengths = 0% failing). However, for both transformed procedures the adult-like simulations required at least 60 trials for 100% valid thresholds.

Overall Success (Summed Accuracy, Variability, and Validity Scores)

For the child-like simulations, the weighted 75% procedure was more successful than the other procedures for lengths of 10 to 40 trials (mean scores = 20.42 to 13.07, at 10 to 40 trials; see Figure 1d). Beyond 40 trials, the transformed 71% was the most successful (mean scores = 12.25 to 10.72, at 50 to 100 trials) and the weighted 75% was the second most successful (mean scores = 12.42 to 11.48, at 50 to 100 trials). The transformed 79% was consistently the least successful of the procedures across all trial lengths.

For the adult-like sample, the weighted 75% was the most successful procedure, having the lowest summed scores for all the trial lengths (mean scores = 12.11 to 2.46, at 10 to 100 trials). For procedures consisting of 10 to 50 trials, the transformed 71% procedure was the second most successful (mean scores = 19.62 to 7.45, at 10 to 50 trials), and the transformed 79% was least successful (mean scores = 25.96 to 7.49). However, for procedures consisting of more than 50 trials, the transformed 79% was second most successful (mean scores = 6.14 to 3.57, at 60 to 100 trials) and the transformed 71% was least successful (mean scores = 6.56 to 4.30).

Figure 1 about here

Discussion

The objective of this experiment was to investigate the staircase procedure that led to the most successful estimates of thresholds with relatively small numbers of trials (less than

30). Previous studies using adults, or simulations of adult-like participants, have identified the weighted 75% procedure as more appropriate than transformed procedures for estimating thresholds with small trial numbers (Kaernbach, 1991; Rammsayer, 1992). Here we found that the weighted 75% procedure was the most successful at estimating thresholds with short procedures (less than 30 trials) of child-like performance (i.e. where attentiveness was variable). In contrast, the transformed 79% was the least reliable procedure for all trial numbers for the child-like sample. All three procedures tested were equally successful in estimating thresholds for adult-like simulations when trial numbers were greater than 30.

In previous studies the transformed 79% procedure has been employed more frequently to estimate thresholds with children than the weighted 75% procedure. However, our results suggest that the transformed 79% procedure is likely to be the least successful procedure for estimating the thresholds of children. The most successful approach from our results, the weighted 75%, has been used with children as young as 7-years-old (e.g. Rickard, Heidtke, & O'Beirne, 2013), whereas younger children, who are likely to be most variable in their attentiveness, have been tested only with transformed staircase procedures (e.g. from 4-years-old; Law, Vandermosten, Ghesquiere, & Wouters, 2017; Banai & Yifat, 2011; Wang, Chen, Chiang, Lai, & Tsao, 2016). Here we find that lengthy procedures, at least in terms of the number of trials, result in less accurate estimates, and fail to increase the overall success of threshold estimates, for simulations with variable attentiveness. Variable attentiveness seems to be the direct cause of the decline in threshold estimation success, since adult-like (fully attentive) simulations result in more accurate threshold estimates for longer procedures. This finding is consistent with the arguments that lengthy procedures are more difficult for young children based on their underdeveloped ability to sustain attention (Klenberg, Korkman, & Lahti-Nuuttila, 2001; Mahone & Schneider, 2012). Most notably, in the current

study, attentiveness seems to have a far greater impact on threshold accuracy than it does on variability or validity.

Particularly for the child-like simulations, procedures consisting of only 10 trials resulted in much less successful threshold estimates than procedures containing 20 or more trials. For young child participants, procedures with only 10 trials are not likely to be long enough to locate their threshold region. If participants respond erroneously early in the procedure, it is highly likely that their staircase will fluctuate around values which are far from their threshold and the few remaining trials will not be enough to rectify this. The relatively high proportion of simulated participants who fail to attain a valid threshold (i.e. two or more reversals) during the procedure supports this proposal.

In this study, three outcomes of a staircase procedure (threshold accuracy, threshold variability and threshold validity) were combined into a measure of threshold estimation success. Each of these component scores can be considered important measures of the success of a procedure, but overall success is dependent on all three. The aim of any procedure is to provide as accurate a threshold estimate (accuracy) as possible, however if this can only be accomplished for a small proportion of the population tested (validity) then it will not be appropriate for widespread usage. Similarly, if the accuracy of a threshold estimate seems accurate for a group (the mean), but varies greatly across the group (variability) then again, usage of the procedure for a large group of participants is questionable. In simulations, large variability in the threshold estimate is also suggestive that re-test reliability (a highly valued characteristic of threshold estimation procedures) is likely to be poor.

The three component scores changed as the procedures increased in their number of trials, especially beyond 20 trials. While both threshold variability and validity improved with

larger trial numbers, threshold accuracy worsened for the child-like sample only. This may well have been an artefact of the initial step size for the procedure. Starting with a large step size and halving the step size after the first error response is likely to bias the staircase towards overshooting the threshold. With the added impact of inattention (random responses) in the child-like simulations this may have resulted in a maintained overshoot of the threshold with more trials.

Based on the results of the child-like simulations, very similar levels of success in estimating thresholds can be achieved with 20 trials as can with 100 trials. Although not comparable with the success of estimating adult-like thresholds at the same number of trials, the weighted 75% is more robust to variations in inattentiveness with fewer trials than the transformed procedures tested. Since a primary concern in designing adaptive auditory discrimination tasks for use with young children is to minimise the duration, the weighted 75% procedure seems like an ideal candidate for use with preschoolers.

Experiment 2: Selecting a step size based on successful threshold estimation with a weighted 75% procedure

Selecting the appropriate step size for a staircase procedure greatly impacts the accuracy and speed at which a threshold can be determined (Cornsweet, 1962; García-Pérez, 1998). For many procedures, large step sizes will increase the risk of overshooting the threshold region. In contrast, small step sizes reduce the risk of overshooting the threshold region, but they do so at the cost of increasing the number of trials required to reach the region in the first place. Smaller step sizes may also result in longer periods of unreliable performance whereby the stimulus differences change only very subtly for a prolonged number of trials, particularly if the participant is very close to their threshold. This may be a

frustrating process particularly for young children who are used to highly positive reward schedules. Generally speaking, positive responses on child focused tasks tend to be followed by positive feedback. For young children in particular, positive feedback often enhances their intrinsic motivation (Henderlong & Lepper, 2002), which subsequently may lead to them engaging longer with a task. Given also that preschool children have substantially shorter attention spans than older children and adults, they may then fail to complete a task if it is too long in duration. Overall, the increased time demands of smaller step sizes may be potentially counterproductive to the accurate estimation of a threshold.

Staircase procedures generally begin with stimulus differences that are much greater than the expected threshold of an individual. Stimuli at these levels are referred to as suprathreshold. Under these circumstances larger step sizes are considered useful at the beginning of procedures as they have the potential to locate the threshold region quickly (Cornsweet, 1962). Introducing a threshold estimation procedure with suprathreshold trials may also be helpful for younger children in order to initially engage their attention and reward them with positive feedback for essentially ‘easy’ trials. The onset of erroneous responses in an attentive participant is then likely to signify that they are approaching their threshold. At this point, the step size is often reduced to limit the risk of overshooting the threshold region and increase the precision of the threshold estimate. In comparison, thresholds estimated using larger step sizes typically are less precise with increased variability (Schlauch & Rose, 1990). There are no guidelines on how to select an appropriate step size for a particular staircase procedure. However, it is common for staircases used with both children and adults to increment (multiply) or decrement (divide) the starting discrimination by a relatively small step size (e.g. $\sqrt{2}$) according to the step rule (Amitay, Irwin, Hawkey, Cowan, & Moore, 2006; Banai & Yuval-Weiss, 2013; Moore et al., 2008; Zhang et al., 2016).

