Can Auditory Objects be Subitized?

Katherine L. Roberts[1], Nicola J. Doherty[2], Elizabeth A. Maylor[2], and Derrick G. Watson[2]

[1]Nottingham Trent University, UK

[2]University of Warwick, UK

Author Note

Correspondence should be addressed to Katherine L. Roberts, Department of Psychology, Nottingham Trent University, Nottingham, UK. Email: kate.roberts@ntu.ac.uk

This work was funded by a British Academy/Leverhulme grant (SG131129) awarded to KLR, EAM and DGW. We thank Katie Jones and Luke Hodson for testing participants in Experiment 2.

Word count: 10,769

Abstract

In vision, humans have the ability to mentally ‘tag’ approximately four objects, allowing us to monitor, attend, and interact with them. As a consequence, we can rapidly and accurately enumerate up to four objects – a process known as subitizing. Here, we investigate whether a similar ability exists for tagging auditory stimuli and find that only two or three auditory stimuli can be enumerated with high accuracy. We assess whether this high accuracy indicates the existence of an auditory subitizing mechanism, and if it is influenced by factors known to influence visual subitizing. Based on accuracy, Experiments 1 and 2 reveal a potential auditory subitizing mechanism only when stimuli are spatially separated, as is the case for visual subitizing. Experiment 3 failed to show any evidence of auditory subitizing when objects were separated in time, rather than space. All three experiments provide only limited evidence for an age-related decline in auditory enumeration of small numbers of objects. This suggests that poor auditory tagging does not contribute significantly to older adults’ difficulties in multi-talker conversations. We hypothesize that although auditory subitizing might occur, it is restricted to approximately two spatially-separated objects due to the difficulty of parsing the auditory scene into its constituent parts.

Keywords: auditory, enumeration, subitizing, aging, location

Public Significance Statement

This study provides initial evidence for an early ‘tagging’ mechanism that allows people to mentally 'tag' multiple sounds in the environment for later processing. Tagging was only possible when sounds were spatially separated, as is the case with visual tagging. Older adults showed similar tagging to young adults, suggesting that this ability does not decline with age and is thus unlikely to contribute to older adults’ difficulties in multi-talker conversations.

Can Auditory Objects be Subitized?

To what extent can we detect and tag multiple objects in the environment? This question has been answered extensively for the visual modality, but we have much less knowledge regarding our awareness of multiple auditory objects. For over a hundred years, since the pioneering work of Jevons (1871), vision researchers have investigated our rapid and potentially preattentive tagging of key objects within a visual scene ('subitizing'; Kaufman, Lord, Reese, & Volkmann, 1949). Such work has addressed how we can individuate identical visual objects, track them over time, and understand their relative spatial locations (Pylyshyn, 1989). The wealth of vision research that has probed this question, including studies of subitizing and multiple object tracking, underlines its importance to visual perception as a whole. Yet we know almost nothing about tagging multiple auditory objects.

Research into awareness of multiple visual objects has demonstrated that we can 'tag', and enumerate, approximately four objects, in parallel (Pylyshyn, 1989; Trick & Pylyshyn, 1993, 1994; but see Olivers & Watson, 2008). These tags, or indexes, provide information about the location of the objects relative to each other and to ourselves, and also provide a link to those objects to allow individual attentional processing of each item (Pylyshyn, 1989, 2001). The ability to simultaneously tag a limited number of items provides many adaptive core and fundamental functions such as allowing us to coordinate and move a limited focus of attention between several identical visual objects or features, determine spatial relationships between items, and coordinate our eye movements (Pylyshyn, 1989). One striking consequence of this tagging system is that, by assigning tags, it is possible to track up to four moving target objects amid an array of identical moving distractor objects (Pylyshyn & Storm, 1988). Theoretically, a tagging system such as this should also prove beneficial in the auditory domain, in which assigning tags to different sound sources (e.g., different talkers, car alarm, radio) could help us to monitor those sound sources over time and to direct attention to (and switch attention between) the sound sources of interest.

A further consequence of this visual tagging system is that approximately four visual objects can be enumerated (‘subitized’) quickly and accurately (Jevons, 1871; Kaufman et al., 1949) by assigning and determining how many of the tags are currently bound to items (Pylyshyn, 1989; Trick & Pylyshyn, 1994). Because the number of tags is limited to approximately four, subitization is also limited to four items. In contrast, enumerating more than four visual objects (typically called counting) requires the disengagement and re-assignment of tags which is more error prone, and results in a relatively large increase in time for each additional item that has to be enumerated (Trick & Pylyshyn, 1994). Complementing the behavioral data, neuroimaging and neuropsychological evidence suggests that rapid visual subitizing and ‘serial’ enumeration beyond the subitizing range (counting) involve separate cortical mechanisms (Demeyere et al., 2010, 2014). In terms of parsing visual input, some obvious applied benefits of visual subitizing include allowing us to recognize large numbers quickly (e.g., 1000000) if the digits are organized into groups of three (1,000,000).

In the present work, we test whether there exists a similar subitizing system for auditory objects. In Experiments 1 and 2, an ‘object’ is loosely defined as a coherent auditory stream arising from a single source, such as bird song, piano music, someone speaking, or a car alarm (Griffiths & Warren, 2004; Kubovy & van Valkenburg, 2001; see below for a more detailed discussion of auditory object formation). In Experiment 3, the auditory objects are sequentially presented pure tones and frequency-modulated tones. As in the visual domain, the ability to rapidly assign individual tags to auditory objects would allow those objects to be subitized, facilitate directing attention to those of interest, and provide an index to monitor future changes.

Age-Related Declines in Visual and Auditory Tagging

In all three experiments, we ask whether there is an age-related deficit in auditory tagging, which might underlie older adults' difficulties in listening situations that are attentionally demanding. Older adults in particular find it difficult to listen amid competing speech or noise, due to age-related declines in auditory perception and cognition (Roberts & Allen, 2016; Schneider et al., 2002). Older adults also report difficulties in multi-talker conversations, such as missing the start of what each new talker is saying, and these difficulties are linked to their feelings of handicap, even when taking into account any hearing loss (Gatehouse & Noble, 2004).

In addition to establishing the limits of auditory enumeration, we also examine whether impaired awareness and tagging of multiple auditory objects might contribute to the difficulties that older adults experience in multi-talker conversations. In simple visual enumeration tasks, older adults are slower overall than young adults, but they have a similar subitizing span and similar response-time slopes (ms per item) in both the subitizing and counting ranges (Watson, Maylor, Allen, & Bruce, 2007; Watson, Maylor, & Bruce, 2005a; Watson, Maylor, & Manson, 2002). An age-related deficit in visual subitizing emerges only when targets must be enumerated among distractors. Under these conditions, in contrast to young adults, older adults are unable to subitize targets (Watson et al., 2002), particularly when the targets and distractors are perceptually similar (Watson et al., 2007). This is likely to be due to older adults' impaired visual attention abilities. Deficits in visual attention processes and/or increased system noise would mean that representations of targets and distractors may not be clearly differentiated. As a consequence, older adults would be less able to apply multiple visual tags in parallel, and would instead have to apply tags in a spatially serial manner (Watson et al., 2007).

Auditory perception and cognition are also impaired in old age (Schneider et al., 2002), making it difficult for older adults to segregate a target auditory stream from distractor streams (Ben-David et al., 2012; Ezzatian et al., 2015). This could well impact on older adults’ ability to subitize auditory objects irrespective of whether or not irrelevant distractor sounds are also present. Weller, Best, Buchholz, and Young (2016) found that older, hearing impaired adults had difficulty enumerating more than two auditory sources, but they did not study the effects of older age per se, independent of hearing impairment. Here we focus on older adults with normal hearing or mild hearing impairment only.

The Role of Perceptual Organization

There are two key requisites that allow visual objects to be rapidly tagged, and therefore subitized. The first is that they must be spatially separated (Pylyshyn, 1989; Watson, Maylor & Bruce, 2005b). For example, the number of shapes present in a scene cannot be subitized if they are placed in a concentric arrangement (Saltzman & Garner, 1948; Trick & Pylyshyn, 1993). Similarly, subitizing of visual properties that do not belong to unique objects (e.g., how many colors are present in a scene) is severely limited to approximately two different features. This may indicate that a scene is parsed preattentively into a foreground color and background colors, and that the background colors are not further segmented (Watson et al., 2005b). This distinction between space-based and feature-based visual subitizing reflects the critical role of spatial location in the visual system, from coding at the retina and in early visual cortex through to visual object formation and selection (Kubovy & van Valkenburg, 2001; Lamy & Tsal, 2000).

The auditory system, on the other hand, is primarily focused on spectral and temporal information. Concurrent sounds enter the ear together and are initially coded according to frequency. A process of auditory scene analysis (Bregman, 1990) is then necessary to integrate frequency components associated with a single sound source (e.g., one person’s voice) and segregate them from different sound sources. The auditory system uses various spectral and temporal cues to achieve this object formation (and segregation), including common time-course, onset and offset times, pitch, and harmonicity. Spatial location does not facilitate individual object formation, but can be useful for streaming and attending to objects over time (Shinn-Cunningham, 2008). Auditory objects are therefore primarily formed and selected on the basis of their spectrotemporal profile (Griffiths & Warren, 2004; Kubovy & van Valkenburg, 2001; Shinn-Cunningham, 2008), but there can be some benefit from spatially separating target sounds from distractors (Freyman et al., 2001; Hawley et al., 2004). In Experiments 1 and 2 of the present work, in addition to the central question of whether or not sounds can be subitized we also assess whether spatial separation is necessary, or even beneficial, to auditory tagging and subitizing. In Experiment 3, we consider the role of temporal separation in the auditory task, and examine enumeration of sequentially presented auditory objects.

The second requisite for efficient visual tagging and subitizing is that it must be possible to identify the target objects without using focal attention (Trick & Pylyshyn, 1993). For example, it is possible to subitize target letter Os amid distractor Xs, but not target Os amid distractor Qs (Trick & Pylyshyn, 1993). The need for targets to be identifiable preattentively could prove to be a limiting factor for tagging concurrent auditory stimuli. In audition, all sounds in the environment enter the ear together, and the auditory system has the non-trivial task of segregating the incoming sounds into their constituent streams (Bregman, 1990). Whereas low-level perceptual grouping is likely to occur preattentively, organizing those sounds into coherent streams over time appears to require attention (Carlyon et al., 2001; Cusack et al., 2004; but cf. Macken et al., 2003; Sussman et al., 2007).

