Estimating working memory capacity for lists of nonverbal sounds

Dawei Li · Nelson Cowan · J. Scott Saults

Abstract Working memory (WM) capacity limit has been extensively studied in the domains of visual and verbal stimuli. Previous studies have suggested a fixed WM capacity of typically about three or four items, on the basis of the number of items in working memory reaching a plateau after several items as the set size increases. However, the fixed WM capacity estimate appears to rely on categorical information in the stimulus set (Olsson & Poom Proceedings of the National Academy of Sciences 102:8776-8780, 2005). We designed a series of experiments to investigate nonverbal auditory WM capacity and its dependence on categorical information. Experiments 1 and 2 used simple tones and revealed capacity limit of up to two tones following a 6-s retention interval. Importantly, performance was significantly higher at set sizes 2, 3, and 4 when the frequency difference between target and test tones was relatively large. In Experiment 3, we added categorical information to the simple tones, and the effect of tone change magnitude decreased. Maximal capacity for each individual was just over three sounds, in the range of typical visual procedures. We propose that two types of information, categorical and detailed acoustic information, are kept in WM and that categorical information is critical for high WM performance.

Keywords Working memory · Audition · Short term memory

Working memory (WM) refers to the cognitive process that involves maintenance and manipulation of a limited amount of information for a short period, usually a few seconds (Baddeley & Hitch, 1974; Cowan, 1995). WM is fundamental to a number of higher-order cognitive functions, such as decision making, language processing, and planning (Cowan, 2005). For decades, researchers have been curious about the limit of WM capacity—how much information can be stored in WM. Although WM has been investigated in great detail using visual and verbal stimuli, WM for tones has received far less attention and will be examined here. We will consider factors that affect performance in other domains in order to assess the capacity of a key attention-related component of WM for nonverbal sounds, uncorrupted by various mnemonic strategies.

Guidance from research on WM in other domains

WM in other domains provides an important context in which to formulate the manner in which to examine WM for tones. On the basis of various empirical and experimental evidence, Miller (1956) proposed that people could keep, in what is now called WM, lists of approximately seven items, plus or minus two. Miller's work elicited many subsequent studies on humans' WM capacity. Some studies, however, indicated that Miller might have overestimated WM capacity (Cowan, 2001; Luck & Vogel, 1997; Sperling, 1960). In one experiment in Sperling's seminal work, participants were briefly presented an array of 12 characters and were instructed to write down the characters that they could remember after the array had disappeared. The results showed that only about 4 characters could be written down, meaning that WM capacity might be more restricted than that estimated by Miller.

Chunking

One reason for the higher estimate obtained by Miller (1956) is that, as he himself pointed out, people sometimes can group

D. Li · N. Cowan (✉) · J. S. Saults
Department of Psychological Sciences, University of Missouri,
18 McAlester Hall,
Columbia, MO 65211, USA
e-mail: CowanN@missouri.edu

several items from a list into a larger meaningful unit, or chunk, and remember the chunk instead of the individual items. In a straightforward illustration, although people usually cannot remember nine random letters, such as *NGJLXISFH*, they can easily remember *IRSFBICIA*, if they are able to chunk these letters into three U.S. government agencies—*IRS*, *FBI*, and *CIA*. In Sperling's experiments, due to the rapid and concurrent presentation of items, one can assume that it was difficult to apply chunking, leading to a smaller estimate of WM capacity. Similar results have been obtained with the recognition of nonverbal items (e.g., Luck & Vogel, 1997). Musical knowledge may allow chunking for musical sounds, which we discouraged by selecting stimuli judiciously.

Rehearsal

Another factor affecting WM capacity estimates is the strategy of rehearsal, or covertly repeating items or labels in WM in order to refresh the representations of the items in WM. Some studies have shown that phonologically similar words, such as *cat*, *bat*, and *mat*, or letters, such as *P*, *V*, *D*, *B*, and so forth, were more difficult to remember than phonologically dissimilar words or letters, a phonological similarity effect, even when the items were visually presented (Conrad, 1964; Conrad & Hull, 1964) and that people could memorize fewer words with longer length than words with shorter length, a word length effect (Baddeley, Thomson, & Buchanan, 1975). According to some theories, these effects reflect the use of rehearsal in verbal WM,[1] which is more error prone when the words are phonologically similar and takes longer when the words are longer. Thus, when participants are required to repeat a simple word, such as "the," while remembering the items (an *articulatory suppression* task), both the phonological similarity effect and the word length effect are greatly diminished or disappear entirely (for a review, see Baddeley, 1986). Because humming might be used to rehearse nonverbal sounds (e.g., Hickock, Buchsbaum, Humphries, & Muftuler, 2003), we used suppression to prevent that possibility (cf. Schendel & Palmer, 2007).

Sensory memory

Another factor that can enhance the estimate of WM is memory for the physical properties of the stimuli—that is,

exactly how they look or sound. Sperling (1960) showed that exposure to the character array led to sensory memory of most of the array items for a short period (under 1 s) and that this sensory memory was available for recall if a partial report cue was provided so that only one row of up to four items had to be recalled on a particular trial. Similar indications of a vivid but short-lived sensory memory for a complex array are obtained in the auditory modality, with auditory sensory memory lasting several seconds (e.g., Darwin, Turvey, & Crowder, 1972). In some experiments, items to be recalled are followed by an interfering item in the same modality in order to overwrite sensory memory, making it necessary that the concepts, rather than sensations, be recalled (e.g., Saults & Cowan, 2007). We adopted that strategy here for nonverbal sounds.