Furthermore, the step size of a staircase may also have a large impact on the degree to which automatic perceptual mechanisms like sensory adaptation are available to influence task performance. For example, smaller step sizes may make sensory adaptation from trial-to-trial more likely, which may then lead to more accurate measures of auditory discrimination capacity.

Step size is such an influencing factor in the accuracy of threshold estimates, but an appropriate step size is not determinable without already knowing an individual threshold. However, the ideal step size can be determined by proxy measures of threshold estimation success in a simulation. Here we determine the ideal step size is one which corresponds to the best summed score of the three aforementioned metrics of threshold estimation success: threshold accuracy, threshold variability, and threshold validity.

This experiment aimed to investigate the most appropriate step size to use in the weighted 75% staircase procedure with preschool children. To accomplish this, child-like and adult-like responses were simulated using the weighted 75% procedure with step sizes ranging from 1%-60% of the value of initial stimulus difference presented. The most successful step size was determined based on the same summed score of accuracy, variability and validity computed in experiment 1. Consistent with previous findings for adult-like simulations (Schlauch & Rose, 1990), we hypothesized smaller step sizes would lead to the most successful threshold estimates in child-like and adult-like simulated participants under the weighted 75% staircase.

Method

Modelling Parameters

The staircase simulation procedure was similar to that used in experiment 1. However, this time, simulations were conducted only under the weighted 75% procedure and consisted of 20 trials per simulated staircase. Sixty step sizes were tested which corresponded to linear steps of the initial stimulus difference presented (from 1% to 60%). All other aspects of the modelling parameters for this study were identical to those described in experiment 1.

Analysis of Simulated Data

The analysis was identical to that in experiment 1. Measures of threshold accuracy, threshold variability, and threshold validity were evaluated at each of the 60 step sizes.

Results

Threshold Accuracy

Threshold accuracy was initially very poor at the smallest step sizes that were tested for both the child-like and adult-like samples (see Figure 2a). Thresholds were most accurate for the child-like sample when the step size was 20% of the initial stimulus difference (Mean = 0.05, 95% CI = 0.004, 0.097). For the adult-like sample, thresholds were most accurate when the step size was 5% of the initial stimulus difference (Mean = 0.04, 95% CI = 0.024, 0.064). Threshold accuracy became poorer as step sizes increased beyond 20% and 5% for child-like and adult-like samples, respectively. On average, across all the step sizes, the child-like threshold estimates were slightly more accurate than those of the adult-like sample (mean child-like = 1.97, SD = 1.37; mean adult-like = 2.02, SD = 1.28).

Threshold Variability

For both the adult- and child-like samples, threshold variability increased gradually as the step size increased (see Figure 2b). Smaller step sizes resulted in lower levels of threshold variability across the sample of simulated participants. The overall influence of step size on threshold variability was very similar for both the child-like (Mean = 5.5, SD: 2.4) and adult-like (Mean = 5.2, SD: 2.4) samples.

Threshold Validity (measured as the percentage of invalid staircase simulations)

No step size resulted in 100% valid thresholds for the child-like or adult-like simulations (child-like maximum valid thresholds = 94.9%, adult-like maximum valid thresholds = 99.99%; see Figure 2c). More child-like simulations failed to attain a valid threshold compared to the adult-like simulations across all of the step sizes (Child-like mean = 8%, 95% CI = 7.22%, 8.77%; adult-like mean = 1.67%, 95% CI = -0.9%, 4.24%). Step sizes that were 1% or 2% of the initial stimulus led to greater than 10% of child-like simulations failing to attain a valid threshold (28.3% and 14% failed, respectively). Step sizes that were 3% and larger led to similar numbers of failures to attain a valid threshold (7.4% and 9.4% failed for step sizes 3 and 60% of the initial stimulus difference, respectively). Similarly, for the adult-like sample, step sizes which were 1% or 2% of the initial stimulus led to greater than 10% of failures to attain a valid threshold (74.7% and 20.3% of failures, respectively), Whereas 3% and larger step sizes led to small numbers of failures (2.4% and 0.02% failed for step sizes 3% and 60% of the initial difference, respectively).

Overall Success (Summed Accuracy, Variability, Validity Scores)

For both child- and adult-like samples, threshold success improved across very small step sizes (i.e. up to 20%) and worsened for very large step sizes (i.e. beyond 40%; see Figure 2d). For the child-like sample, threshold success improved when step sizes increased from 1% to 14% (mean scores = 30.02 to 15.18). For step sizes greater than 14%, threshold success worsened (mean scores = 15.23 to 53.89). However, for step sizes that were 8% to 20% success scores varied very little (by less than 1 out of 100, mean scores = 15.99 to 15.76 at 8% to 20%).

For the adult-like sample, threshold success improved for step sizes 1% to 6% (mean scores = 39.87 to 6.58), whilst larger step sizes from 7% to 60% had poorer threshold success (mean scores = 7.38 to 47.95). Threshold success varied very little (less than 1 out of 100) for step sizes from 4% to 6% (mean scores = 6.93 to 6.57, at 4% to 6%).

Figure 2 about here

Discussion

The aim of this experiment was to investigate which step size led to the most successful threshold estimates in child-like simulations under the weighted 75% staircase procedure. Previous studies have suggested that staircase procedures consisting of smaller step sizes lead to more reliable threshold estimates in adults, and in simulations of adult-like participants (Saber & Green, 1996). Although these previous studies were conducted with the transformed rather than the weighted staircase procedure, the weighted procedure has been suggested to be more robust to differences in step size (Kaernbach, 1991). However, there have been no previous studies that have investigated how step size influences the

successfulness of threshold estimates when used specifically with children, or child-like participants, under the weighted procedure. The hypothesis for this experiment was that, for both child- and adult-like samples, procedures with small step sizes would lead to more successful threshold estimates compared to procedures with large step sizes. The findings confirmed this hypothesis showing that thresholds were generally more successful whenever step sizes were smaller rather than larger for both the simulated samples. Smaller step sizes as outlined by previous studies, allow the threshold to be targeted quite precisely during the procedure. The results of this experiment showed this to be the case even when there was variation in the attentiveness of the simulated participants.

The findings of this experiment also showed that extremely small step sizes (1%–2% of the initial stimulus difference) led to threshold estimates with very poor levels of success. Similar to the findings in experiment 1 with procedures consisting of extremely low numbers of trials (10 trials), step sizes of 1%–2% of the initial stimulus difference are likely to be too small to reach the true threshold within the duration of the procedure. This may especially be the case for the child-like sample who are more likely than the adult sample to respond erroneously very early on in the procedure. While it is tempting to assume that these step size values will not be large enough to lead to a series of reversals and calculate a valid threshold, the high number of valid thresholds (at least 2 reversals) attained for these very small step sizes by both the child-like and adult-like simulations, suggest this is not the case.

In both simulated samples, as step sizes increased beyond approximately 20% of the initial stimulus, thresholds became less successful. These findings replicate previous studies that have found that larger step sizes lead to thresholds with poorer reliability (Saber & Green, 1996; Schlauch & Rose, 1990), and this was found to be the case particularly with the child-like sample. For the child-like sample who are likely to make more erroneous responses than the adult-like sample, larger step sizes are likely to result in them ending up further away

from their true threshold. The child-like sample may then not proceed to make enough correct responses, due to variation in attentiveness, to enable them to approach the threshold region. The adult-like sample will likely respond attentively more often than the child-like sample and complete trials with stimulus values which gradually enclose their threshold, but larger step sizes will still result in increased variability of reversals around the threshold.