Cusack et al. (2004) presented multiple auditory streams to their participants and found that the data were consistent with a ‘hierarchical decomposition’ model. According to this model, participants are initially aware of broad categories of the sounds currently in the environment (e.g., music, speech, traffic), but they only have access to sub-streams (e.g., guitar, drums, singers) when focal attention is directed toward that specific stream (in this case, the music). It is likely that several factors will determine the number of streams available at the highest level of the hierarchy, including frequency separation (Brochard et al., 1999; Cusack et al., 2004), stimulus intensity (Botte et al., 1997), and top-down cognition such as attention (Dowling et al., 1987). The hierarchical decomposition model suggests a slightly more elaborate scene analysis than the simple foreground/background distinction proposed for feature-based visual subitizing (Watson et al., 2005b), implying that more than two concurrent sounds might be identifiable preattentively. It is also possible for listeners to be aware of the number of auditory objects (sounds or sound sources) in the environment without segregating each individual stream. In the example above, recognizing the sounds of a guitar and a drum would provide evidence of two auditory objects without it being necessary to perceptually segregate those streams.

Auditory Enumeration

Few previous studies have investigated the enumeration of concurrent auditory stimuli. Two studies have suggested that concurrent auditory stimuli cannot be subitized, and that even counting accuracy is poor for two or more stimuli (McLachlan et al., 2012; Thurlow & Rawlings, 1959). However, in both of these studies it is not clear whether the limiting factor was participants’ ability to enumerate the objects, or simply to segregate the objects, which were pure tones (Thurlow & Rawlings, 1959) and harmonic complexes (McLachlan et al., 2012). More recent studies (Kawashima & Sato, 2015; Vitevitch & Siew, 2016; Weller et al., 2016; Zhong & Yost, 2017) investigated enumeration of concurrent talkers and found that only between three and five talkers could be accurately counted (with accuracy of more than 50%). Although Kawashima and Sato’s (2015) work did not consider auditory subitizing, their data indicate a potentially bilinear enumeration function, consistent with fast and accurate enumeration of two or three talkers, followed by slower and less accurate enumeration of larger numbers of talkers. In contrast, Zhong and Yost's (2017) enumeration data show that enumeration accuracy decreases linearly with increasing numbers of sound sources before levelling off for five or more sound sources.

Here, we present three experiments that specifically investigate whether auditory objects can be subitized, and if so, determine the subitizing span for auditory objects, the factors that influence auditory subitizing, and whether there is an age-related decline in auditory subitizing. Experiments 1 and 2 explore enumeration of concurrent auditory stimuli. The stimuli were a set of auditory clips (e.g., hens clucking, piano solo) that have previously been used in auditory search tasks (Eramudugolla et al., 2005, 2008). They have distinct spectro-temporal profiles and each sound is clearly discriminable against a background of the other sounds (Eramudugolla et al., 2005). Experiment 3 investigates enumeration of sequential auditory stimuli, by asking participants to enumerate target tones within a rapidly presented sequence of target and distractor tones.

General Methods

Participants

Young participants were recruited from the University of Warwick's student population. Older adults were recruited from the Warwick Age Study Panel of healthy community-dwelling volunteers. Pure tone audiometry was used to assess hearing thresholds at frequencies between 250 and 8000 Hz (Maico MA25 screening audiometer with DD45 headset). Young adults were excluded if their thresholds exceeded 25 dB HL at any individual frequency (two participants in Experiment 1 and one each in Experiments 2 and 3). Older adults were recruited who reported 'fair' or better hearing, but were then included regardless of their audiometric thresholds. A measure of hearing impairment was obtained by averaging over five frequencies (250, 500, 1000, 2000 and 4000 Hz) for the better ear. The average threshold was then used to determine the impact of mild hearing impairment on auditory enumeration.

In all three experiments we tested 20 young participants. This sample size was based on our earlier research that indicated that 18 participants would give a strong test of feature versus object-based visual subitizing (Watson et al., 2005b) and Kawashima and Sato's (2015) research that showed that 12 participants were sufficient to detect differences in counting accuracy when auditory stimuli were presented from the same or different locations. Watson et al. (2007) found that a sample of 20 young and 20 older adults was sufficient to detect age-related differences in subitizing ability when targets were presented amid distractors. We initially recruited a larger sample (n = 30) to allow older participants with severe age-related hearing loss to be excluded. However, we found that we were able to recruit older adults with comparatively good hearing and so recruited only 20 older participants in Experiment 2 (conducted after Experiments 1 and 3).

One young and one older adult participated in both Experiments 1 and 2; one young and three older adults participated in Experiments 2 and 3; two young and seven older adults participated in Experiments 1 and 3.

Ethical approval was granted by the University of Warwick's Humanities and Social Sciences Research Ethics Committee. All participants gave written, informed consent. Young participants received £6 compensation; older participants received £10 inconvenience allowance plus travel expenses.

Stimuli and Apparatus

All experiments were conducted in sound-attenuated testing booths at the University of Warwick. Stimuli were presented via Sennheiser HD518 headphones at comfortable volume levels. In Experiments 1 and 2, the stimuli were 10-second clips of eight distinctive sounds taken from Eramudugolla et al. (2005). The sounds were hens clucking, Gregorian chant, piano solo, cello solo, male horse-race commentator (English), female news reader (Hindi), police siren, and alarm-clock ring, with equalized RMS sound pressure levels. Each sound clip was 5-s in duration and was immediately repeated once, to create 10-s clips.

Procedure

In all three experiments, participants were familiarized with the stimuli and then completed a short practice session before beginning the experimental trials. Participants pressed the space bar to initiate each trial, in response to an instruction screen ("Press the space bar to continue"). The screen went immediately blank and the sounds were played after a 1-s delay. The task was always to decide how many sounds were present. When participants believed they knew the answer, they pressed the space bar. The sounds then stopped and the question "How many?" appeared on screen. The participant entered their response by pressing a number on the keypad. On-screen feedback indicated accuracy and the correct number of sounds (e.g., "Correct! There were 2 sounds."). Feedback was presented for 800 ms and was followed by a 1-s blank screen before the instruction screen appeared for the next trial. Participants were instructed to respond with the space bar as quickly and accurately as possible. Response times (RTs) were calculated as the time from sound onset to the space bar being pressed to ensure that RTs were not affected by the time taken to find the correct response key (see Watson et al., 2002, for a discussion of this method).

Older participants additionally completed the Speech, Spatial and Qualities of Hearing questionnaire (SSQ; Gatehouse & Noble, 2004). This contains 14 questions regarding the participants' speech perception in different situations (Speech), 17 questions about their ability to localize sounds (Spatial), and 18 questions relating to the quality of the sounds that they hear (Qualities). Each question is answered by marking a point on a line anchored between 0 (no ability) and 10 (perfect ability). An example Speech question is:

"You are in a group of about five people in a busy restaurant. You can see everyone else in the group. Can you follow the conversation?" (response line anchored with 0 'not at all' and 10 'perfectly').

Data Analysis

Accuracy and RT data were entered into analyses of variance (ANOVAs). RTs were included for correct trials only, and excluded if they were more than three *SD*s above the participant's mean for that cell of the design. When there was only one correct RT for a condition/numerosity, it was included if it fell within three *SD*s of the participant's overall mean on correct trials. These exclusion rules led to the removal of less than 1% of the RT data. Where Mauchley's test of sphericity indicated that sphericity could not be assumed, a Greenhouse-Geisser correction was applied. This is indicated by non-integer degrees of freedom. Estimated effect sizes are indicated by partial eta squared values (η^2_p).

Experiment 1

In Experiment 1, we investigated young and older adults' ability to correctly enumerate concurrent auditory clips that varied in their spectrotemporal profile. We looked for evidence of auditory subitizing when stimuli were presented at the same location, and we additionally tested whether the first requisite of visual subitizing – that targets must be spatially separated – also applies to the auditory domain.

Method

Participants. Participants were 20 young adults (7 male, mean age 21 years, range 18-29) and 30 older adults (10 male, mean age 72 years, range 63-84). For the older participants, better-ear averages were 20 dB HL or below for 19 participants and between 20 and 40 dB HL for 11 participants, indicating a mild hearing loss (BSA guidelines, 2011). Young adults had an average BEA of 4.5 dB HL whereas older adults with normal hearing had an average BEA of 15.4 dB HL. All but one of the older participants had approximately symmetric thresholds (10 dB HL or less between the average for each ear). The remaining participant had an asymmetry of 24 dB HL.

Stimuli and apparatus. On each trial, between one and six sounds were presented simultaneously. Interaural time differences (ITDs) were used to lateralize the sounds to eight different locations, from approximately 90° to the left to 90° to the right (+/- 590, 454, 272 and 91 μs; exact lateralization depends on head size). Sounds lateralized using ITDs appear to arise from locations along an imaginary line between the two ears. In the ‘different locations’ condition, the stimuli were presented from up to six of the eight locations (selected at random, with each stimulus occupying a different location). In the ‘same location’ condition, one of the eight locations was selected at random and all sounds originated from that location.

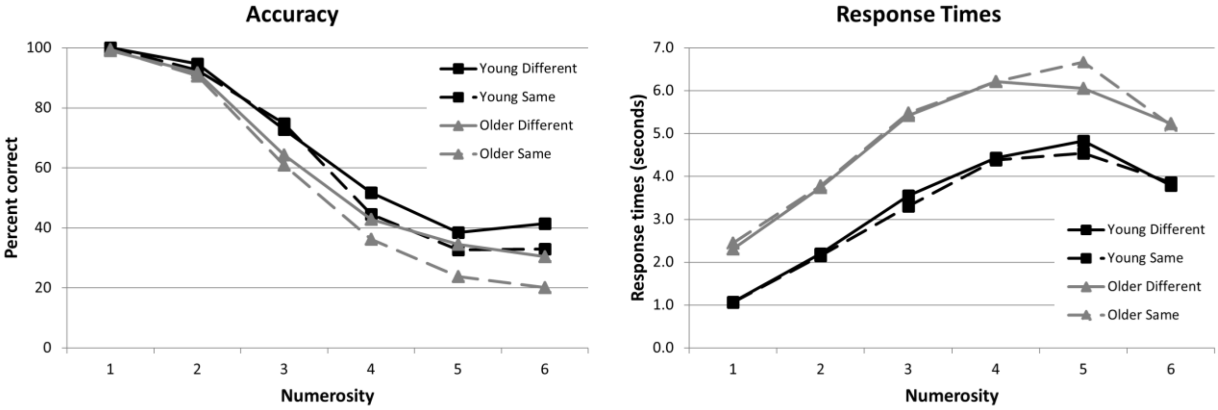
Procedure. Participants were initially played a 5-s clip of each sound with an accompanying label on screen (e.g., ‘piano solo’). They were then played the sounds again and asked to name them (with any plausible name accepted), to ensure that they were familiar with the identity of all stimuli.