Core WM capacity

Cowan (2001) suggested that the smaller limit of three to five items in WM is obtained under conditions in which the items retained are chunks (meaningful units) based on already-known information. This is presumably the case when the items cannot be further grouped into larger meaningful units at the time when they are presented in the to-be-remembered materials, cannot be rehearsed verbally, and cannot be retained in a sensory form. To ensure that these conditions apply, known patterns across stimuli can be avoided, articulation can be suppressed, and the items to be retained can be followed by a sensory mask.

A key example upon which the present work is based is the recognition memory for colored squares, examined by Luck and Vogel (1997). In one experiment, they instructed the participants to memorize a briefly presented array of a few colored squares for several seconds, followed by the presentation of a second, probe array in which one square may have changed color; in another experiment yielding similar results, one item in the second array was marked to indicate which square might have changed. The task was to decide whether the new square had the same color as the previous square in that location. With this change detection paradigm, they estimated the participants' visual WM capacity at about four items. The brief presentation of the first array made the items difficult to chunk, and a secondary memory load of two digits further discouraged rehearsal. Sensory memory presumably could not be used to great advantage either, inasmuch as the probe array would have overwritten the critical sensory information before a judgment could be made. Given these restrictions, it is suggested that the results are indicative of a core WM capacity (Cowan, 2001).

This core WM capacity was observed also when participants were taught pairs of words, in which case participants recalled about three chunks from a list in the presence of

---

[1] Some studies have shown evidence for a visual similarity effect for visually presented verbal stimuli and have shown that the visual similarity effect is not affected by articulatory suppression (Logie, Della Sala, Wynn, & Baddeley, 2000). Given that we used acoustically presented nonverbal sounds and that our visual symbols carried no specific information except for the recall signal, with the other visual symbols just marking the presence of a tone in the list and not of discriminatory value for tone recognition, visual similarity is unlikely to be an important factor in this study.

articulatory suppression, no matter whether the chunks in the list were singletons or learned pairs (Chen & Cowan, 2009). Given the considerable evidence for a small core capacity for information from stimuli that can be labeled, we wished to examine WM for tonal stimuli that could not easily be labeled.

Cowan (2001) proposed a measure that can be used to estimate the number of items held in WM. This measure applies to the experimental situation in which the test probe display clearly indicates which item changed, if any of them did (Rouder, Morey, Morey, & Cowan, 2011). It assumes that the array includes N items and that k items fit in WM. Then when $N \geq k$, the proportion of correct detections of a change, or *hits*, can be estimated as hits $= k/N + (1 − k/N)g$, where g is the rate of guessing that there has been a change, in the absence of WM information. Guessing takes place only if the tested item was not in WM, so when there is no change, false alarms $= (1 − k/N)g$. Combining these equations yields the estimate $k = (\text{hits} − \text{false alarms})N$. We apply this formula to recognition memory for lists of tones.

WM and categorical information

Although a number of studies directly support the theory of core WM capacity, there are a few alternative theories and experiment results. Some studies found lower capacity estimates in WM tasks. For example, Alvarez and Cavanagh (2004) found that participants remembered fewer items when complexity of the stimulus set increased. However, Awh and colleges suggested that WM capacity was limited by sample–test similarity, instead of stimulus complexity (Awh, Barton, & Vogel, 2007). They presented the same complex objects as those used in Alverez and Cavanagh but manipulated the test object to have either high or low similarity with the sample object. They found that when sample–test similarity was low, capacity for complex objects was identical to capacity for simple objects. These results indicated that stimulus complexity is unlikely to be the main factor influencing WM capacity, although it affects the conditions under which the representations in WM will be adequate for a comparison with the test stimulus.

Another line of studies used a different approach to study WM capacity. These studies used various modified change detection tasks that measure the precision of a memory trace. Zhang and Luck (2008) presented participants an array of colored squares. In the test, participants were presented with a continuous color wheel and were required to choose a color that matched the color of a target square. Zhang and Luck (2008) manipulated the number of squares in the sample array, and they found that memory capacity was high at set size 1, 2, and 3 but dropped drastically from set size 3–6, whereas the precision of participants' response decreased from set size 1–3 but stayed unchanged from set size 3–6. In support of the core WM capacity theory, these results indicated that when set size was high, participants devoted their mental resources to only a subset of stimuli. However, using similar paradigms, some studies have concluded that WM capacity is constrained by a limited pool of continuous mental resources that can be spread among any number of items, rather than a fixed number of slots (e.g., Bays & Husain, 2008). In Bays and Husain's study, participants were briefly presented an array of colored squares or colored arrows, and after a short delay, they were required to judge whether a new square or arrow had the same spatial location or orientation as the previous square or arrow with the same color. The amount of location and orientation displacement was varied across trials. Results showed that memory precision decreased when the array size increased, even for the smallest set sizes. Bays and Husain also showed that WM resources could be allocated flexibly by manipulating eye movements during encoding. They proposed that WM resources could be distributed sparsely to all objects, instead of to a fixed number of objects. The debate between slot-based and resource-based models of WM capacity has yet to be resolved to everyone's satisfaction (e.g., Anderson, Vogel, & Awh, 2011; Bays, Catalau & Husain 2009; Zhang & Luck, 2011), although we strongly favor the fixed slots view; for perhaps the most compelling rebuttals of the continuous resource view to date, see Anderson and Awh (2012) and Thiele, Pratte, and Rouder (2011).