As for experiment 1, we considered the three component scores as well as the overall summed score in order to evaluate threshold success. Whilst threshold variability worsened as step sizes increased and threshold validity improved with increasing step size (besides the identified initial, extremely small step sizes), threshold accuracy followed quite a different trend. For both samples, threshold accuracy initially was quite poor with small step sizes and gradually improved up until the optimum step sizes identified above (threshold accuracy: 20% for child-like sample and 5% for adult-like sample). For step sizes larger than these optimum points, the accuracy of thresholds began to gradually decrease again. This trend was more pronounced in the child-like sample than in the adult-like sample. Step sizes which were too small led to participants not reaching their true threshold within the duration of the procedure. Due to the increased likelihood for the child-like simulations to respond erroneously, the staircase was more likely to make more steps up and subsequently move further away from the true threshold. However, very large step sizes mean that the values of the stimulus difference on consecutive trials in the procedure will vary quite significantly, and potentially miss the threshold. Thus the accuracy of threshold estimates will not be very precise.

Based on the results of the child-like and adult-like simulations in this experiment, it was found that when smaller step sizes are used in a staircase procedure, there are good levels of threshold success. Smaller step sizes were most robust to variations in attentiveness. It seems that the weighted 75% procedure lasting 20 trials and using a relatively small step size

would be appropriate to use with young children who vary in attentiveness like the child-like simulations used in this study.

Experiment 3: Measuring the impact of age, response paradigm and level of task engagement on the reliability of threshold estimates

Experiments 1 and 2 used a simple model of child-like performance by simulating variations of attentiveness during an auditory discrimination task. This modelling approach identified the most appropriate staircase procedure, number of trials, and step size to present to young children. However, this model assumed young children to be variable only in their attentiveness, not their other cognitive capacities like working memory, or their sensitivity to task demands. While differences in attentiveness were modelled as random episodes of distraction, differences in working memory are likely to impact each trial in a less binary manner. It is therefore less clear precisely how working memory differences should be modelled to emulate the reduced working memory capacities of younger children compared with older children and adults. Instead, measurements of responses from real participants may be necessary to identify procedures most robust to the working memory resources of typical preschool aged children.

Auditory discrimination tasks can be presented in a range of paradigms differing in their working memory demands, as well as how comprehensible and engaging they are for young children to sustain their attention. The response paradigm itself is typically changed in two distinct ways to modify the level of sustained attention and working memory required: (1) the number of sounds (intervals) presented are adjusted; and (2) the number of alternatives that could be the correct response are adjusted. Previous studies with children between the ages of 3–4 years old have typically involved only two or three intervals and two

or three response alternatives. Two interval, two alternative forced choice paradigms (2I-2AFC) have been most widely used with children of preschool age (3-4 years old; Anvari, Trainor, Woodside, & Levy, 2002; Banai & Yifat, 2011; Corriveau et al., 2010). In these paradigms two sounds are presented and children are typically instructed to say whether the sounds are the same or different (Anvari et al., 2002; Banai & Yifat, 2011); or they may be asked to identify the sound which is characterised by the label presented by the experimenter, for example, 'which sound was the louder tone?'. Although it is tempting to assume that 2-interval paradigms require less cognitive resources than 3-interval tasks, purely based on the duration of a trial and thus the potential for reduced demands on working memory and sustained attention, these tasks can actually increase the demands over three interval two alternative forced choice paradigms (3I-2AFC), which do not require the comprehension of verbal labels for a specific stimulus parameter (e.g. loudness; Sutcliffe & Bishop, 2005). Instead, 3I-2AFC paradigms are typically presented as odd-one-out tasks where participants are required to identify which tone differs from two other tones that are identical. Finding labels for the acoustic properties of the stimulus difference that are easy for young children to comprehend can be quite difficult. In fact it is during the preschooler years between 3 and 5 years old that children start frequently and appropriately producing the types of comparative adjectives typically used as labels in two interval paradigms (e.g. 'louder', 'quieter'; Gathercole, 1985). Two interval paradigms may also be susceptible to verbal response biases by the participants, for example the tendency to report 'same' rather than 'different', or 'louder' rather than 'quieter', similar to the bias to report 'yes' rather than 'no' that is often observed among preschoolers (2-5 year olds; Fritzley & Lee, 2003). This is not to discount the potential presence of biases in three interval tasks as well. In fact, bias to detect changes more on the first (primacy) or last (recency) sounds in the sequence, or to select one

particular interval over a range of alternative choices are equally likely on two and three interval paradigms.

Although used less frequently with children than 2I-2AFC paradigms, 3I-3AFC paradigms have been used with children as young as 3 years of age (Wightman et al., 1989). However, there are much fewer preschool studies consisting of 3I-2AFC and 3I-3AFC tasks compared to 2I-2AFC tasks. Some studies have shown that threshold estimates are significantly lower and less variable when adults and children over 6 years old are presented with more than two auditory intervals in forced choice tasks (Schlauch & Rose, 1990; Sutcliffe & Bishop, 2005).

The ability to sustain attention on a task is not driven purely by the attentional capacities of the individual being tested, but is also likely influenced by the attention-grabbing properties of the task being presented. Visually presented cartoons linked to the sounds being presented have typically been used to draw attention to position of the stimulus presentation in time (giving them a spatial referent for response) and to provide rewarding feedback that is aimed at enhancing the motivation of children to respond as accurately as they can. In addition, it has been reported across a number of previous threshold estimation tasks with preschoolers that the children do not actually interact with the task themselves by making their own responses. Instead, they are required to point at the interval that they wish to select for their response and the experimenter then performs the response via a computer mouse, or other device (Banai & Yifat, 2011; Bavin, Grayden, Scotti, & Stefanakis, 2010; Corriveau et al., 2010; Creel, 2015; Thompson, Cranford, & Hoyer, 1999). By having this barrier between participants actually being able to interact with the task themselves, as well as preschoolers' restricted attentional spans, this may increase the likelihood of them losing interest and disengaging from the task.

This experiment aimed to establish the impact of response paradigm (3I-2AFC, 3I-3AFC), level of task engagement (2D touch screen selection and 2D computerised feedback, touch 3D figurines and 3D object feedback) and age (3 year olds, 4 year olds, adults), on the accuracy of threshold estimates from the weighted 1-up 1-down procedure, starting with a relatively large difference between the sounds and with a relatively small step-size, as established from experiments 1 and 2. Unlike the simulation experiments (1 and 2), the ‘true’ threshold for the real participants were unknown. So instead of computing the difference between the true and estimated thresholds as a measure of accuracy, the accuracy of the threshold estimate was assessed in terms of its test re-test reliability across different test sessions. If the task is accurately estimating the threshold, then the threshold should be very similar across test sessions. Variability was measured in terms of the difference in individual threshold estimates across test runs within and between sessions. Validity was measured in terms of the proportion of participants reaching the targeted correct performance or above, in contrast to the simulation studies (experiments 1 and 2) where validity was measured as those achieving at least two reversals. The success of the threshold estimates was estimated as the sum of the normalised (0 to 100) accuracy in terms of reliability (1-reliability coefficient), variability (SD across all runs), and validity (percent of participants performing at less than 75% correct). We hypothesised that: (1) response paradigms that have been argued to place less demands on working memory capacity (i.e. 2AFC compared with 3AFC) would result in lower threshold estimates and improved measures of threshold accuracy, variability and validity for preschoolers, but not adults; (2) increased levels of physical engagement (3D figurines compared with a 2D touchscreen) would result in improved measures of threshold accuracy, variability and validity for preschoolers, but not adults; (3) the most successful threshold estimates will be attained with the preschoolers for the 3I-2AFC paradigm and the highest level of physical engagement (3D figurines).