Participants first completed 12 practice trials (two trials for each numerosity). The experiment then comprised eight blocks of 30 trials (5 trials for each of the 6 numerosities, in random order). The blocks alternated between the ‘different location’ (four blocks) and ‘same location’ (four blocks) conditions, with the initial condition counterbalanced across participants.

Results

Accuracy (proportion correct) and mean RTs on correct trials were entered into mixed analyses of variance (ANOVAs) including age group (young, older), location (same, different), and numerosity (1 to 6). See Figure 1 for accuracy and RT data.

Figure 1. Accuracy and response times in Experiment 1, for each numerosity (1 to 6 auditory objects), for young (black) and older (gray) participants, and when sounds were lateralized to different locations using interaural timing differences (solid lines) or from the same location (dashed lines).



Participants became less accurate as numerosity increased, $F(2.7, 128.9) = 340.19$, $p < .001$, $\eta^2_p = .876$, and were less accurate when the sounds came from the same location, $F(1, 48) = 24.66$, $p < .001$, $\eta^2_p = .339$. There was also an interaction between numerosity and location, $F(3.5, 168.5) = 4.64$, $p = .002$, $\eta^2_p = .088$. Paired *t*-tests with a Bonferroni correction for multiple comparisons (critical $p = .008$) showed that presenting the sounds from different locations improved enumeration for between 4 and 6 auditory objects, but not for smaller numbers of auditory objects ($t(49) = -1.00$, 1.43, 0.61, 3.33, and 3.72, for 1 - 6 sounds, respectively, $p = .32$, .16, .54, .002, .002, and .001).

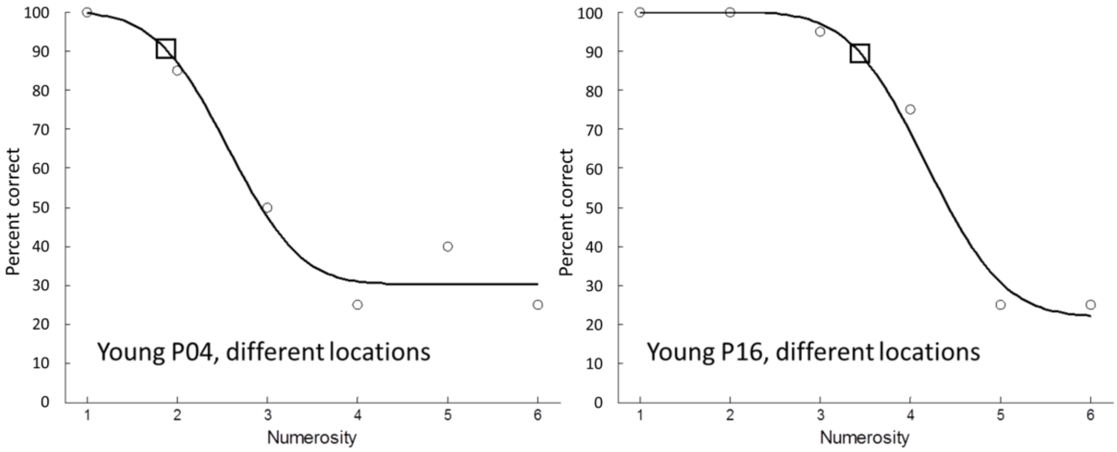
Older adults were significantly less accurate overall, $F(1, 48) = 16.17$, $p < .001$, $\eta^2_p = .252$, but age group did not interact significantly with numerosity or location (all *p*s > .1).

Results from the ANOVA on the RT data showed a similar pattern to the accuracy data: there was slowing with older age, $F(1, 41) = 8.68$, $p = .005$, $\eta^2_p = .18$, and increasing numerosity, $F(1.6, 63.6) = 73.16$, $p < .001$, $\eta^2_p = .64$. Although older participants were slower overall this did not interact with numerosity, $F < 1$. There was no significant effect of location, no interaction between numerosity and location, and no three-way interaction between numerosity, location and age (all *p*s > .1).

Subitizing span. The maximum number of items that can be subitized is often estimated in visual studies by fitting a bilinear function to the RT or accuracy data. The subitizing span is then indicated by the flex point between the relatively flat subitizing slope and the steeper counting slope. Because auditory enumeration was especially poor with larger numbers of items, it does not produce a linear counting slope. Instead, as can be seen in Figure 1, the accuracy data form a sigmoid even when the largest numerosity is removed to prevent any potential influence of ‘end’ effects (Mandler & Shebo, 1982; Trick & Pylyshyn, 1994; Watson & Humphreys, 1999).

To estimate a subitizing span, we therefore used Psignifit 3.0 (Fründ et al., 2011) in Matlab (The Mathworks: Natick, MA) to fit a sigmoidal (Gaussian) function to the accuracy data from all six numerosities (see Figure 2 for examples). For two young and three older participants we obtained a bad fit to the data (observed deviance outside the 95% confidence interval derived from bootstrapping with 1000 samples). These participants were removed from the following analyses. We then calculated the point of maximum curvature in the left-hand section of the function (constrained to >= 0 objects), to estimate an upper limit for the subitizing span. The average results across participants are shown in Table 1. Note that a non-integer subitizing span would indicate that a subitizing mechanism is used on a proportion of trials with the higher integer numerosity (e.g., a subitizing span of 2.5 might suggest that participants are able to subitize two items on every trial, and three items on half the trials).

Figure 2. Example individual data from Experiment 1. Plots show individual participants' accuracy at each numerosity (open circles), the fitted Gaussian function (solid line), and the point of maximum curvature (open square). Participant 4 (left plot) has an estimated subitizing span of 1.9; Participant 16 (right plot) has an estimated subitizing span of 3.4.



Plots of the RT data showed clearly linear slopes for numerosities between 1 and 4 (see Figure 1). Nonetheless, for completeness we also fit the sigmoid function to the RT data. In some conditions, at some numerosities, participants failed to make any correct responses. Due to these missing data, functions could only be fitted to RT data from 23 of the older adults. There was also a poor fit for three young adults and one older adult. For the remaining participants, estimated 'subitizing spans' based on RTs were less than two in all conditions (see Table 1).

Table 1

Average Subitizing Spans Estimated from the Point of Maximum Curvature of a Gaussian Function Fitted to the Accuracy and Response-Time Data from Experiment 1

		Subitizing span	
Age	Condition	Accuracy	Response Times
Young	Different	2.56 (2.33 – 2.80)	1.36 (1.03 – 1.69)
Young	Same	2.71 (2.50 – 2.92)	1.34 (1.01 – 1.68)
Older	Different	2.38 (2.19 – 2.58)	1.09 (0.80 – 1.37)
Older	Same	2.29 (2.11 – 2.46)	1.24 (0.97 – 1.56)

Note. 95% confidence intervals are shown in parentheses.

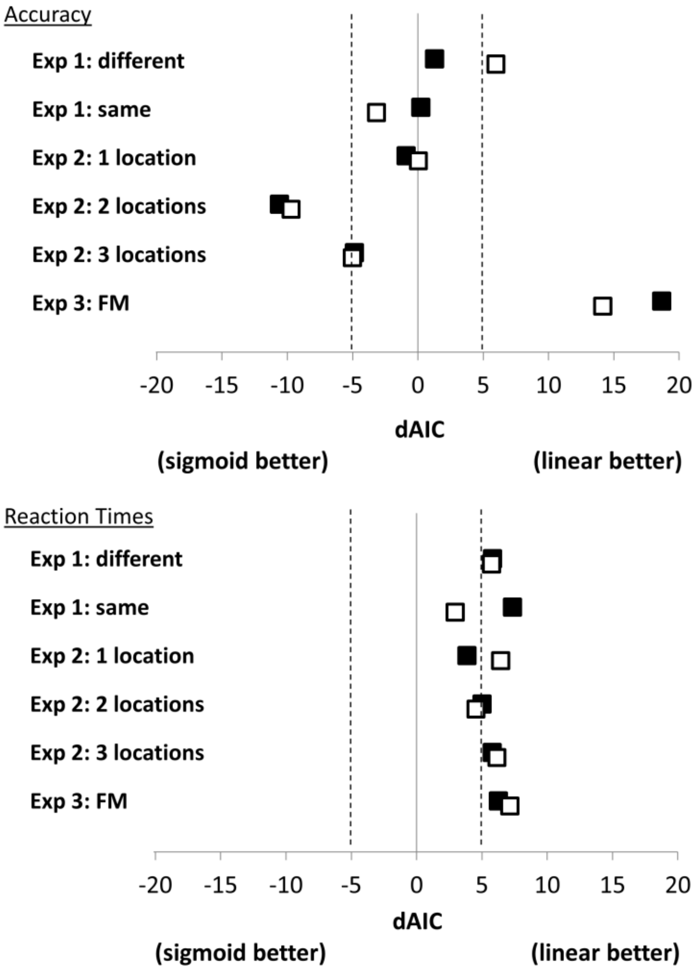
Direct comparison of linear and nonlinear functions. In visual enumeration studies, evidence for separate subitizing and counting mechanisms often comes from fitting linear and bilinear functions to the data and assessing which provides the better fit. If a bilinear function fits the data better than a linear function, this provides evidence consistent with the existence of two separate enumeration mechanisms (subitizing and counting).