Olsson and Poom (2005) manipulated the amount of categorical information in the stimulus sets. They found that visual WM capacity was as low as only one in a stimulus set with little categorical information, even if the visual items were easy to distinguish. When categorical information, such as discrete colors and shapes, was added to the stimulus set, the estimated visual WM capacity increased to slightly below three. The authors concluded that categorical information stored in long-term memory was crucial to visual WM performance. Considering this study along with Zhang and Luck (2008), it is possible that the core WM capacity must make use of categorical information in the stimulus set. When the set size is small (e.g., less than three items), participants are able to retain categorical information of all the stimuli, as well as some information about stimulus details. When the set size increases to a certain point (e.g., more than four items), participants may have to devote all their mental resources to the categorical information of only a subset of the stimuli, leaving less capability for storing object details. Categorical information is also likely the information stored in the focus of attention and is less susceptible to decay or interference than is information about details (Saults & Cowan, 2007). Therefore, if a stimulus set contains little categorical information, low WM capacity is expected, especially after a long delay.

Indeed, a few studies have suggested the role of categorical information in the auditory domain (Fujisaki & Kawashima, 1970; Nairne, 1990; Surprenant & Neath, 1996). Nairne proposed a feature model that consists of two features of memory trace: modality dependent and modality independent. Modality-dependent features refer to physical properties that are modality specific, whereas modality-independent features refer to categorized information that does not depend on a specific modality (Nairne, 1990). Consequently, if a stimulus set is difficult to categorize, only modality-dependent features can be used in a memory task, leading to low performance.

In this study, we manipulated the amount of categorical information in our stimulus sets to study its influence on WM capacity for nonverbal sounds.

Prior research on WM for tones

In contrast to the extensively investigated domains of visual and verbal WM, few studies have investigated the capacity limit of nonverbal auditory items in WM that are uncontaminated in that they contain little verbal or verbalizable information, are difficult to visualize, and contain no familiar structure. It is possible that such auditory items could be more difficult to remember due to their acoustic nature, which can be retained in WM only through their sound properties, instead of phonological, visual, or semantic properties.

Studies of music sequence production suggest the use of two levels of structure: the scales of discrete pitch relationships, or intervals (such as the 12-tone chromatic scale of equal temperament in Western music), and the seven-interval subsets of the chromatic scale called diatonic scales (Burns & Ward, 1999; Davies, 1979). People can use familiarity with these scales to encode the melodic contour of a musical sequence, grouping or chunking intervals to achieve better memory of musical sequences, as compared with random tone sequences (Dewar, Cuddy, & Mewhort, 1977; Idson & Massaro, 1976). In our stimuli, we avoided these familiar structures in order to assess WM capability without chunking. Even within structured stimuli, however, there is some evidence of a core capacity limit. In particular, in music sequence production, pitch-ordering errors (musical sequences reproduced with tones in the wrong order) suggest a WM capacity limitation: There is typically confusion between tones no more than three to four tones apart in the intended sequence (Drake & Palmer, 2000; Palmer, 2005; Palmer & Pfordresher, 2003).

Capacity for lists of tones

There has been little research in which the number of tones in a sequence has been varied in order to assess the effect of

that manipulation on the ability to detect a change in one tone. Some studies studied serial recall of nonverbal auditory stimuli that were presented in different spatial locations (Lehnert & Zimmer, 2006; Parmentier & Jones, 2000). Lehnert and Zimmer asked participants to remember arrays of visual, auditory, or mixed visual and auditory stimuli at different spatial locations. The visual and auditory stimuli were from the same objects (e.g., an image of an airplane and sound produced by an airplane). The results showed that the hit rate was approximately .66, .57, and .52 at set size 4, 6, and 8, respectively, for auditory arrays and .90, .79, and .67 at set size 4, 6, and 8, respectively, for visual arrays. Importantly, even in the mixed array condition, performance was significant lower for auditory than for visual items. These results suggested lower capacity for spatial auditory than for spatial visual information. However, it is likely that binding of spatial and auditory features is more difficult than binding of spatial and visual features. Some studies also showed that sound localization was impaired during a spatial WM task, but not phonological WM tasks, indicating that spatial sound localization involves more spatial memory than does phonological memory (Merat & Groeger, 2003). Therefore, although the studies on auditory spatial WM provide insightful results for the organization of representations in WM, they are not direct measures of nonverbal auditory WM capacity, because of the involvement of spatial locations.

Watson, Foyle, and Kidd (1990) varied the number of component tones widely. They chose tones in a manner that eliminated conventional musical cues, dividing the frequency range 300–3 kHz into N tones based on logarithmically equal intervals, where N was the list length, and shuffling the order of the resulting tones. Clearly, the number of tones made a very large difference for performance, although no estimate of the WM capacity for tones could be obtained from their procedure. Note that, using this method, the number of tones in the list is confounded with the frequency difference between adjacent tones.