Method

Participants

Sixty-six 3-4 year old participants were recruited for this study. Nineteen participants were either absent or did not wish to participate for one or more of the test sessions. Complete data used in the analyses was thus available for twenty-four 3 year olds (3 years 1 month – 3 years 10 months; mean age = 3 years 5 months, SD = 2.5 months, 11 females), twenty-three 4 year olds (4 years 2 months – 4 years 10 months; mean age = 4 years 6 months, SD = 3.5 months, 10 females) and twenty adults (20 years 10 months – 35 years 3 months; mean age = 25 years 2 months, SD = 3 years 7 months, 14 females). Child participants were recruited from several preschool, day-care and playgroup settings in Northern Ireland and adult participants were recruited from the student population at Queen's University Belfast. Informed written consent was provided by all adult participants and parents or guardians of the child participants. Ethical approval for the study was granted by the School of Psychology Research Ethics Committee, Queen's University Belfast.

Design and Procedure

All participants completed a brief hearing screen, a single session of discrimination training (visual) and the same frequency discrimination task four times over the course of a three-week period. The frequency discrimination task was presented in four separate sessions. All testing sessions were conducted in a quiet room away from distractions by the same experimenter. Equal numbers of children in each age group (3 years old and 4 years old) were initially randomly assigned to the 2AFC or 3AFC response paradigm. However, several drop-outs (n=19 preschoolers) due to preschool absence resulting in failure to complete one or

more of the measures, resulted in unequal numbers of children in each group. After removing the 19 participants who failed to complete all tasks this resulted in eleven 3 year-olds and thirteen 4 year olds completing the 2AFC format, and thirteen 3 year-olds and ten 4 year olds completing the 3AFC format.

After the visual discrimination task, participants were explained the instructions of the frequency discrimination task. Participants completed four practice trials at the beginning of the procedure. These trials consisted of the maximum frequency difference between the stimulus intervals. Regardless of accuracy on the practice trials, participants immediately entered into the experimental trials afterwards. Preschool participants completed each presentation condition in two separate sessions, within the same week, then again two weeks later in order to measure test-retest reliability. Adult participants completed each condition in the one session, then again two weeks later. The orders in which the presentation conditions were completed was counter-balanced across participants.

Hearing Screen

All participants completed a brief hearing screen consisting of the detection of pure tones at octave frequencies from 250-8000 Hz in both ears presented using a clinically calibrated manual audiometer. Participants had to identify when they detected a sound in either their left or right ear. In the case of adult participants, this was accomplished by the raising of a hand immediately upon the detection of a sound. To aid engagement, child participants indicated their detection of a sound by placing a building block in a box as quickly as possible on hearing the sound. All participants attained detection thresholds of ≤ 30 dB HL bilaterally.

Discrimination Training

Prior to administering the first frequency discrimination task all children completed a brief discrimination training task on a single occasion. This task presented children with geometric shapes on a touchscreen computer to familiarise children with the concepts of odd-one-out tasks. The visual discrimination task was presented to participants on the touch screen laptop in the response paradigm that they had been assigned to (3I-2AFC or 3I-3AFC). The task consisted of three blocks with four trials in each. On each block geometric shapes were presented that differed by a single dimension; colour, shape or size. Participants received visual feedback in the form of happy or sad faces for correct and incorrect responses, respectively.

Frequency Discrimination Task

The frequency discrimination task was presented as a game with three pig characters. For the 3I-2AFC paradigm, the game consisted of a mummy pig ('X') and her two identical baby pigs ('A', 'B'). Mummy pig acted as a reference tone and produced the same sound throughout the task (500 Hz). The two baby pigs deviated in frequency to mummy pig for half of the trials each, starting at a difference of 1000 Hz. Participants were required to identify the baby pig that produced the deviant sound to mummy pig. The 3I-3AFC paradigm task consisted of three identical baby pigs. One of the baby pigs produced a sound that was deviant in frequency from the other two, again starting with a 1000 Hz difference. Each pig made the deviant sound on a third of the trials. Participants were required to identify the baby pig that made the deviant sound from the others ('odd-one-out').

Participants completed the frequency discrimination task in two presentation conditions (see Figure 3). Firstly, on a touch-screen laptop as the pigs produced their sounds, the saturation of the images altered slightly to give the effect that the pigs were lighting up. All participants identified the baby pig they perceived to make the deviant sound by pressing

its image on the laptop touch screen. Following correct responses, a brown box, which was displayed at the top of the screen, disappeared to reveal a number of vegetables behind it. Participants could select one vegetable, using the touch-screen, which then appeared at the bottom of the screen in the pigs' feeding trough. The trough and its contents remained on screen for the duration of the task, representing the number of correct responses that had been made. Following incorrect responses, the brown box remained on screen and a cursor appeared to prompt participants to proceed to the next trial. Children were encouraged to find the "hungry baby pig" who made the different sound, and to try to get this pig as much food as possible. In addition, the pigs all had red woolly hats that remained on for as long as participants made correct responses. If the participant made an incorrect response, the hats disappeared from the screen. Child participants were encouraged to try to get the pigs their hats back by correctly identifying the hungry baby pig that made the different noise. Participants were told the pigs were cold without their hats. This aspect of the game was emphasized to encourage accurate responses providing a penalty when incorrect responses were made. Secondly, the participants also completed the task with toy pig figurines. The figurines were silicon, hollow pig toys (mummy pig: 6cm width, 7cm height; baby pigs: 3cm width, 4.5cm height) and were attached to a block base (26cm width, 7.5 cm height) that was connected to a laptop. Underneath each of the pig toys were individual touch-sensitive pads, so that when slight pressure was applied by the participant to one of the toys, the pad would feedback to the software on the laptop which pig had been pressed. Each pig also lit up via an LED light attached to the base whenever they produced their sound and also when the participant pressed one of the baby pigs as their response. The pig images in the laptop condition were drawn to resemble the pig toys and the same farmyard backdrop scene from the laptop condition was provided as the backdrop for the figurines. Toy vegetables and a basket to resemble the feeding trough were used in the figurines condition. The basket

remained visible to participants throughout the task and following correct responses the experimenter would lift the lid from another box containing the toy vegetables and present them to participants to select one piece to put into the basket. Red knitted hats were also used in the figurines condition and the experimenter manually removed and replaced them following incorrect and correct responses respectively.

Figure 3 about here

Auditory Stimuli and Staircase Procedure

Stimuli were pure tones lasting 200 ms (10 ms rise and fall times), with 500 ms inter-stimulus intervals and were played at 75 dB SPL through binaural headphones. The minimum frequency, and the frequency that deviant tones were compared against was 500 Hz. Deviant stimuli began at a difference of 1000 Hz. This difference was halved following every correct response until the participant made their first error, after which the procedure converted to the weighted up-down staircase targeting the 75% correct performance level on the psychometric function (Kaernbach, 1991). After every correct response the staircase would step down and the frequency difference would decrease by one step ($\sqrt{2}$ of the previous trial frequency), and after every incorrect response the staircase would step up and the frequency difference would increase by three steps ($[\sqrt{2}]^3$ of the previous trial frequency). Participants completed a total of twenty experimental trials. Thresholds were estimated as the average of the largest even number of reversals in the procedure. Participants completed two runs of the task at each time point and the best (lowest) estimate of each time point was selected for statistical analysis.