In the auditory enumeration task, this approach is complicated by the limit on the number of auditory objects that can be enumerated accurately, which leads to an asymptote in the data after approximately four or five auditory objects. Therefore, in order to compare the sigmoidal and linear functions, we fitted linear functions to the first four data points, in addition to the sigmoid functions described above. We then calculated the residual sum of squares (RSS) for the linear and sigmoidal functions over those four data points, for each individual participant and experimental condition, to determine which function provided the best fit. If the sigmoid provided a better fit, this would be suggestive of an auditory subitizing mechanism. Comparison of goodness of fit was evaluated using Akaike Information Criterion

(AIC) to control for differences in the number of parameters in the linear and sigmoidal functions. Note that this approach is somewhat conservative: if participants can subitize four auditory objects then the linear function will provide an excellent fit to the data, despite the existence of a subitizing mechanism.

Figure 3 shows the mean sigmoidal-linear AIC difference (dAIC) across participants in each experiment, age group, and condition, for the accuracy and RT data. A dAIC of 0 indicates that the linear and sigmoidal functions provide a similar fit to the data. A dAIC of less than -5 would provide reasonably strong evidence that the sigmoid provides a better fit than the linear function, whereas a dAIC of more than 5 would indicate that the linear function is superior (Baguley, 2012). The result of this analysis shows that the sigmoid does not provide a better fit than the linear function in any of the conditions in Experiment 1. Therefore there is no evidence that participants are using an auditory subitizing mechanism in Experiment 1.

Figure 3. Comparison of the linear and sigmoid functions, for the accuracy and response-time data. Residuals were compared for the first four data points, taking into account the number of parameters (Akaike Information Criterion; AIC). The difference between the AIC values (dAIC: sigmoidal minus linear) is plotted, for all conditions and experiments. Filled squares: young participants; white squares: older participants.



Effect of audiometric hearing status. Data from the older adults were entered into an ANOVA with hearing status (normal/mild impairment) as a between-participants factor and numerosity and location as within-participants factors. There was no significant effect of hearing status, $F(1, 28) = 2.31$, $p = .140$, $\eta^2_p = .08$, and no significant interactions involving hearing status (all *p*s > .1).

Summary

Participants were able to enumerate approximately two auditory objects with high accuracy (> 90%), indicating worse enumeration accuracy than is found with visual objects. Older adults were slower and less accurate overall, but this did not worsen with increasing numbers of objects.

Lateralizing the auditory objects to different locations using ITDs improved enumeration of larger numbers of auditory objects slightly (four to six), but did not influence the enumeration of smaller numbers of auditory objects. Audiometric hearing thresholds did not influence older adults' enumeration accuracy.

Experiment 2

In Experiment 2 we investigated further the effect of spatial separation on auditory enumeration. Unlike the visual system, auditory information is not processed in spatiotopic maps in the cortex. The location of auditory stimuli is calculated based on differences in the arrival time and level of the signal at the two ears (interaural time differences (ITDs) and interaural level differences (ILDs)), and spectral changes introduced by the head and external ears. Recent evidence suggests that auditory localization can be based on the relative activation within three spatial channels: left, midline and right (Briley et al., 2016). In Experiment 1, stimuli were separated using ITDs only. However, effects of spatial attention can be stronger when ILDs are also present, as this enables attention to be directed toward a particular spatial channel (Roberts et al., 2009). In Experiment 2 we tested the hypothesis that auditory stimuli can be subitized only if they fall within separate spatial channels. We presented between one and five concurrent sound clips (using the same sound clips as in Experiment 1), lateralized to different locations using generic head-related transfer functions (HRTFs) (Gardner & Martin, 1994). HRTFs include ITDs and ILDs, as well as spectral cues introduced by the head and external ears. Stimuli were either presented to one spatial location

(90° left, midline, or 90° right), two locations (left and midline, left and right, or midline and right) or three locations (left, midline and right). Each location (left, midline, right) corresponds to a spatial channel (Briley et al., 2016).

Method

Participants. Participants were 20 young adults (7 male, mean age 24 years, range 19-30) and 20 older adults (8 male, mean age 76 years, range 67-87). For the older participants, better-ear averages over five frequencies were below 20 dB HL for 10 participants, between 20 and 40 dB HL for nine participants indicating a mild hearing loss, and 43 dB HL for one participant, indicating a moderate hearing loss. Young adults had an average BEA of 6.0 dB HL whereas older adults with normal hearing had an average BEA of 13.9 dB HL. All but six of the older participants had approximately symmetric thresholds (<= 10 dB HL difference). Three had asymmetries between 10 and 15 dB HL, two had asymmetries between 20 and 25 dB HL, and one had an asymmetry of 40 dB HL.

Stimuli and apparatus. On each trial, between one and five sounds were presented simultaneously. Stimuli were convolved with generic HRTFs in Matlab, to lateralize the sounds to three possible locations (90° left, midline, 90° right). Sounds lateralized using individualized HRTFs appear to arise from an external sound source. With generic HRTFs the percept varies depending on head shape and size. Sounds were either presented from one, two or three locations, as described above. When the number of sound clips exceeded the target number of locations, more than one sound clip was presented from one or more of the locations, distributed evenly between the available locations. Participants completed 36 trials at each numerosity. A maximum of five, rather than six, concurrent stimuli were presented in Experiment 2 to maximize the number of trials in each condition. This followed from the finding in Experiment 1 that six concurrent stimuli could not be reliably enumerated.

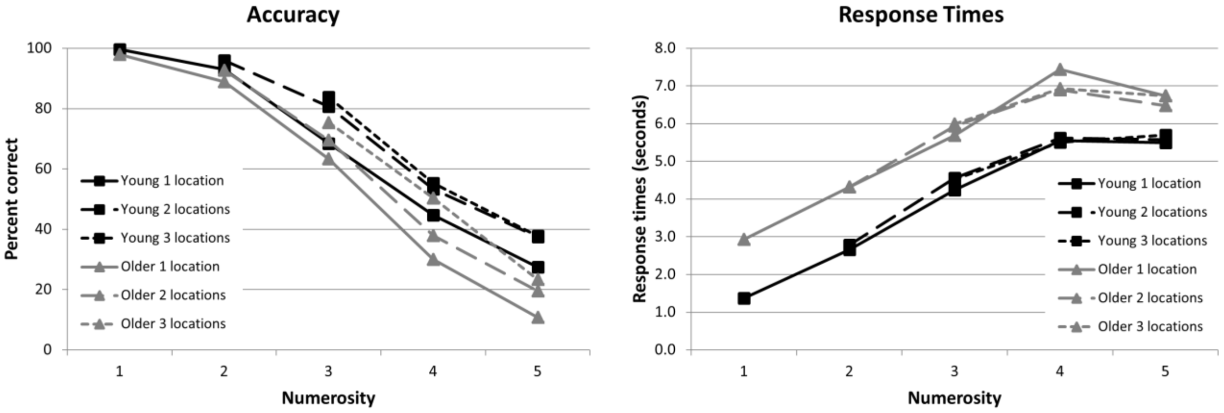
Procedure. Participants were familiarized with the stimuli as in Experiment 1. Participants first completed ten practice trials. The experiment then comprised four blocks of 45 trials (9 trials for each of the 5 numerosities, presented in a random order).

Results

Two separate analyses were conducted to investigate the effect of the number of locations on enumeration performance. Data for two locations were only available for numerosities of two or more, and data for three locations were only available for numerosities of three or more. We first compared performance when stimuli were presented from one or two locations, using data from numerosities of between two and five. We then compared performance when stimuli were presented from two or three locations, using data from numerosities between three and five.

Accuracy data (see Figure 4) were first entered into a mixed ANOVA including age group (young, older), numerosity (2 to 5) and number of locations (1 or 2). This analysis includes all numerosities for which sounds were presented from 1 location and 2 locations. Accuracy decreased with increasing numerosity, $F(2.2, 85.2) = 327.80$, $p < .001$, $\eta^2_p = .90$, and was worse when stimuli were presented from 1 location compared with 2 locations, $F(1, 38) = 42.29$, $p < .001$, $\eta^2_p = .53$, but there was no interaction between numerosity and number of locations, $F(2.5, 94.9) = 1.06$, $p = .37$, $\eta^2_p = .03$, suggesting that presenting the stimuli from two different locations had the same benefit at each numerosity between 2 and 5. Accuracy was worse for older adults, $F(1, 38) = 14.53$, $p < .001$, $\eta^2_p = .28$, and there was a significant interaction between age group and numerosity, $F(3, 114) = 3.48$, $p = .018$, $\eta^2_p = .08$, such that older adults showed a bigger decrease in accuracy with each additional sound clip (see Figure 4). Age group did not interact with the number of locations, $F < 1$, and there was no three-way interaction between age group, numerosity and locations, $F < 1$.

Figure 4. Accuracy and response times in Experiment 2. Data are shown for each numerosity (1 to 5), for young and older participants (black, gray), with stimuli from 1, 2 or 3 locations.



To evaluate whether there was an additional benefit for presenting stimuli from 3 spatial locations, accuracy data were entered into a mixed ANOVA including age group (young, older), numerosity (3 to 5) and number of locations (2 or 3). As before, accuracy was worse for older adults, $F(1, 38) = 11.59$, $p = .002$, $\eta^2_p = .23$, decreased with numerosity, $F(1.5, 56.5) = 144.00$, $p < .001$, $\eta^2_p = .79$, and when stimuli were presented from 2 locations compared with 3 locations, $F(1, 38) = 11.00$, $p = .002$, $\eta^2_p = .23$. There was an interaction between age group and the number of locations, $F(1, 38) = 4.15$, $p = .049$, $\eta^2_p = .10$. Post-hoc comparisons revealed that older, but not young, adults benefitted when the stimuli were presented from 3 locations compared with just 2 locations (young: mean difference = .018, 95% confidence interval = -.019 to .054; older: mean difference = .074, 95% CI = .029 to .118).

Similar ANOVAs conducted on the RT data indicated that for 1 and 2 locations, RTs increased with increasing numerosity, $F(1.6, 42.4) = 79.09$, $p < .001$, $\eta^2_p = .75$, and older participants had significantly longer RTs, $F(1, 26) = 6.37$, $p = .018$, $\eta^2_p = .20$. There were no other significant effects or interactions in the RT data (all *p*s > .14). A similar pattern was found when the RT data were analyzed for 2 and 3 locations: effects of numerosity, $F(1.4, 44.2) = 14.44$, $p < .001$, $\eta^2_p = .32$, and age (albeit marginal), $F(1, 31) = 3.16$, $p = .085$, $\eta^2_p = .09$, but there was no effect of the number of locations and no significant interactions (all *p*s > .5).