Kidd and Watson (1992) found that what was important was not the number of tones per se but the proportion of the tone list taken up by the target tone (proportion of the total duration [PTD]). In their procedure, however, participants were held responsible for only one tone per series, the one in the middle of the pattern (or in one experiment, two tones flanking the middle tone and changing together), which would not place a load on WM commensurate with the list length. Surprenant (2001) replicated this finding and further showed that both PTD and relative distinctiveness account for memory effects in three tone sequence recognition tasks. These findings indicate that besides list length, stimulus properties are important factors affecting auditory WM performance, similar to the case of visual WM (Alvarez & Cavanagh 2004; Awh et al., 2007).

In the closest precursor to the present study that we could find, Prosser (1995) chose 14 tones that were selected to avoid a musical scale and presented lists of 2, 4, or 6 randomly selected tones per trial. The list was followed by a tone probe to be judged present or absent from the sequence. To evaluate the results, we apply the formula of Cowan (2001) to the means shown in Prosser's Fig. 1. Doing so, using data for a short (1-s) retention interval, for lists of 2, 4, and 6 tones yields estimates of $k = 1.5$, 2.2, and 2.9 tones in WM, respectively. These estimates are roughly consistent with past evidence on nontonal stimuli or are slightly lower. The shift across list lengths is found also for visual arrays and may occur because certain individuals have a capacity lower than set size N, resulting in ceiling effects that limit the estimates for the smaller set sizes. However, due to the short retention interval and lack of a mask sound, these estimates were likely affected by sensory memory. Indeed, capacity estimates fell to 1.5, 1.7, and 1.7 items for set sizes of 2, 4, and 6, respectively, when Prosser used a 7-s retention interval.

Some previous studies investigated the effect of perceptual organization of auditory sequences on memory performance (Deutsch, 1970; Jones, Macken, & Harries, 1997; Warren & Obusek, 1972). Warren and Obusek found that participants were unable to report serial order of auditory sequence with three or four sounds, including but not limited to tones, when the duration of a sound was 200 ms. They also found that for proper serial order identification, stimulus duration should be at least 670 ms for oral response and 300 ms for card-ordering response. In this study, we used a relatively long stimulus presentation (500 ms) to avoid the limitations found with short stimulus durations.

Capacity, attention, and time

Cowan (2001) suggested that a limited number of items making up the core contents of WM is held in the focus of attention. The primary function of holding information that way would be to make the item representations resistant to interference or decay. In that regard, it is useful to examine the items in WM after a several-second retention interval, so that features and items susceptible to decay already would have decayed, and what remains is the items held firmly in mind. Cowan et al. (2011) presented a combination of colored squares and spoken letters, followed by a mask and then an 8-s retention interval, and after that period still observed a capacity of 2.9–3.6 items.

Capacity might be lower, however, for tonal stimuli that do not correspond to known musical categories. Prosser (1995) included a 7-s retention interval and, for lists of two, four, and six tones, we estimate from his Fig. 1 that $k = 1.5$, 1.7, and 1.7 items, respectively.

The present study

We wished to explore further these rough estimates of tones in WM, derived from the findings of Prosser (1995) at a long retention interval, more systematically

Fig. 1 Illustration of the procedure in Experiment 1. See the text for details
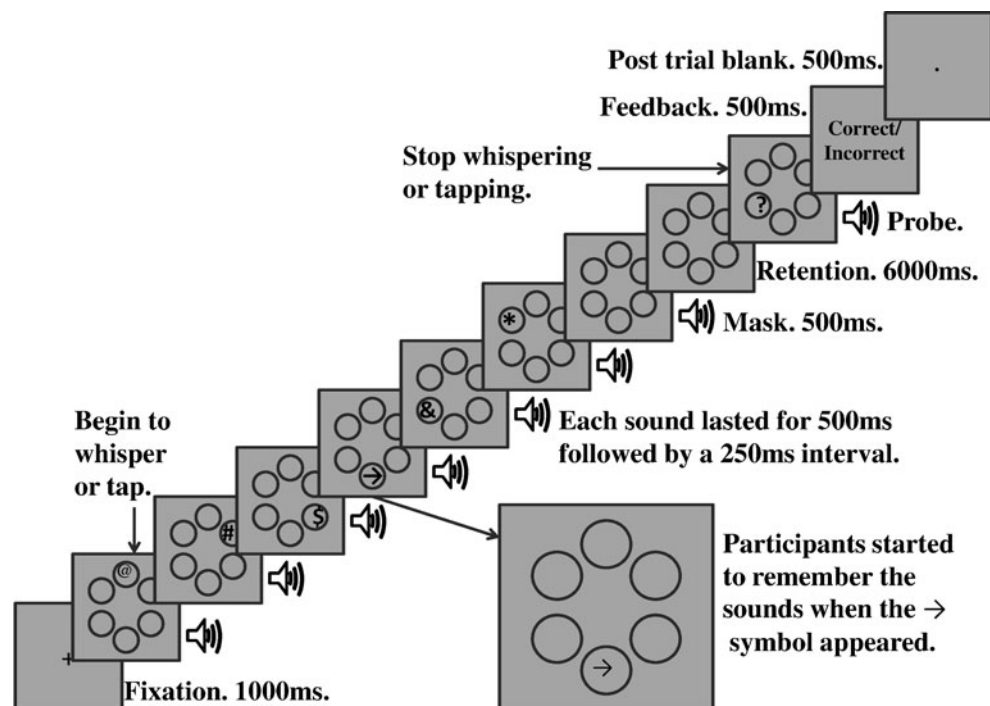
in order to understand WM capacity limits. We adapted the change detection procedure by presenting sequences of tones, followed by a probe tone or probe tone list to be recognized as the same as the original list or as changed (see Fig. 1). To identify the WM capacity limit for individual tones without any familiar musical structure, we used lists of tones randomly selected from a nonmusical scale of 12 pitches that differ from the notes of the chromatic scale and span several octaves. We used a retention interval of 6 s (following a list-final masking stimulus), which is long enough that any residual sensory memory that somehow survived the mask should already have decayed before the probe (see Darwin et al., 1972), leaving behind information that resists decay.