Results

Participant performance was analysed in terms of differences between age groups, presentation format and response paradigm. Differences were analysed in terms of: (1) threshold estimates; (2) test re-test reliability of thresholds within an age group; (3) variability of individual threshold estimates across runs; and (4) the proportion of participants reaching the targeted proportion of correct performance in each age group. All these measures were computed across all runs that a participant completed. Some child participants were sufficiently fatigued by the task that they did not assent to participate in the second run of the procedure in the same session. This amounted to 3 or 10 runs for the 3 year olds, for the first and second time points, respectively, and 1 run for the 4 year old in the second time point only. Adult participants completed all runs at all time points. In order to prevent biasing the computation of variance across the 4 runs of each presentation type and response paradigm, missing values were replaced with the mean threshold of the participants of the same age and response paradigm that did complete the second run. In contrast, the mean threshold estimates and test re-test reliability coefficients were computed based solely on the runs completed by the participants.

Thresholds

The mean estimated threshold across runs was submitted to a 2x3x2 mixed ANOVA: level of physical engagement (touchscreen, figurines) within x age (3 years old, 4 years old, adult) between x response paradigm (2AFC, 3AFC) between. Significant main effects of level of physical engagement ($F[1,61] = 16.87, p < 0.001, \eta_p^2 = 0.22$; see Figure 4a) and age ($F[2, 61] = 99.58, p < 0.001, \eta_p^2 = 0.77$) were found. Significantly lower thresholds were found for higher levels of physical engagement (figurines) across response paradigm and age (main effect of level of physical engagement). A significant linear trend in lower thresholds with

increasing age was all found across response paradigm and level of physical engagement ($F[2, 61] = 99.58, p < 0.001, \eta_p^2 = 0.77$). Adults obtained lower thresholds than 3 and 4 year olds ($t[23.30] = -25.37, p < 0.001$ and $t[22.08] = -9.51, p < 0.001$ corrected for unequal variances, respectively) and 4 years olds obtained lower thresholds than 3 year olds ($t[33.45] = -3.45, p < 0.01$ corrected for unequal variances). A significant interaction of level of physical engagement by age was found ($F[2, 61] = 4.54, p < 0.05, \eta_p^2 = 0.13$), where lower thresholds were obtained for higher levels of physical engagement (figurines) for the both 3 and 4 year olds ($t[23] = 2.37, p < 0.05$ and $t[22] = 3.94, p < 0.01$, respectively), but thresholds were unaffected by level of physical engagement in adults ($t[19] = 0.24, p > 0.1$). All other main effects and interactions failed to reach significance (all $p > 0.1$).

Figure 4 about here

Threshold Accuracy (Test re-test reliability)

Reliabilities of the threshold estimates were computed for each age group intraclass correlations (ICC) between first and second sessions averaged across runs in the same session. ICCs were not computed for the adult results because the sample of thresholds showed high positive kurtosis (leptokurtotic distribution), and thus not meeting the assumptions for correlation. Reliabilities overall did not differ substantially by age for the preschool participants (see Table 1). However, reliabilities were higher for the greater level of physical engagement condition (3D figurines compared with 2D touchscreen) for both age groups of preschoolers. Furthermore, a substantial difference in 95% confidence intervals

between the reliabilities of the 2D touchscreen condition (0.05 – 0.82) and the 3D figurines condition (0.38 – 0.88) was restricted to the 3 year old group.

Table 1 about here

Threshold Variability

Variability scores of the threshold estimates were computed for each participant and condition as the standard deviation of the four runs of each presentation type. Variability scores were then submitted to a 2x3x2 mixed ANOVA: level of physical engagement (touchscreen, figurines) within x age (3 years old, 4 years old, adult) between x response paradigm (2AFC, 3AFC) between. There was a significant main effect of age ($F[2, 61] = 33.20, p < 0.001, \eta_p^2 = 0.52$; see Figure 4b). Adult thresholds had less variability than 3 and 4 year olds ($t[25.31] = -8.11, p < 0.001$ and $t[23.25] = -8.38, p < 0.001$ corrected for unequal variances, respectively). There was a trend for the thresholds of 3 year olds to be less variable than those of 4 year olds ($t[45] = -1.84, p = 0.07$). A borderline level of physical engagement x age interaction was also found ($F[2, 61] = 3.14, p = 0.05, \eta_p^2 = 0.09$). This effect was the result of tendency for reduced threshold variability for the lower physical engagement task (touchscreen) with the 3 year olds only. All other main effects and interactions failed to reach significance.

Threshold Validity

The proportion of participants that attained at least 75% correct (target level of performance) on at least one staircase run was computed. Separate chi-square analyses were conducted in order to investigate whether there were significant associations between age group and achieving target level of performance, response paradigm group and achieving target level of performance, as well as level of physical engagement and achieving target level of performance. There was a significant effect of age with 100% of adults reaching target level of performance or above, 54% of 4 year olds and 19% of 3 year olds ($\chi^2(2) = 58.27, p=0.001$). There was no significant association between response paradigm and reaching target level of performance. There was no significant association between level of physical engagement and meeting target level of performance.

Separate Chi-Square analyses were conducted for each age group to examine the association of response paradigm and achieving target level of performance, as well as presentation type and achieving target level of performance. There was a significant association found between the level of physical engagement and meeting target level of performance for the 3 year olds only ($\chi^2(1) = 6.70, p=0.02$). For the condition with greatest levels of physical engagement (3D figurines), 33.3% of the 3 year olds reached the target level of performance, while only 4.2% reached the target level in the laptop condition. No significant effects were found for the other age groups.

Overall Success (Summed Accuracy, Variability and Validity Scores)

For both the response paradigm and level of physical engagement in the task, threshold estimation success improved with age. For all age groups the task with the greatest level of physical engagement (3D figurines) resulted in the most successful threshold estimates (see Figure 5). The choice of response paradigm had the greatest impact on the

successful estimation of adult thresholds, where threshold estimates were most successful with the 3AFC version. Similarly, the 3AFC procedure resulted in more successful threshold estimates for the 4 year olds, but success using the 3AFC procedure only differed slightly from the 2AFC procedure (difference of 1.82) and was much smaller than the difference found between the same procedures for the adults (difference of 10.14). In contrast, the 2AFC procedure resulted in slightly better threshold estimates for 3 year olds than the 3AFC (difference of 4.26).

Figure 5 about here

Discussion

The aim of this experiment was to investigate the impact of response paradigm (3I-2AFC vs 3I-3AFC), level of physical engagement (2D touchscreen laptop vs 3D figurines) and age (3- and 4-year-old preschool children and adults) on the estimation of discrimination thresholds. We hypothesised that response paradigms which are likely to place less demands on working memory capacity (i.e. 3I-2AFC) and tasks which involve greater levels of physical engagement (i.e. 3D figurines) would lead to lower and more reliable threshold estimates for preschool children. We also hypothesised that the overall most successful response paradigm would be 2AFC and the overall most successful presentation would be the figurines.