Subitizing span. As in Experiment 1, we estimated the subitizing span by fitting sigmoid (Gaussian) functions to the accuracy data for the 1-location, 2-location, and 3-location conditions and extracting the point of maximum curvature (Table 2). When the number of locations exceeded the numerosity, data for a lower number of locations were included (e.g., all three functions were fitted using data for 1 numerosity from 1 location). This allows the subitizing span to be directly compared across all three numbers of locations. Three older participants were excluded: one because the sigmoidal function was a bad fit to the data and two because of accuracy of less than 90% for enumerating a single sound clip.

Functions were also fitted to the RT data. In some conditions, at some numerosities, participants failed to make any correct responses. Due to these missing data, functions could only be fitted to RT data from 18 young adults and 9 older adults. There was also a poor fit for one young adult and two older adults. For the remaining participants, estimated ‘subitizing spans’ were less than two in all conditions (Table 2).

Table 2

Average Subitizing Spans Estimated from the Point of Maximum Curvature of a Gaussian Function Fitted to the Accuracy and Response-Time Data from Experiment 2

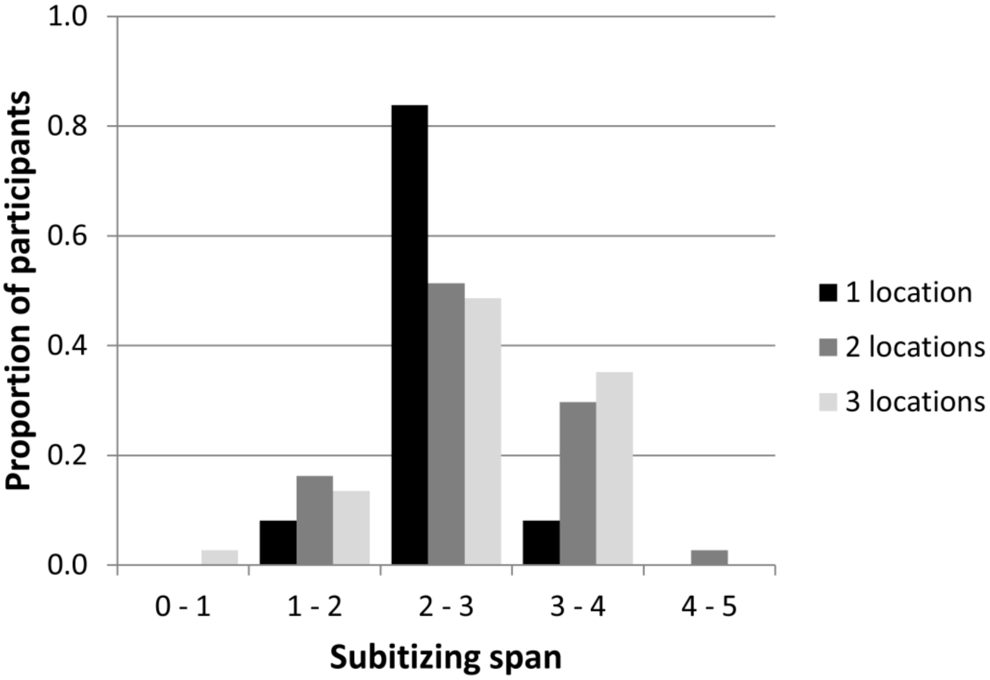
		Subitizing span	
Age	Condition	Accuracy	Response Times
Young	1 location	2.43 (2.26 – 2.60)	1.50 (1.10 – 1.01)
	2 locations	2.90 (2.62 – 3.18)	1.75 (1.56 – 1.94)
	3 locations	2.83 (2.48 – 3.18)	1.58 (1.24 – 1.93)
Older	1 location	2.44 (2.25 – 2.63)	1.75 (1.12 – 2.38)
	2 locations	2.69 (2.39 – 2.99)	1.52 (1.24 – 1.79)
	3 locations	2.65 (2.27 – 3.03)	1.75 (1.52 – 1.98)

Note. 95% confidence intervals are shown in parentheses.

Comparison of linear and nonlinear functions. As described in Experiment 1, we directly compared linear and sigmoidal functions to test for separate subitizing and counting mechanisms. Figure 3 shows the mean dAIC (sigmoidal – linear) for each age group and condition, for the accuracy and RT data. For the accuracy data, the sigmoid provides a significantly better fit to the data than the linear function, but only when the auditory objects are presented from two or more locations. In contrast, the linear function appears to provide a better fit to the RT data in all three conditions. The same pattern is found for the young and older adults.

Effects of age and location conditions on subitizing spans. The points of maximum curvature were entered into a mixed ANOVA including age group (young, older) and number of locations (1, 2, and 3). There was a significant main effect of the number of locations, *F*(1.7, 58.4) = 4.61, *p* = .019, η^2_p = .12. Post-hoc *t*-tests revealed a significant difference in the point of maximum curvature between 1 and 2 locations, *t*(36) = -3.69, *p* = .001, and between 1 and 3 locations, *t*(36) = -2.47, *p* = .018, but not between 2 and 3 locations, *t*(36) = 0.38, *p* = .71. There was no effect of age group, *F* < 1, and no interaction between number of locations and age group, *F* < 1. See Figure 5 for the distribution of subitizing spans, collapsed across age groups.

Figure 5. Distribution of subitizing spans in Experiment 2, for the different location conditions, collapsed across young and older participants. Subitizing spans were estimated by finding the point of maximum curvature of a fitted Gaussian function.



Effect of audiometric and self-reported hearing status. Older participants were divided into those with normal hearing (n = 10) and those with a mild or moderate hearing impairment (n = 10). Adding hearing status to the Numerosity x Locations ANOVAs did not reveal any significant effects of hearing.

We investigated whether there is a link between auditory subitizing (based on the accuracy data) and audiometric or self-reported hearing ability. Average SSQ responses were 6.98 (*SD* = 1.6) for Speech, 7.0 (1.5) for Spatial and 8.0 (1.3) for Qualities of hearing, on a scale from 0 to 10 where 10 indicates no self-reported hearing difficulties. There were no significant correlations between either hearing or SSQ scores and the maximum curvature with one, two or three locations, following Bonferroni correction for multiple comparisons (critical *p* = .004).

Summary

As in Experiment 1, participants were able to enumerate approximately two auditory objects with high accuracy. However, in this experiment, when stimuli were lateralized to different locations using generic HRTFs rather than ITDs, we did find an increase in enumeration accuracy when stimuli were presented from more than one location. When sounds were presented from more than one location, we found that a sigmoid function provided a better fit than a linear function to the accuracy (but not the RT) data, potentially indicating the existence of separate subitizing and counting mechanisms. The accuracy-based estimated subitizing span was greater when sounds were presented from more than one location, but young adults did not gain an additional benefit when sounds were presented from three locations.

Older adults were less accurate overall, and showed a larger decrease in accuracy with each additional auditory object compared with young adults. Note that older, but not young, adults became more accurate when stimuli were presented from three locations compared with two. In this condition, older adults’ performance approached that of young adults.

Experiment 3

In Experiment 3 we consider the role of temporal separation of auditory stimuli and address a second requisite for subitizing: that target stimuli must be available at preattentive levels of processing.

Whereas visual subitizing relies on spatial separation, the emphasis on spectrotemporal information in audition may indicate that auditory subitizing would be facilitated by temporal, rather than spatial, separation. Camos and Tillmann (2008) suggested that subitizing of sequential stimuli is possible if the stimuli can be held within a ‘single focalization’ of attention. They investigated enumeration of sequential auditory stimuli and found a discontinuity after two items. However, this work used a rapid sequence of events (80-ms stimulus onset asynchrony) that may have resulted in masking, and moreover, numerosity could be estimated from the length of each sequence. In contrast, here we keep sequence length the same but vary the relative number of targets and distractors (analogous to the approach used previously in visual enumeration studies; see Watson et al., 2002, for a discussion). Two other studies (ten Hoopen & Vos, 1979; Repp, 2007) have found that enumeration of auditory sequences improves when the stimuli are organized into groups of two (Repp, 2007), or two to five tones (ten Hoopen & Vos, 1979) using location or pitch as a grouping cue. These studies suggest that participants may have been able to subitize tones within a group, and then count the number of groups.

Generally, in visual search tasks, search for a target that has the absence of a feature is less efficient than search for a target that has the presence of a feature – a search asymmetry (Treisman & Souther, 1985). Thus a letter Q target can be detected preattentively among letter O distractors, but detection of a target O among Q distractors results in slow, inefficient search. Applied to enumeration, target Qs can be subitized amid distractor Os, but target Os cannot be subitized amid distractor Qs (Trick & Pylyshyn, 2003). We exploited a similar asymmetry that occurs in the auditory modality (Cusack & Carlyon, 2003) and investigated whether participants could subitize target frequency-modulated (FM) tones amid distractor pure tones, but not target pure tones amid distractor FM tones. Stimuli were 100-ms pure and frequency-modulated tones at different frequencies, to reduce forward and backward masking and reduce the likelihood that target tones were perceived as oddballs (Camos & Tillmann, 2008).

Method

Participants. Participants were 20 young adults (5 male, mean age 22 years, range 18-30) and 30 older adults (13 male, mean age 72 years, range 66-79). Pure tone audiometry indicated that older adults' better-ear averages were below 20 dB HL for 23 participants and between 20 and 40 dB HL for 7 participants, indicating a mild hearing loss. Young adults had an average BEA of 9.2 dB HL whereas older adults with normal hearing had an average BEA of 14.3 dB HL. All older participants had approximately symmetric thresholds (<= 10 dB HL difference).

Stimuli and apparatus. The stimuli were 100-ms pure and frequency-modulated tones at frequencies between 440 and 570 Hz, in 10-Hz steps. Stimuli were cosine gated for 10 ms at the start and end. FM tones had a modulation frequency of 10 Hz and a maximum frequency change of 200 Hz. The sampling frequency was 44,100 Hz.

On each trial, participants heard a series of 14 tones, with 50-ms inter-stimulus intervals.

Procedure. Participants were initially played the pure ("beep") and FM ("raindrop") tones to familiarize them with the stimuli.