Several features distinguish our study from the past work of Prosser (1995) or any other study, to our knowledge. First, as one step to eliminate sensory memory information, we presented a masking sound after each list. There is a long history of auditory backward masking of recognition using interstimulus target–mask intervals of a fraction of a second (e.g., Massaro, 1975), but our purpose here was not to prevent recognition. Rather, similar to Saults and Cowan (2007), we waited long enough for recognition of all tones in the list to be completed and then presented a mask, in order to force participants to rely on the recognized abstract information in WM, rather than on a sensory memory trace, which otherwise might have persisted for several seconds (Cowan, 1984; Darwin et al., 1972).

Second, unlike most prior studies, on half of the trials, we suppressed articulation in case participants were able to vocalize tones covertly and rely on that process as subvocal rehearsal. To equate attention demand, we asked participants to tap their right index finger on the desk on the other half of the trials (Ricker, Cowan, & Morey, 2010).

Third, to equate the amount of intertone interference in memory (Cowan et al., 2005), we included conditions in which the number of tones stayed the same across different memory loads, which was accomplished by presenting six tones and requiring memorization starting at a variable point in the middle of the list (Fig. 1).

Fourth, and finally, we provided visual cues to indicate which serial position in the tone series was being probed. We did this because it is required for the k measure of items in working memory, which is based on the assumption that the participant needs to compare the probe with the memory of only one item. This measure of items in working memory has been psychometrically validated much more fully than any other measure; the data conform to a receiver operating characteristic function expected according to the model (Rouder et al., 2008; Rouder et al., 2011).

All of these precautions, taken together, should allow us to examine WM capacity for abstract information about tones without any prelearned categories for the tones.

Experiment 1

Method

Participants

Twenty-seven undergraduate University of Missouri students (12 male, 15 female) participated in the experiment to fulfill introductory psychology course requirements. In both Experiment 1 and the following two experiments, we included only individuals without special music training, defined as participation in a band or an orchestra or music instruction at a college level.

Apparatus and stimuli

The stimuli were presented with E-Prime (Schneider, Eschman, & Zuccolotto, 2002) in soundproof booths, using loudspeakers. Twelve simple tones (sine waves) were generated by Praat software (Boersma & Weenink, 2009), with a lowest frequency of 200 Hz and a highest frequency of 3900 Hz. There was a 31 % frequency difference between each two adjacent tones. Each tone had a duration of 500 ms and included 25-ms linear onset and offset ramps.

We wanted the pitches of our 12 tones to be as far apart as possible, so they would be easy to discriminate, but still within a range with similar difference limens for frequency change, which increases sharply beyond 4000 Hz (Sek & Moore, 1995). We also wanted them to differ from familiar musical notes. Thus, our lowest tone was about 35 cents above the G below middle C (G3), while our highest tone was about 23 cents below B7, the second highest note on an 88-key piano (100 cents = 1 semitone). A 31 % difference between tones avoids familiar musical intervals and harmonic relationships between tones. Adjacent semitones in music differ by about 5.9 % (precisely $2^{1/12}$) in 12-tone equal temperament, the common tuning system for Western music (Burns & Ward, 1999). Although our stimuli spanned about four octaves, no tone in our set had a simple harmonic relationship with another tone. For example, the second harmonic of 200 Hz is 800 Hz, but the closest frequency to that in our set was 771.6 Hz. Avoiding octaves minimizes the tendency to confuse two tones with different pitch height but equal chroma on the basis of octave generalization (Shepard, 1982).

Six circles were presented in the center of the screen on a gray background, as shown in Fig. 1. The participants were seated approximately 50 cm from the screen. The sounds were presented through two speakers (left and right) in front of the participants, with intensities between 60 and 70 dB(A) as measured by a sound level meter.