In line with findings from previous threshold estimation studies with preschoolers, the 3- and 4-year-old preschoolers in this study had significantly larger frequency discrimination thresholds than adults (Banai & Yifat, 2011; Jensen & Neff, 1993). Importantly, the younger 3-year-old preschoolers had significantly larger threshold estimates than other preschoolers

just 4 to 17 months older (4-year-old group). These findings parallel those of Jensen and Neff who identified significant improvements in frequency discrimination thresholds between 4 and 6 years old as evidence of the development of auditory discrimination resolution. Direct comparisons between discrimination thresholds from other studies are complicated by the differences in measurement approaches between tasks. However, Jensen and Neff, presenting typically between 2.5-3 times more trials than presented in the current study, found average discrimination thresholds for a group of 4 year olds that were substantially lower than the ones reported in the current study for the same aged participants. Snowling, Gooch, McArthur and Hulme (2018), presenting 3 times more trials than the current study, found larger discrimination thresholds for 4 year olds than Jensen and Neff, but still slightly smaller ones than found in the current study. One key difference between the current study and previous studies with 3 and 4 year olds, was the size of the starting discrimination difference. Previous frequency discrimination studies with preschoolers (3 – 4 year old) have all begun with suprathreshold starting discrimination differences of 400 Hz or 500 Hz (Banai & Yifat, 2011; Jensen & Neff, 1993; Snowling, Gooch, McArthur, & Hulme, 2018). In contrast, the current study used a relatively large initial discrimination (1000 Hz), which has the advantage of typically resulting in correct responses by children for the first few trials. Starting closer to the expected threshold region will obviously enable reaching a threshold more rapidly than starting further away. However, having a starting discrimination (ceiling) that is closer to the expected threshold region also means that even particularly inattentive participants can end up with lower estimated thresholds, albeit unreliable ones. In the absence of repeated measurements to establish the reliability of these values, we cannot be certain whether the thresholds found by Jensen and Neff are similarly representative of the combined sensory and non-sensory capacities that seem to have impacted the threshold estimates found in the current study.

Unsurprisingly the threshold estimates for the preschoolers resulted from more variable (less stable) threshold estimates than the threshold estimates for adults. There was a surprising trend (borderline statistical significance) for thresholds to be more stable across staircase runs for the less physically engaging task (touch screen) than the more physically engaging task (figurines) for the 3 year olds only. This should be interpreted with caution since the size of the effect was very small and the 3 year olds assented to less runs per session than the older 4 year olds. This effect could be the result of the higher and less accurate thresholds in the touch screen condition. If the 3 year olds consistently failed to reach their threshold region, then responses near ceiling would likely have resulted in less variability when error responses were made. In contrast, greater movement in the staircase towards the threshold region of any individual would allow for greater variability when error responses are made. If this was the case then an association between the number of 3 year olds reaching target performance (75% correct) and the level of physical engagement of the task would be expected, as found.

All previous auditory discrimination tasks with preschoolers have been presented on digital screens. Despite most studies involving cartoon-based images to represent the task stimuli (e.g. cartoon animal characters producing the auditory stimuli) in order to present the task as an engaging game for children, these studies still report larger and more variable threshold estimates in comparison to older children and adults (Banai & Yifat, 2011; Jensen & Neff, 1993; Moore et al., 2008). One potential explanation for these variable thresholds is that these tasks are still not engaging enough to sustain the attention of preschoolers.

Furthermore, responses to these tasks are typically collected by participants motioning their responses to an experimenter rather than engaging directly with the task and making their own responses (e.g. pressing the response buttons themselves). This lack of physical engagement may also be a significant barrier to sustaining attention on the task and thus lead

to unreliable threshold estimates. We found some evidence for this, in that threshold estimates were significantly lower for both groups of preschoolers when higher levels of physical engagement were required by the task. In contrast, the adult thresholds were unaffected by this manipulation in the simple effects analysis. Measures of test re-test reliability support the interpretation of threshold estimates being more accurate when measured using 3D figurines than a 2D touch screen laptop for both the 3 year olds and 4 year olds. This highlights that while development of auditory discrimination resolution may have some impact on the higher thresholds attained for preschoolers than adults, non-sensory factors related to task engagement also have a substantial impact. This finding confirmed our hypothesis and fits with the understanding that whenever younger children are more physically engaged in a task, they attend more (e.g. Bertrand & Camos, 2015). Interestingly, the level of task engagement did not have a significant impact on the proportion of participants reaching the target level of performance. In the context of staircase procedures, the number of correct responses made by participants around the target performance level can serve as a proxy for how well they attended to the procedure (i.e. fewer correct responses suggest that participants were attending less during the procedure and did not complete the procedure listening to trials around their true threshold). If as proposed, the proportion of participants reaching the target level of performance is a proxy measure of the group level of attention, then failure to find an association between the figurines condition and higher correct performance suggests the groups as a whole were not less attentive in the figurines than the touch screen condition. It is possible that the impact of level of engagement on reducing thresholds and improving reliability were derived from differences in metacognitive strategies and motivation to succeed on the task, rather than differences in attention.

Previous studies have argued that the task response paradigm is an important factor in achieving accurate threshold estimates with young children especially (e.g. Sutcliffe &

Bishop, 2005). Previous behavioural and computation studies with adult-like participants, adults and older children (above 6 years of age) have found that presenting a greater number of alternative choices to discriminate in a task is more preferential for accurate and reliable threshold estimation (Sutcliffe & Bishop, 2005; Schlauch & Rose, 1990). However, the majority of preschool studies have used just two alternative choices believing perhaps, potentially erroneously, that this is an appropriate paradigm to use with the restricted working memory capacities and attention spans of this young age group. Incorporating a third interval, but still maintaining two alternative choices is possible by designating one of the intervals to be a reference tone (e.g. AXB), which remains fixed throughout the procedure. This requires participants to discriminate which interval (A or B) deviates in a particular dimension from the reference tone (X). Banai and Yifat (2011) found that the discrimination thresholds of preschoolers are significantly related to working memory scores in a task without a reference tone ($r = -0.34$, $p < 0.01$) and not in a task with a reference tone ($r = -0.22$, $p > 0.05$). Even if the association between working memory and discrimination thresholds had been significant for the task with a reference tone, the change in explained variance between auditory discrimination thresholds and working memory when a reference tone is used (4%), and when a reference tone is not used (12%), is substantial. In the current study we found no significant impact of response paradigm on the threshold estimates, variability in the thresholds or the proportion of valid threshold estimates achieved. When looking at the overall success of the procedure, the 2AFC paradigm was most successful for the youngest age group, in line with our hypotheses. However, there was very little difference in success of the response paradigms for 4 year olds and success with adults was substantially improved using the 3AFC response paradigm.

Overall, high levels of physical engagement (the figurines condition) was found to be the most successful procedure to generate threshold estimates for all participant age groups

compared to the laptop condition. The success of the procedure seemed to be more impacted by the level of physical task engagement, rather the task response paradigm. While it was initially expected that the 2AFC paradigm would be less taxing on the working memory resources of the preschool groups, according to Banai & Yifat's (2011) previous findings, the current methodology did have some differences which may have led to the failure to differences between the 2AFC and 3AFC paradigms. It should be noted that both levels of physical engagement (figurines and touch screen) in this task were higher than used in previous studies of this age group and may have modulated the memory demands of the task as well as the attentional demands. Similarly, the task used in the current study employed both positive reinforcement like previous studies (food for the cartoon pigs when correct responses were achieved), but also a penalty for incorrect responses (removal of the pigs' hats so they got cold when incorrect responses were made). Reinforcement schedules are known to have an impact on the task performance of preschool children. For example, the presentation of positive reinforcement following trial completion of a visual discrimination task has been shown to lead to 5 and 6 year old children completing more trials, compared to a task where trials were followed by a penalty (Van Den Boomen & Peters, 2017).

General Conclusions

The overall aim of this study was to investigate the impact of key procedural parameters to measuring the auditory discrimination thresholds of preschool children. Reliable thresholds are difficult to measure in preschool-aged children and changes in some procedural elements of auditory discrimination tasks have the potential to improve the estimation of discrimination thresholds in these children. Establishing how much the discrimination threshold estimates from any particular procedure (and set of parameters) are

an accurate measure of sensory capacities and how much they are biased by non-sensory factors is an important goal for researchers. Simulations can be used to assess the impact of very controlled non-sensory variables like attention on threshold estimation from any procedure and set of parameters. However, the relevance of these simulations relies on manipulating non-sensory factors in a way that the target population (in the current case preschoolers) actually function.