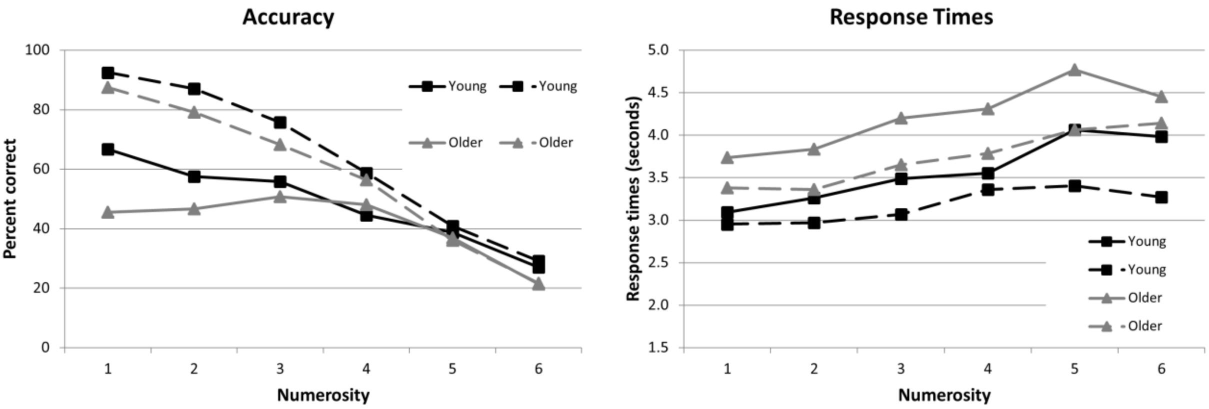
On each block of trials, participants were instructed to count either the pure tones ("beeps") or FM tones ("raindrops"). Each sequence of 14 tones included between 1 and 6 target sounds. When participants were ready to respond, they pressed the space bar and the text ‘How many beeps?’ or ‘How many raindrops?’ appeared on screen.

Participants first completed six practice trials for each block type (count pure tones/FM tones). The experiment then comprised six blocks of 12 trials per condition (2 trials for each of the 6 numerosities, presented in a random order). The blocks alternated between the pure and FM conditions, with the initial condition counterbalanced across participants.

Results

Accuracy and RT data are shown in Figure 6. Accuracy was entered into an ANOVA including age group (young, older), target type (count pure/FM tones), and numerosity (1-6). Participants were significantly more accurate when counting FM tones than pure tones, $F(1, 48) = 69.42$, $p < .001$, $\eta^2_p = .59$, and with smaller numerosities, $F(5, 240) = 158.54$, $p < .001$, $\eta^2_p = .77$. The accuracy benefit for counting FM tones was greater at smaller numerosities, resulting in a significant interaction between condition and numerosity, $F(3.3, 159.3) = 22.33$, $p < .001$, $\eta^2_p = .32$. Paired *t*-tests with a Bonferroni correction for multiple comparisons (critical $p = .008$) showed that accuracy was better for FM targets than pure targets for numerosities up to 4 ($t(49) = 8.34$, 8.67, 5.95, 3.14, 0.11, and 0.50, for 1 – 6 targets, respectively, $p < .001$, $< .001$, $< .001$, .003, .915 and .620).

Figure 6. Accuracy and response times in Experiment 3. Data are shown for each numerosity (1 to 6), for young and older participants (black, gray), and when the task was to enumerate pure tones amid frequency-modulated (FM) distractors (Pure), or FM tones amid pure-tone distractors (FM).



Older adults were not significantly less accurate overall, $F(1, 48) = 2.15$, $p = .15$, $\eta^2_p = .04$, but age group did interact with numerosity, $F(5, 240) = 2.56$, $p = .03$, $\eta^2_p = .05$. Young participants were more accurate than older participants at small numerosities but performance was similar at larger numerosities, resulting in a near-significant difference (Bonferroni-corrected critical $p = .008$ (two tailed) or $p = .017$ (one tailed)) between the age groups at numerosities 1, $F(1, 48) = 6.68$, $p = .013$, $\eta^2_p = .12$, and 2, $F(1, 48) = 3.84$, $p = .056$, $\eta^2_p = .07$, but not at larger numerosities (all ps > .2).

RT data showed a similar pattern of results. Participants responded more quickly when counting FM tones compared with pure tones, $F(1, 29) = 10.89$, $p = .003$, $\eta^2_p = .27$, and were faster at smaller numerosities, $F(2.3, 66.2) = 9.55$, $p < .001$, $\eta^2_p = .25$. Older adults were slower overall, $F(1, 29) = 4.19$, $p = .050$, $\eta^2_p = .13$, but age did not interact with target type (pure/FM) or numerosity (all ps > .3).

Subitizing span. Participants were unable to reliably enumerate small numbers of pure tones amid FM tones, and so we did not attempt to estimate a subitizing span in this condition. For the FM-tone enumeration task, we fitted sigmoid (Gaussian) functions to the accuracy data and extracted the point of maximum curvature (Table 3). Three young and six older participants were excluded due to accuracy below 80% when enumerating a single target.

Functions were also fitted to the RT data. In some conditions, at some numerosities, participants failed to make any correct responses. Due to these missing data, functions could only be fitted to RT data from 18 young adults and 23 older adults. There was also a poor fit for one young adult. For the remaining participants, estimated ‘subitizing spans’ were less than two for both age groups (Table 3).

Table 3

Average Subitizing Spans Estimated from the Point of Maximum Curvature of a Gaussian Function Fitted to the Accuracy and Response-Time Data from Experiment 3, when the Task was to Enumerate Frequency-modulated Tones

	Subitizing span	
Age	Accuracy	Response Times
Young	2.71 (2.22 – 3.21)	0.93 (0.14 – 1.73)
Older	2.54 (2.21 – 2.87)	1.53 (0.71 – 2.35)

Note. 95% confidence intervals are shown in parentheses.

Comparison of linear and nonlinear functions. Figure 3 shows the mean dAIC (sigmoidal – linear) for participants in each age group, for the accuracy and RT data. Both the accuracy and RT data indicate that the linear function provides a better fit to the data, for both young and older adults.

Effect of audiometric hearing status. Accuracy data from the older adults were entered into an ANOVA including target condition (count pure/FM), numerosity (1-6), and hearing status (normal/mild impairment). There was no main effect of hearing status, $F < 1$, but there was a significant interaction between numerosity and hearing status, $F(5, 140) = 3.14$, $p = .010$, $\eta^2_p = .10$. Older adults with mild hearing impairment were less accurate at smaller numerosities, leading to a significant difference between hearing groups at the first numerosity, $F(1, 28) = 4.70$, $p = .039$, $\eta^2_p = .14$, but not larger numerosities (all ps > .2).

When only participants with normal hearing were included in the Age group × Target type × Numerosity ANOVA for accuracy (see above), there was still no significant effect of age group, $F(1, 41) = 1.77$, $p = .191$, $\eta^2_p = .10$, but there was no longer a significant interaction between age group and numerosity, $F(5, 205) = 1.54$, $p = .179$, $\eta^2_p = .04$.

Summary

In Experiment 3, we found highly accurate enumeration of one or two FM tones when presented within a stream of pure tones, but no evidence for auditory subitizing. This suggests that separating auditory objects in time, rather than space, does not provide conditions compatible with auditory subitizing. We did however find that accurate enumeration of small numbers of objects was only possible when target tones could be clearly identified amid distractor tones (enumeration of FM tones amid pure tones, but not pure tones amid FM tones). This meshes with findings from visual enumeration studies (e.g., Trick & Pylyshyn, 2003) in which only targets that are individuated at preattentive levels of processing can be subitized.

Older adults were slower overall and had worse accuracy when enumerating small numbers of auditory objects. This was associated with poor audiometric hearing thresholds. There was no longer a difference in accuracy between young and older participants when hearing-impaired older adults were excluded.

General Discussion

We conducted three auditory enumeration studies designed to assess whether one of the fundamental mechanisms within the visual domain (subitizing) also generalized to the auditory domain. In doing so, we probed numerous aspects of auditory enumeration producing a number of key findings.

Auditory Subitizing is Limited to Approximately Two, Spatially-Separated Objects

Across all three experiments, approximately two auditory objects could be enumerated with the high accuracy that is typically associated with the subitizing mechanism. After this point, enumeration accuracy began to decline, indicating the operation of a more error-prone mechanism or set of processes. In contrast, the RT data from all experiments and conditions show linear slopes, consistent with a serial counting mechanism being engaged for all numerosities.

In order to provide *strong* evidence for separate subitizing and counting mechanisms in audition, it would be necessary to prove that a nonlinear function provides a better fit to both the accuracy and RT data than a linear function. This was not the case in Experiment 1, in which auditory objects were separated using ITDs, nor in Experiment 3 in which auditory objects were separated in time. In Experiment 2 we found that a nonlinear function provided the better fit to the accuracy data than a linear function; however, a linear function provided the better fit to the RT data.

Contrast Between Accuracy and RT Data

Visual subitizing is characterized by enumeration that is both fast and accurate, resulting in flatter enumeration functions within the subitizing range for both RTs and accuracy. In the present study, flatter subitizing functions were found for accuracy but not RTs. A similar dissociation arises in studies investigating haptic/tactile enumeration, where evidence for subitizing is mixed (Gallace, Tan, & Spence, 2008). Some studies do show a bilinear RT function, but the ‘flatter’ subitizing slopes are much steeper than those found in visual enumeration studies (Plaisier, Bergmann Tiest, & Kappers, 2009), and so are not entirely compatible with the notion of tags being assigned in parallel (or indeed rapidly). If we consider subitizing to require the rapid enumeration of items with high accuracy then our findings suggest that there is little if any evidence for the subitization of auditory stimuli. However, if we consider subitizing to reflect the ability to process small numbers of items in a different way to large numbers then there is some evidence that up to two auditory items can be subitized, at least in some relatively limited circumstances. Irrespective of the nuances in definitions, our work shows that at least in some circumstances, up to two auditory items can be perceived/tagged with high accuracy even if this is not achieved in a parallel manner.

That said, one clear difference between the current study and previous studies of visual enumeration is that the stimuli in our experiments varied over time. As noted above, linear RT functions could indicate that participants used a serial enumeration process for all numerosities (i.e., no evidence of subitizing). Alternatively, participants might have become more conservative as numerosity increased. That is, they might have rechecked or confirmed an initial (and rapid) estimate of numerosity more often when larger numbers of auditory objects were present. One possible way to determine this would be to present the auditory stimuli for a relatively short amount of time, thus limiting the possibility for re-checking and assessing performance purely on accuracy measures. Analogously, future work could ask participants to enumerate non-stationary visual stimuli.