Procedure

On each trial, participants had to try to remember 2, 3, 4, 5, or 6 tones and then perform a recognition task. At the beginning of each trial, a "+"appeared at the center of the screen for 1,000 ms, which indicated the onset of a trial and provided a fixation point for the participant. Next, six circles were presented at the center of the screen, as shown in Fig. 1. Six tones, randomly selected without replacement from the set of 12, were sequentially presented at a rate of one item every 750 ms. A printed character (*, &, $, @, #, %, or →) accompanied each tone, and the characters were presented sequentially, with each character in one of the circles, always starting from the circle at the top. The character disappeared as soon as its corresponding tone ended. The participants were instructed to start remembering tones, starting with the one accompanied by a forward arrow (→) and continuing until the end of the series. They were also instructed to ignore the characters except for the forward arrow (→) to minimize any additional processing load (Lavie, 1995). The position of the forward arrow (→) was manipulated such that the memory load was set to include five levels: 2, 3, 4, 5, and 6 tones. The other characters were randomly arranged, and there was no constant association between particular characters and particular tones.

Two additional types of trials were included in the experiment and were the same as those in the other conditions, except that the participants heard only two or four tones and saw two or four characters, respectively, during the encoding phase. The characters were presented sequentially, each in one circle, starting from the circle on the top, and the first character was always a forward arrow (→). We included these additional conditions to estimate to what extent the different stimulus presentation methods would affect the participants' performance. In the following text, we will denote these trials as "presentation method 2" (PM2) and the other trials as "presentation method 1" (PM1).

A masking tone, which was produced by the simultaneous combination of all 12 stimulus tones, was presented for 500 ms after the last 1 of the 6 tones, in the same temporal rhythm as these tones, to eliminate sensory memory. After a 6,000-ms retention interval, a probe tone was presented, accompanied by a "?" symbol in one of the circles corresponding to a tone that was to be remembered. The participants were to decide whether the probe tone corresponding to the "?" location was the same as the one at that location during encoding or was different. If the tone was different, it did not match any of the tones in the presented series, and the participants were made aware of that. On half of the trials, the correct answer would be "same," and on the other half of the trials, the correct answer would be "'different," and the test tone was randomly selected from the 12 tones other than the target tone. The

participants were instructed to press "s" for "same" and "d" for "different," and they had unlimited time to respond. Feedback that lasted for 500 ms was provided after the participant made a response. A blank period with a dot in the center of the screen lasted for 1,000 ms before the next trial started.

The trials were allocated into 10 blocks. Each block contained 4 trials for each condition, adding up to 28 trials per block. Each trial lasted for 16 s, and the experiment lasted for 1.5 h.

In half of the blocks, the participants were instructed to whisper "the" twice a second during the encoding and maintenance phases ("whisper" sessions); to equate attention cost, in the other half of the sessions, they were instructed to tap the right index finger on the table twice a second during these phases ("tap" sessions) (Ricker et al., 2010). The "whisper" and "tap" sessions were arranged in a consistent order, *whisper–tap–tap–whisper–whisper–tap–tap–whisper–whisper–tap*.

At the beginning of the experiment, participants were trained to whisper "the" and tap their right index finger on the desk, each for 1 min. During the practice, there was a beep every second to help the participants keep the pace. After the training, participants performed two practice memory blocks, each consisting of seven trials (one trial per condition). The first practice session was a "whisper" block, and the second practice was a "tap" block.

Results and discussion

Accuracy

A two-way repeated measure ANOVA of PM1 response accuracy with the set size of tones to be remembered (two, three, four, five, or six) and articulation condition ("whisper" and "tap") as within-participants factors revealed significant main effects of set size, $F(4, 104) = 15.39$, $\eta_{p}^2 = .37$, $p < .01$, and articulation, $F(1, 26) = 4.88$, $\eta_{p}^2 = .15$, $p < .05$. The interaction between set size and articulation was not significant, $F(4, 104) = 1.27$, $p > .05$ (see Fig. 2, top left). The main effect of set size showed better performance at smaller set sizes, and the main effect of articulation suggested better performance in the "tap" condition, meaning that repeating a simple word could interrupt rehearsal of tones, replicating the results of previous work (Schendel & Palmer, 2007). Full ANOVA tables of all three experiments are provided in the supplementary material (Supplementary Fig. 1).

Tone change magnitude effects The frequency difference between target and test tones varied drastically among the "change" trials. It is possible that magnitude of frequency change might have influenced WM capacity, in that