The choice of adaptive procedure (including the step rule), number of trials and step size all have an impact on the how successfully threshold estimates can be computed for populations with varying levels of attention, like preschoolers. However, even with the clear selection of parameters (adaptive procedure, number of trials, and step-size) based on simulations of variability in attention, threshold reliabilities are not as high as would be expected if all non-sensory factors were minimised. Some non-sensory factors like motivation are not well understood in information processing terms (if they even can be), and are therefore not easy to model. However, the influence of motivational factors, like the level of physical engagement within a task should not be underestimated. Direct empirical study of the way attention, working memory and motivation affects auditory discrimination performance will be vital to the design of the most successful procedures for threshold estimation in preschoolers.

Although the current study focused on establishing the most appropriate auditory discrimination procedures for preschoolers, the methods described here lay the foundations for designing auditory discrimination procedures for a range of non-typical adult populations. For example, populations with developmental delays may respond positively to the adjustments recommended for younger children by the current study. Enabling the reliable measurement of auditory discrimination in populations of older children with attention deficits may also be an application for the procedures designed in the current study.

Acknowledgements: NG and TF co-wrote the paper. NG and TA designed the modelling procedures. NG and TF designed the experimental procedures. All authors contributed independently to different programming and analyses, and all authors discussed the results and reviewed the final manuscript. This project was funded by a Department for Education and Learning studentship grant awarded to TF and TA for the supervision of NG.

Declaration of Interest: The authors declare no competing financial interest.

Endnotes

¹The percentage sign could represent a wide range of different stimulus parameters that can be incremented and decremented by addition and subtraction, such as decibels or log-frequency. In the model, the “10%” step is the parameter difference that would lead to a 10% improvement in performance, correcting for chance, at the steepest part of the psychometric function. Relating the model to Experiment 3, the model steps are equivalent to $m \log_2(\Delta F / F_{ref})$, where ΔF is the frequency difference between the pure tones in the first and second interval, F_{ref} is an arbitrary reference frequency, and m is an arbitrary constant related to the gradient of the psychometric function. Setting $m = 20\%$ and $F_{ref} = 30.25$ Hz would correspond to an initial difference of 1000 Hz, a 50% threshold (corrected for chance) at a frequency difference of 125 Hz, and a psychometric function whose peak gradient is equivalent to a 20% improvement in performance (corrected for chance) for each doubling of frequency.

References

- Amitay, S., Irwin, A., Hawkey, D. J. C., Cowan, J. A., & Moore, D. R. (2006). A comparison of adaptive procedures for rapid and reliable threshold assessment and training in naive listeners. *The Journal of the Acoustical Society of America*, *119*, 1616–1625.
- Anvari, S. H., Trainor, L. J., Woodside, J., & Levy, B. A. (2002). Relations among musical skills, phonological processing, and early reading ability in preschool children. *Journal of Experimental Child Psychology*, *83*, 111–130.
- Banai, K., & Yifat, R. (2011). Perceptual anchoring in preschool children: Not adultlike, but there. *PLoS ONE*, *6*(5), e19769.
- Banai, K., & Yuval-Weiss, N. (2013). Prolonged development of auditory skills: A role for perceptual anchoring? *Cognitive Development*, *28*(3), 300–311.
- Bavin, E. L., Grayden, D. B., Scotti, K., & Stefanakis, T. (2010). Testing auditory processing skills and their associations with language in 4-5-year-olds. *Language and Speech*, *53*(Pt 1), 31–47.
- Bertrand, R., & Camos, V. (2015, January). The role of attention in preschoolers' working memory. *Cognitive Development*, *33*, 14-27.
- Betts, J., Mckay, J., Maruff, P., & Anderson, V. (2006). The Development of Sustained Attention in Children: The Effect of Age and Task Load. *Child Neuropsychology*, *12*(3), 205–221.
- Boets, B., Vandermosten, M., Poelmans, H., Luts, H., Wouters, J., & Ghesquière, P. (2011). Preschool impairments in auditory processing and speech perception uniquely predict future reading problems. *Research in Developmental Disabilities*, *32*(2), 560–570.

- Carlson, S. M., Moses, L. J., & Breton, C. (2002). How Specific is the Relation between Executive Function and Theory of Mind? Contributions of Inhibitory Control and Working Memory. *Infant and Child Development*, 11(2), 73–92.
- Chermak, G. D., Somers, E. K., & Seikel, J. A. (1998). Behavioral Signs of Central Auditory Processing Disorder and Attention Deficit Hyperactivity Disorder. *Journal of the American Academy of Audiology*, 9(1), 78–84.
- Cornsweet, T. M. (1962). The staircase method in psychophysics. *The American Journal of Psychology*, 75(3), 485–491.
- Corriveau, K. H., Goswami, U., & Thomson, J. M. (2010). Auditory processing and early literacy skills in a preschool and kindergarten population. *Journal of Learning Disabilities*, 43(4), 369–382.
- Creel, S. C. (2015). Ups and Downs in Auditory Development: Preschoolers' Sensitivity to Pitch Contour and Timbre. *Cognitive Science*, 1–31.
- Dawes, P., Sirimanna, T., Burton, M., Vanniasagaram, I., Tweedy, F., & Bishop, D. V. M. (2009). Temporal auditory and visual motion processing of children diagnosed with auditory processing disorder and dyslexia. *Ear and Hearing*, 30, 675–686.
- Ferguson, M. A., Hall, R. L., Riley, A., & Moore, D. R. (2011). Communication, Listening, Cognitive and Speech Perception Skills in Children With Auditory Processing Disorder (APD) or Specific Language Impairment (SLI). *Journal of Speech, Language, and Hearing Research*, 54(1), 211–227.
- Fitch, R. H., Miller, S., & Tallal, P. (1997). Neurobiology of speech perception. *Annual Review of Neuroscience*, 20, 331–353.

- France, S. J., Rosner, B. S., Hansen, P. C., Calvin, C., Talcott, J. B., Richardson, A. J., & Stein, J. F. (2002). Auditory frequency discrimination in adult developmental dyslexics. *Perception & Psychophysics*, 64(2), 169–179.
- Fritzley, V. H., & Lee, K. (2003). Do young children always say yes to yes-no questions? A metadevelopmental study of the affirmation bias. *Child Development*, 74(5), 1297–1313.
- García-Pérez, M. A. (1998). Forced-choice staircases with fixed step sizes: Asymptotic and small-sample properties. *Vision Research*, 38(12), 1861–1881.
- Gathercole, V. C. (1985). More and more and more and more. *Journal of Experimental Child Psychology*, 40(1), 73–104.
- Gomes, H., Molholm, S., Christodoulou, C., Ritter, W., & Cowan, N. (2000). The development of auditory attention in children. *Frontiers in Bioscience*, d108-120.
- Henderlong, J., & Lepper, M. R. (2002). The effects of praise on children's intrinsic motivation: A review and synthesis. *Psychological Bulletin*, 128(5), 774–795.
- Jensen, J. K., & Neff, D. L. (1993). Development of basic auditory discrimination in preschool children. *Psychological Science*, 42(2), 104–107.
- Kaernbach, C. (1991). Simple adaptive testing with the weighted up-down method. *Perception & Psychophysics*, 49, 227–229.
- Killan, C. F., Royle, N., Totten, C. L., Raine, C. H., & Lovett, R. E. S. (2015). The effect of early auditory experience on the spatial listening skills of children with bilateral cochlear implants. *International Journal of Pediatric Otorhinolaryngology*, 79(12), 2159–2165.
- Klenberg, L., Korkman, M., & Lahti-Nuuttila, P. (2001). Differential Development of Attention and Executive Functions in 3- to 12-Year-Old Finnish Children.