Auditory Subitizing: Potential Mechanisms

An accuracy-based subitizing span of approximately two auditory objects would be consistent with that found in feature-based visual enumeration studies in which targets are defined by their color (Watson et al., 2005b). The visual feature-based subitizing span of around two visual objects is thought to reflect segregation of the visual scene into a foreground and background. In this case, it would be simple to enumerate the presence of a background only, or a background plus foreground, resulting in highly accurate performance. A similar mechanism could operate for auditory subitizing, in which the auditory scene is parsed into a target object plus background. However, the subitizing spans in Experiment 2 exceeded two auditory objects, suggesting some limited ability to further decompose the 'background' stream. Cusack et al.'s (2004) hierarchical decomposition model would support this hypothesis, proposing that participants are initially (preattentively) aware of broad categories of current sounds in the environment, and not just a target and background. However, any further decomposition of these broad categories of sounds would require focal attention, thereby limiting the number of auditory objects that can be subitized to around only two or three.

Spatial separation is critical to visual subitizing. In Experiments 1 and 2 we asked whether spatial separation also facilitates auditory subitizing. Experiment 1 revealed that lateralizing auditory objects to different locations using ITDs only improved counting accuracy for four or more objects, but did not improve accuracy when enumerating small numbers of auditory objects. Nor did it lead to nonlinear enumeration functions, in either the accuracy or RT data. In contrast, in Experiment 2 we found that presenting auditory objects from different locations using generic HRTFs improved accuracy for all numerosities, and the accuracy data were better fit by a nonlinear function.

Improved accuracy at all numerosities when sounds were lateralized using HRTFs rather than ITDs alone could be due to factors relating to auditory scene analysis. First, sounds in Experiment 2 were presented at greater eccentricities, and from fewer locations, than in Experiment 1 (-90, 0, and 90° azimuth, compared with 8 evenly-spaced horizontal lateralizations in Experiment 1). It is therefore possible that the increased spatial separation in Experiment 2 was responsible for the increased accuracy. Second, HRTFs include ILDs, and thus each signal is more strongly represented in the contralateral auditory cortex than in the ipsilateral auditory cortex. This allows auditory spatial attention to enhance the signal in the target auditory cortex, providing increased spatial attention benefits compared with when stimuli are lateralized using ITDs alone (Roberts et al., 2009). It is therefore likely that participants found it easier to direct their attention to the auditory objects when the sounds were lateralized using HRTFs compared with ITDs only. Third, spatially separating the stimuli using HRTFs could produce ‘spatial unmasking’, a process whereby target identification is improved when a target and distractor are spatially separated (Shinn-Cunningham, Schickler, Kopco, & Litovsky, 2001). A release from energetic masking is provided because the target to distractor ratio is improved at one ear. Spatial unmasking could potentially speed a serial enumeration process, by allowing each target to be identified more easily amid distractors.

Potentially, these mechanisms could also account for the change from a linear to nonlinear accuracy function. A further possibility relates to how the auditory system codes spatial location. Visual subitizing is achieved by determining the number of tags that are currently assigned to objects in the environment (Pylyshyn, 1989; Trick & Pylyshyn, 1994). In Experiment 2, we speculated that auditory subitizing could operate in a similar way by determining the number of spatial channels that were currently activated. This remains a potential explanation. However, there are methodological issues regarding the increased spatial separation in Experiment 2 compared with Experiment 1, and the presentation of more than one auditory object from each location in Experiment 2.

Future research could further investigate auditory tagging through use of a multiple object tracking task. If the accuracy data in Experiment 2 do indeed indicate that two or three auditory objects are tagged, then it should be possible to track two or three moving target auditory objects amid identical moving distractor objects. Although this proposed study would be methodologically challenging, it would provide an independent test of an auditory tagging mechanism.

Accurate (>50%) Auditory Enumeration is Limited to Three to Four Auditory Objects

Consistent with previous auditory enumeration studies (Kawashima & Sato, 2015; Weller et al., 2016; Zhong & Yost, 2017), we found that between three and four auditory objects could be enumerated with 50% accuracy. This was true when enumerating both spatially separated concurrent auditory objects in Experiments 1 and 2, and temporally separated sequential auditory objects in Experiment 3. Kawashima and Sato (2015) considered the possibility that their findings, with voices, might not generalize to other types of natural sounds. Here we find that the limit on accurate auditory enumeration holds for other types of auditory stimuli, including environmental sounds and pure/FM tones. Although in our study stimuli were presented for only 10 seconds, it does not seem likely that longer stimulus durations would result in increased numbers of stimuli being enumerated accurately. For example, Weller et al. (2016) presented stimuli for up to 45 seconds and still found that normally-hearing listeners could only accurately identify up to four auditory sources.

One possibility is that participants use alternative cues to numerosity (e.g., loudness) to determine the number of auditory objects that are present. This is also an issue in visual enumeration studies, where the density or overall luminance of the display contains useful cues to numerosity, and it is not always possible to dissociate cues associated with magnitude from those associated with numerosity. However, in the present study these magnitude cues are less reliable than in other studies. In Experiments 1 and 2 the auditory objects varied in intensity over time, making intensity an unreliable cue to numerosity. In Experiment 3, the same number of stimuli were presented on every trial, with the task being to enumerate targets amid distractors. This approach has also been used in visual studies to control the overall size of the display (e.g., Watson et al., 2005a).

Targets Must be Individuated Preattentively to be Accurately Enumerated

In visual enumeration studies, participants are unable to subitize visual objects in parallel if focused attention is required to separate target items from distractors (Trick & Pylyshyn, 1993). Analogously, in Experiment 3 we compared enumeration performance when participants enumerated pure tones amid distractor FM tones and FM tones amid distractor pure tones. The FM tones required less focal attention to be identified than the pure tones. We found that participants were able to enumerate FM tones presented among pure tone distractors (equivalent to enumerating preattentively available visual targets) but had lower accuracy and longer RTs for enumerating pure tones among FM distractors (equivalent to enumerating visual targets that require serial attention to detect). The gap between pure-tone and FM-tone enumeration accuracy was greatest for smaller numerosities. The pattern of results differs from that found in visual enumeration studies, in which being unable to identify the targets preattentively eliminates subitizing but participants are still able to identify a single target with high accuracy. Potentially, this difference between visual and auditory enumeration of targets amid distractors reflects the specific visual/auditory tasks and stimuli, or the change from enumeration of concurrent to sequential stimuli.

For the FM task, we did not find any evidence for an auditory subitizing mechanism – either based on accuracy or RTs – indicating that separating auditory objects in time, rather than space, is not sufficient to allow auditory subitizing to occur. One possibility is that participants perceived the rapid sequence of tones as a single stream, and therefore had difficulty enumerating target items within the stream. Previous studies (e.g., Taubman, 1950) suggest that the interval between temporally-separated auditory stimuli can be critical to participants' ability to enumerate those stimuli. In addition, the total duration of the auditory stream may affect enumeration performance, as streaming builds up over time (e.g., Moore & Gockel, 2012).

Auditory Enumeration is Only Minimally Affected by Healthy Aging

As previously found in visual enumeration studies (e.g., Watson et al., 2002), older adults were slower and less accurate in all three auditory enumeration tasks. Visual subitizing is typically unaffected by healthy aging, but here we asked whether poor auditory subitizing might partially account for difficulties that older adults report in multi-talker conversations (Gatehouse & Noble, 2004). In Experiment 1, older adults were slower and less accurate than young adults, but there was no interaction between age group and numerosity in either the accuracy or RT data, suggesting that older adults had a similar cost to young adults for each additional auditory object.

In Experiment 2, where we found evidence of subitizing, older adults had similar subitizing spans to young adults but had a larger drop in accuracy for each additional auditory object in the counting range (3 to 5 auditory objects). Older, but not young, participants showed a small additional benefit when stimuli were lateralized to three spatial locations, over and above the benefit when stimuli were lateralized to two spatial locations. This additional benefit affected enumeration at all numerosities (3-5) but did not influence the subitizing span when stimuli were presented from 3 rather than 2 locations. The additional benefit brought older adults’ accuracy closer to, but still below, the accuracy of young adults when enumerating spatially separated auditory objects.

In Experiment 3, older adults were slower than young adults and were less accurate, particularly with smaller numerosities. However, this was entirely accounted for by hearing loss in the older participants – only those participants with mild hearing impairment showed the reduced accuracy at smaller numerosities. An enumeration deficit for hearing-impaired older adults was also found by Weller et al. (2016). In Experiment 3 here, the deficit for older adults is attributable to perceptual loss rather than any age-related cognitive deficit, underlining the importance of accounting for perceptual deficits when assessing older adults’ cognitive ability (Allen & Roberts, 2016).

Conclusion

Across three experiments, participants could enumerate only two or three auditory objects with high accuracy. We found evidence consistent with different subitizing and counting mechanisms in only one experiment, when auditory objects were separated using generic HRTFs which contain ILDs as well as ITDs. Accuracy-based average estimated subitizing spans were between two and three, suggesting a subitizing limit that is noticeably smaller than that found with visual objects. Consistent with previous research, across the experiments we found that only up to between three and four auditory objects could be counted with accuracy greater than 50%. Older adults were slower and less accurate than young adults, but there was only limited evidence for an age-related decline in enumeration of auditory objects. We propose that any putative auditory subitizing mechanism is limited by the need for focal attention to decompose the auditory scene into its constituent auditory objects.

References

Allen, H. A., & Roberts, K. L. (2016). Editorial: Perception and cognition: Interactions in the aging brain. *Frontiers in Aging Neuroscience,* 8:130. doi: 10.3389/fnagi.2016.00130.

Baguley, T. (2012). *Serious stats: A guide to advanced statistics for the behavioural sciences.* Basingstoke, UK: Palgrave Macmillan.

Ben-David, B. M., Tse, V. Y. Y., & Schneider, B. A. (2012). Does it take older adults longer than younger adults to perceptually segregate a speech target from a background masker? *Hearing Research, 290,* 55-63.

Botte, M.-C., Drake, C., Brochard, R., & McAdams, S. (1997). Preliminary measures of the focusing of attention on auditory streams. *Perception & Psychophysics, 59,* 419-425.