Fig. 2 Results for presentation method 1, in which a list of six tones was presented and an arrow cue indicated the first tone that was to be remembered. Top panels: Accuracy rates. Bottom panels: *k* value estimates. Left-hand panels: Results for Experiment 1. Middle panels: Results for Experiment 2. Right-hand panels: Results for Experiment 3. The *k* values are calculated with Cowan's *k* formula (Cowan, 2001). The solid lines with open triangles denote the "tap" trials, and the dashed lines with solid squares denote the "whisper" trials. The error bars represents 95 % repeated measure confidence intervals (Hollands & Jarmasz, 2010)
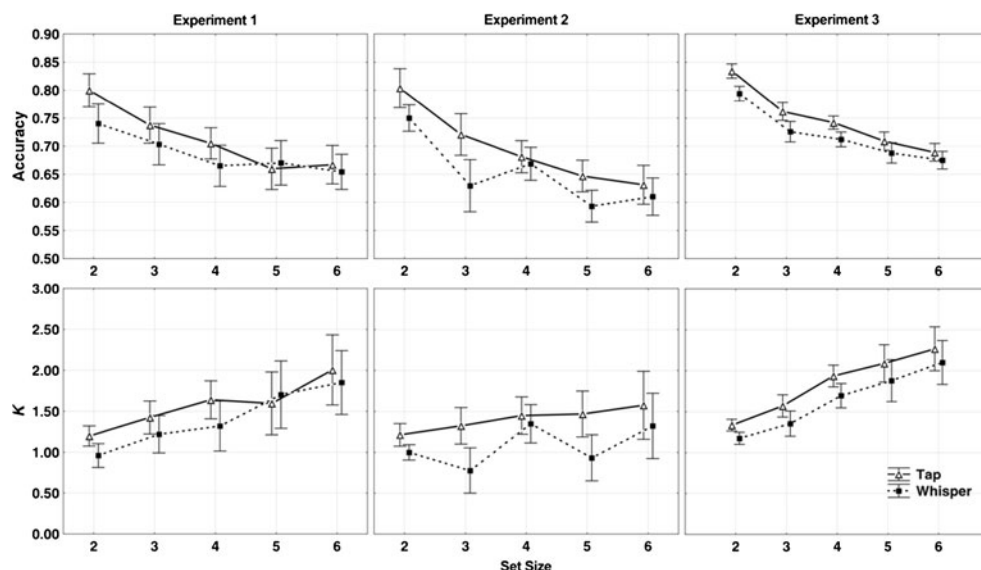
participants performed worse when frequency change was relatively small (although still clearly discriminable). Therefore, we examined the effect of tone change magnitude on memory performance. Only "change" trials were included in the analysis. All the "change" trials were sorted according to the frequency difference between target and test tones, ranging from 1 to 11 tones apart. Trials with frequency differences from 1 to 4 tones apart were combined as one condition ("small"), and the remaining, larger-difference trials were combined as another condition ("large"). Groups were divided in this way so that "small" and "large" groups would have approximately the same number of trials.[2] A three-way repeated measure ANOVA of the hit rate data, with set size (two, three, four, five, and six), articulation condition ("whisper" and "tap"), and tone change condition ("small" and "large") as within-subjects factors, revealed significant main effects of set size, $F(4, 104) = 4.43$, $\eta_p^2 = .15$, $p < .01$, and tone change condition, $F(1, 26) = 30.47$, $\eta_p^2 = .54$, $p < .001$. The hit rate was 0.62 ± 0.06 for small tone differences and 0.72 ± 0.06 for large tone differences; the interval represents 95 % repeated measure confidence intervals (Hollands & Jarmasz, 2010).

The only significant interaction was between set size and tone change magnitude, $F(4, 104) = 2.86$, $\eta_p^2 = .10$, $p < .05$. Post hoc Newman–Keuls tests revealed that the effect of tone change magnitude was significant ($p < .05$) at set size 2, 3, and 4, but not at set size 5 and 6 (see Fig. 3, left panel). This effect is unlikely to have been due solely to stimulus

discriminability, because if stimulus discriminability was the cause, performance would have been better at large than at small tone change across all set sizes, rather than set sizes no bigger than four. Instead, it is possible that participants remembered two types of information in this task. The first type is categorical information. For example, participants could sort a tone to a certain category, such as "high tone," "medium tone," or "low tone," and keep the categorical information in WM. The second type is more detailed acoustic information. When target and test tones were different enough, participants would be able to use both categorical and detail information during the test; when the difference was small; however, participants would be unable to use categorical information, because target and test tones would be sorted into the same category. The effect of categorical information diminished at set size 5 and 6, when performance in the large change trials fell down to the same level as in the small change trials. One explanation is that, at large set sizes, participants were unable to use any of their capacity to encode the more detailed information. Moreover, a higher proportion of correct responses at large set sizes come from lucky guesses, which do not depend on the tone frequency change magnitude.

Items in WM

For each set size, we calculated the participants' WM capacity using Cowan's *k* formula as noted above. The results are shown in Fig. 2 (bottom left). The highest capacity estimate (mean ± 95 % within-participants confidence interval) was 2.01 ± 0.42 tones at set size 6, lower than those estimated in simple visual or verbal WM tasks (Chen & Cowan, 2009; Luck & Vogel, 1997), while similar to the

---

[2] In another analysis, we assigned trials with a frequency difference of 1–5 tones to a "small" group and trials with a frequency difference of 7–11 tones to a "large" group. The results showed a similar trend as the original analysis (see supplementary material, Fig. 1), suggesting that the effect of tone change magnitude is not specific to a special grouping criterion.
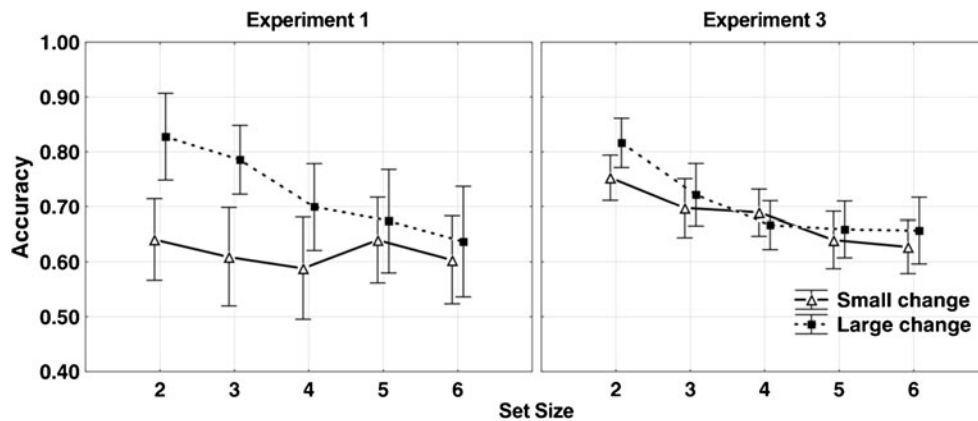
Fig. 3 Effects of tone change magnitude in Experiments 1 and 3. Experiment 2 was not included in this figure, because its procedure was different from that in Experiments 1 and 3 and, thus, would not contribute to a fair comparison. Left-hand panel: Experiment 1. Right-hand panel: Experiment 3. The solid lines with open triangles denote the small tone change magnitude, and the dashed lines with solid squares denote the large tone change magnitude. The error bars represents 95 % repeated measure confidence intervals (Hollands & Jarmasz, 2010)