- Developmental Neuropsychology*, 20(1), 407–428.
- Law, J., Vandermosten, M., Ghesquière, P., & Wouters, J. (2017). Predicting future reading problems based on pre-reading auditory measures: A longitudinal study of children with a familial risk of dyslexia. *Frontiers in Psychology*, 8(124).
- Levitt, H. C. C. . (1971). Transformed Up-Down Methods in Psychoacoustics. *The Journal of the Acoustical Society of America*, 49(2B), 467–477.
- Lin, C. H., Hsiao, C., & Chen, W. (1999). Development of Sustained Attention Assessed Using the Continuous Performance Test among Children 6–15 Years of Age. *Journal of Abnormal Child Psychology*, 27(5), 403–412.
- Litovsky, R. (2015). Development of the auditory system. *Handbook of Clinical Neurology*, (Vol. 129, pp. 55-72). Elsevier.
- Luciana, M., Conklin, H. M., Hooper, C. J., & Yarger, R. S. (2005). The Development of Nonverbal Working Memory and Executive Control Processes in Adolescents. *Child Development*, 76(3), 697–712.
- Mahone, E. M., & Schneider, H. E. (2012). Assessment of Attention in Preschoolers. *Neuropsychology Review*, 22(4), 361–383.
- Moore, D. R., Ferguson, M. A., Edmondson-Jones, A. M., Ratib, S., & Riley, A. (2010). Nature of Auditory Processing Disorder in Children. *Pediatrics*, 126(2), E382–E390.
- Moore, D. R., Ferguson, M. A., Halliday, L. F., & Riley, A. (2008). Frequency discrimination in children: perception, learning and attention. *Hearing Research*, 238(1–2), 147–154.
- Olsho, L. W., Koch, E. G., & Halpin, C. F. (1987). Level and age effects in infant frequency discrimination. *The Journal of the Acoustical Society of America*, 82(2), 454–464.

- Olsho, L. W., Schoon, C., Sakai, R., Turpin, R., & Sperduto, V. (1982). Auditory Frequency Discrimination in Infancy. *Developmental Psychology*, 18(5), 721–726.
- Porter, H. L., Grantham, D. W., Ashmead, D. H., & Tharpe, A. M. (2014). Binaural Masking Release in Children With Down Syndrome. *Ear and Hearing*, 35(4), e134–e142.
- Rammsayer, T. H. (1992). An experimental comparison of the weighted up-down method and the transformed up-down method. *Bulletin of the Psychonomic Society*, 30(5), 425–427.
- Rickard, N. A., Heidtke, U. J., & O’Beirne, G. A. (2013). Assessment of auditory processing disorder in children using an adaptive filtered speech test. *International Journal of Audiology*, 52(10), 687–697.
- Rose, R. M., Teller, D. Y., & Rendleman, P. (1970). Statistical properties of staircase estimates. *Perception & Psychophysics*, 8(4), 199–204.
- Saberi, K., & Green, D. M. (1996). Adaptive psychophysical procedures and imbalance in the psychometric function. *J Acoust Soc Am*, 100(1), 528–536.
- Schlauch, R. S., & Rose, R. M. (1990). Two-, three-, and four-, interval forced-choice staircase procedures: Estimator bias and efficiency. *Journal of the Acoustical Society of America*, 88(2), 732–740.
- Schneider, B. A., Trehub, S. E., Morrongiello, B. A., & Thorpe, L. A. (1989). Developmental changes in masked thresholds. *The Journal of the Acoustical Society of America*, 86(5), 1732–1742.
- Snowling, M. J., Gooch, D., McArthur, G., & Hulme, C. (2018). Language Skills, but Not Frequency Discrimination, Predict Reading Skills in Children at Risk of Dyslexia. *Psychological Science*, 0956797618763090.

- Sutcliffe, P., & Bishop, D. (2005). Psychophysical design influences frequency discrimination performance in young children. *Journal of Experimental Child Psychology*, 91, 249–270.
- Talcott, J. B., Witton, C., Hebb, G. S., Stoodley, C. J., Westwood, E. A., France, S. J., Hansen, P. E., & Stein, J. F. (2002). On the relationship between dynamic visual and auditory processing and literacy skills; results from a large primary-school study. *Dyslexia*, 8(4), 204–205.
- Thompson, N. C., Cranford, J. L., & Hoyer, E. (1999). Brief-tone frequency discrimination by children. *Journal of Speech, Language, and Hearing Research*, 42(5), 1061–1068.
- Van Den Boomen, C., & Peters, J. C. (2017). Spatial frequency discrimination: Effects of age, reward, and practice. *PLoS ONE*, 12(1), 1–13.
- Wang, H.-L. S., Chen, I.-C., Chiang, C.-H., Lai, Y.-H., & Tsao, Y. (2016). Auditory perception, suprasegmental speech processing, and vocabulary development in Chinese preschoolers. *Perceptual and Motor Skills*, 123(2), 365–382.
- Wightman, F., & Allen, P. (1992). Individual differences in auditory capability among preschool children. In *Developmental Psychoacoustics* (pp. 113–133).
- Wightman, F., Dolan, T., Allen, P., Jamieson, D., & Kistler, D. (1989). Temporal resolution in children. *Journal of the Acoustical Society of America*, 60(3), 611–624.
- Witton, C., Talcott, J. B., & Henning, G. B. (2017). Psychophysical measurements in children: challenges, pitfalls, and considerations. *PeerJ*, 5(5), e3231.
- Zhang, Y., Moore, D., Guiraud, J., Molloy, K., Yan, T., & Amitay, S. (2016). Auditory Discrimination Learning : Role of Working Memory. *PLoS ONE*, 11(1), e0147320 (1-

18).

Tables

Table 1. Threshold estimate test re-test reliabilities by level of physical engagement and age.

Age	Level of Physical Engagement							
	Medium (touch screen)				High (Figurines)			
	t score	ICC	95% CI	F	t score	ICC	95% CI	F
3 year olds	0.82	0.59	0.05 - 0.82	2.42*	0.79	0.73	0.38-0.88	3.72**
4 year olds	1.51	0.79	0.50 - 0.91	4.68***	1.07	0.88	0.73 - 0.95	8.62***
Adults	0.86	-	-	-	-0.36	0.52	-0.22 - 0.81	2.07

Note. *p<0.05, **p<0.01, ***p<0.001

Figure Legends

Figure 1. Results of performance for child-like and adult-like simulations based on changes in staircase and number of trials. (a) Mean threshold accuracy, (b) Mean threshold variability, (c) Threshold validity measured as the % of invalid staircase simulations, (d) Overall success (summed scores). Smaller summed scores indicate greater success. Error bars depict 95% confidence intervals.

Figure 2. Results of performance for child-like and adult-like simulations based on changes in step size¹. (a) Mean threshold accuracy, (b) Mean threshold variability, (c) Threshold validity measured as the % of invalid staircase simulations, (d) Overall success (summed scores). Smaller summed scores indicate greater success. Error bars depict 95% confidence intervals.

Figure 3. Factorial experimental design measuring response paradigm (2AFC; 3AFC) by level of physical engagement (high = 3D figurines; medium = 2D touch screen images). Illustrations and figurines were matched as closely as possible.

Figure 4. (a) Log mean (Hz) difference threshold. (b) Log variability (Hz) of difference threshold across runs. Error bars depict the standard error of the mean.

Figure 5. Measures of threshold estimation success for the 3 year olds, 4 year olds and adults.

(a) Overall success (summed scores) collapsed across response paradigm for the medium and high levels of physical engagement. (b) Overall success (summed scores) collapsed across levels of engagement for the 2AFC and 3AFC response paradigms. Smaller summed scores indicate greater success.