Bregman, A.S. (1990). *Auditory scene analysis*. MIT Press: Cambridge, MA.

Briley, P. M., Gorman, A. M., & Summerfield, A. Q. (2016). Physiological evidence for a midline spatial channel in human auditory cortex. *Journal of the Association for Research in Otolaryngology, 17,* 331-340.

Brochard, R., Drake, C., Botte, M.-C., & McAdams, S. (1999). Perceptual organization of complex auditory sequences: Effect of number of simultaneous subsequences and frequency separation. *Journal of Experimental Psychology: Human Perception and Performance, 25,* 1742-1759.

Camos, V., & Tillmann, B. (2008). Discontinuity in the enumeration of sequentially presented auditory and visual stimuli. *Cognition, 107*, 1135-1143.

Carlyon, R. P., Cusack, R., Foxton, J. M., & Robertson, I. A. (2001). Effects of attention and unilateral neglect on auditory stream segregation. *Journal of Experimental Psychology: Human Perception and Performance, 27,* 115-127.

Cusack, R., & Carlyon, R. P. (2003). Perceptual asymmetries in audition. *Journal of Experimental Psychology: Human Perception and Performance, 29,* 713-725.

Cusack, R., Deeks, J., Aikman, G., & Carlyon, R. P. (2004). Effects of location, frequency region, and time course of selective attention on auditory scene analysis. *Journal of Experimental Psychology: Human Perception and Performance, 30,* 643-656.

Demeyere, N., Lestou, V., & Humphreys, G. W. (2010). Neuropsychological evidence for a dissociation in counting and subitizing. *Neurocase, 16,* 219-237.

Demeyere, N., Rotshtein, P., & Humphreys, G. W. (2014). Common and dissociated mechanisms for estimating large and small dot arrays: Value-specific fMRI adaptation. *Human Brain Mapping, 35,* 3988-4001.

Dowling, W. J., Lung, K., & Herrbold, S. (1987). Aiming attention in pitch and time in the perception of interleaved melodies. *Perception & Psychophysics, 41,* 642-656.

Eramudugolla, R., Irvine, D. R. F., McAnally, K. I., Martin, R. L., & Mattingley, J. B. (2005). Directed attention eliminates ‘change deafness’ in complex auditory scenes. *Current Biology, 15,* 1108-1113.

Eramudugolla, R., McAnally, K. I., Martin, R. L., Irvine, D. R. F., & Mattingley, J. B. (2008). The role of spatial location in auditory search. *Hearing Research, 238,* 139-146.

Ezzatian, P., Li, L., Pichora-Fuller, K., & Schneider, B. A. (2015). Delayed stream segregation in older adults: More than just informational masking. *Ear and Hearing, 36,* 482-484.

Freyman, R. L., Balakrishnan, U., & Helfer, K. S. (2001). Spatial release from informational masking in speech recognition. *Journal of the Acoustical Society of America, 109,* 2112-2122.

Fründ, I., Haenel, N. V., & Wichmann, F. A. (2011). Inference for psychometric functions in the presence of nonstationary behavior. *Journal of Vision,* 11:16. doi: 10.1167/11.6.16.

Gallace, A., Tan, H. Z., & Spence, C. (2008). Can tactile stimuli be subitised? An unresolved controversy within the literature on numerosity judgments. *Perception, 37,* 782-800.

Gardner, B., & Martin, K. (1994). HRTF measurements of a KEMAR dummy-head microphone. Retrieved on April 2, 2015 at http://sound.media.mit.edu/resources/KEMAR.html

Gatehouse, S., & Noble, W. (2004). The Speech, Spatial and Qualities of Hearing Scale (SSQ). *International Journal of Audiology, 43,* 85-99.

Griffiths, T. D., & Warren, J. D. (2004). What is an auditory object? *Nature Reviews Neuroscience, 5,* 887-892.

Hawley, M. L., Litovsky, R. Y., & Culling, J. F. (2004). The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer. *Journal of the Acoustical Society of America, 115,* 833-843.

Jevons, W. S. (1871). The power of numerical discrimination. *Nature, 3,* 281-282.

Kaufman, E. L., Lord, M. W., Reese, T. W., & Volkmann, J. (1949). The discrimination of visual number. *The American Journal of Psychology, 62,* 498-525.

Kawashima, T., & Sato, T. (2015). Perceptual limits in a simulated ‘Cocktail party’. *Attention, Perception, & Psychophysics, 77,* 2108-2120.

Kubovy, M., & van Valkenburg, D. (2001). Auditory and visual objects. *Cognition, 80,* 97-126.

McLachlan, N. M., Marco, D. J. T., & Wilson, S. J. (2012). Pitch enumeration: Failure to subitize in audition. *PLoS ONE,* 7(4): e33661. doi:10.1371/journal.pone.0033661.

Macken, W. J., Tremblay, S., Houghton, R. J., Nicholls, A. P., & Jones, D. M. (2003). Does auditory streaming require attention? Evidence from attentional selectivity in short-term memory. *Journal of Experimental Psychology: Human Perception and Performance, 29*, 43-51.

Mandler, G., & Shebo, B. J. (1982). Subitizing: An analysis of its component processes. *Journal of Experimental Psychology: General, 111,* 1-22.

Moore, B. C. J., & Gockel, H. E. (2012). Properties of auditory stream formation. *Philosophical Transactions of the Royal Society B – Biological Sciences, 367*, 919-931.

Olivers, C. N. L., & Watson, D. G. (2008). Subitizing requires attention. *Visual Cognition, 16*, 439-462.

Plaisier, M. A., Bergmann Tiest, W. M., & Kappers, A. M. L. (2009). One, two, three, many – Subitizing in active touch. *Acta Psychologica, 131*, 163-170.

Pylyshyn, Z. (1989). The role of location indexes in spatial perception: A sketch of the FINST spatial-index model. *Cognition, 32,* 65-97.

Pylyshyn, Z. W. (2001). Visual indexes, preconceptual objects, and situated vision. *Cognition, 80,* 127-158.

Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision, 3,* 1-19.

Repp, B. H. (2007). Perceiving the numerosity of rapidly occurring auditory events in metrical and nonmetrical contexts. *Perception & Psychophysics, 69,* 529-543.

Roberts, K. L., & Allen, H. A. (2016). Perception and cognition in the ageing brain: A brief review of the short- and long-term links between perceptual and cognitive decline. *Frontiers in Aging Neuroscience,* 8:39. doi: 10.3389/fnagi.2016.00039.

Roberts, K. L., Summerfield, A. Q., & Hall, D. A. (2009). Covert auditory spatial orienting: An evaluation of the spatial relevance hypothesis. *Journal of Experimental Psychology: Human Perception and Performance, 35,* 1178-1191.

Saltzman, I., & Garner, W. (1948). Reaction time as a measure of the span of attention. *Journal of Psychology*, 25, 227-241.

Schneider, B. A., Daneman, M., & Pichora-Fuller, M. K. (2002). Listening in aging adults: From discourse comprehension to psychoacoustics. *Canadian Journal of Experimental Psychology, 56*, 139-152.

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in Cognitive Sciences, 12,* 182-186.

Shinn-Cunningham, B. G., Schickler, J., Kopco, N., & Litovsky, R. (2001). Spatial unmasking of nearby speech source in a simulated anechoic environment. *Journal of the Acoustical Society of America, 110,* 1118-1129.

Sussman, E. S., Horváth, J., Winkler, I., & Orr, M. (2007). The role of attention in the formation of auditory streams. *Perception & Psychophysics, 69,* 136-152.

Taubman, R. E. (1950). Studies in judged number: I. The judgment of auditory number. *Journal of General Psychology, 43*, 167-194.

ten Hoopen, G., & Vos, J. (1979). Effect on numerosity judgment of grouping of tones by auditory channels. *Perception & Psychophysics, 26,* 374-380.

Thurlow, W. R., & Rawlings, I. L. (1959). Discrimination of number of simultaneously sounding tones. *Journal of the Acoustical Society of America, 31,* 1332-1336.

Treisman, A., & Souther, J. (1985). Search asymmetry: A diagnostic for preattentive processing of separable features. *Journal of Experimental Psychology: General, 114,* 285-310.

Trick, L. M., & Pylyshyn, Z. W. (1993). What enumeration studies can show us about spatial attention: Evidence for limited capacity preattentive processing. *Journal of Experimental Psychology: Human Perception and Performance, 19,* 331-351.

Trick, L. M., & Pylyshyn, Z. W. (1994). Why are small and large numbers enumerated differently? A limited-capacity preattentive stage in vision. *Psychological Review, 101,* 80-102.

Vitevitch, M. S., & Siew, C. S. Q. (2016). Estimating group size from human speech: Three's a conversation but four's a crowd. *Quarterly Journal of Experimental Psychology, 70,* 1-35. DOI: 10.1080/17470218.2015.1122070

Watson, D. G., & Humphreys, G. W. (1999). The magic number four and temporo-parietal damage: Neurological impairments in counting targets amongst distractors. *Cognitive Neuropsychology, 16,* 609-629.

Watson, D. G., Maylor, E. A., Allen, G. E. J., & Bruce, L. A. M. (2007). Early visual tagging: Effects of target-distractor similarity and old age on search, subitization, and counting. *Journal of Experimental Psychology: Human Perception and Performance, 33,* 549-569.

Watson, D. G., Maylor, E. A., & Bruce, L. A. M. (2005a). Effects of age on searching for and enumerating targets that cannot be detected efficiently. *Quarterly Journal of Experimental Psychology, 58A,* 1119-1143.

Watson, D. G., Maylor, E. A., & Bruce, L. A. M. (2005b). The efficiency of feature-based subitization and counting. *Journal of Experimental Psychology: Human Perception and Performance, 31,* 1449-1462.

Watson, D. G., Maylor, E. A., & Manson, N. J. (2002). Aging and enumeration: A selective deficit for the subitization of targets among distractors. *Psychology and Aging, 17,* 496-504.

Weller, T., Best, V., Buchholz, J. M., & Young, T. (2016). A method for assessing auditory spatial analysis in reverberant multitalker environments. *Journal of the American Academy of Audiology*, *27*, 601-611.

Zhong, X., & Yost, W. A. (2017). How many images are in an auditory scene? *Journal of the Acoustical Society of America*, *141*, 2882-2892.