capacity limit found in the previous studies on memory for tone sequences (Prosser, 1995).

To investigate the influence of the different stimulus presentation methods, we used only the capacity (k) of set size 2 and 4, but both presentation methods as a dependent variable, and conducted a two-way repeated measure ANOVA with set size (2 and 4) and presentation method (PM1 and PM2) as within-participants factors. The results revealed significant main effects of set size, $F(1, 26) = 11.83$, $\eta_p^2 = .31$, $p < .01$, and presentation method, $F(1, 26) = 5.11$, $\eta_p^2 = .16$, $p < .05$. The interaction was not significant, $F(1, 26) = 0.02$, $p > .05$. The main effect of presentation method suggested that people performed slightly better when they memorized all the stimuli that they heard, instead of starting to remember from a specific stimulus. Nevertheless, the average k values for PM2 were still low, with 1.17 ± 0.13 at set size 2 and 1.56 ± 0.25 at set size 4, as compared with 0.99 ± 0.14 at set size 2 and 1.40 ± 0.26 at set size 4 for PM1. The results indicate that changing the presentation method would not induce much improvement in terms of k value estimates.

It is possible that tones along a frequency continuum introduce sequential interference (e.g., Deutsch, 1970) that sometimes results in the inefficient use of WM capacity. In order to assess participants' best performance, we examined the maximum k value for each individual, no matter which combination of set size and articulation condition produced that maximum. Trials in PM1 and PM2 were combined to calculate k values at set size 2 and 4. This produced a mean maximum of 3.11 items ($SD = 0.94$), within the range of mean capacities observed in studies with categorical stimuli.

Stimulus discriminability is unlikely to have been the cause for the low WM capacity in this experiment. Adjacent tones have a 31 % frequency difference, which not only is a large difference, but also avoids any pair of tones with different pitches but equal chroma. Therefore, it is unlikely

that the stimuli in this experiment are difficult to discriminate from each other.

Experiment 2

The estimated auditory WM capacity in Experiment 1 was lower than the measured visual or verbal WM capacity in previous studies, even in the presence of a long retention interval (Cowan et al., 2011). In this next experiment, we tried one method that might allow more categorical information about the tones to be extracted. An early study on short-term memory for tone lists found better accuracy rates when the context tones were also presented together with the probe tone, as compared with the single-tone probe (Dewar et al., 1977). The authors suggested that higher-order information, such as relational or pattern information, aided in the WM performance. Accordingly, in the second experiment, we re-presented the entire studied list as a test probe, with or without a change in one tone. This method should allow any contextual information encoded from the studied list to be of use in the test.

Method

Participants

Twenty-four undergraduate students (7 male and 17 female) at the University of Missouri participated in the experiment to fulfill introductory psychology course requirements.

Apparatus and stimuli

The apparatus and stimuli were the same as those used in Experiment 1.

Procedure

The procedure was similar to that in Experiment 1, except for one difference during the test phase. Instead of presenting only one test tone and a question mark, in the present experiment, a list of tones was presented during the test. The number of the tones during the test was the same as the number of tones that the participants were supposed to remember. Characters were also presented sequentially, one per tone, each in one of the circles, starting from the circle marked with the →, simultaneously with the stimuli presentations. One of the characters was a question mark ("?"). The participants were instructed to decide whether the test tone corresponding to the "?" was the same as the tone at the same specific location during stimulus presentation or whether it was different from any of the tones that they remembered. When the correct answer was "different," the test tone was randomly selected from the 12 tones other than the target.

Results and discussion

Accuracy

For the trials of PM1, the same two-way repeated measure ANOVA was conducted and revealed significant main effects of set size, $F(4, 92) = 18.24$, $\eta_p^2 = .44$, $p < .01$, and articulation condition, $F(1, 23) = 12.74$, $\eta_p^2 = .36$, $p < .01$. The main effect of set size indicated higher performance at low set sizes, and the main effect of articulation condition reflected higher performance when there was no articulatory suppression. The interaction between set size and articulation was not significant, $F(4, 92) = 1.59$, $p > .05$ (Fig. 2, top middle).

Tone change magnitude effects As in Experiment 1, we analyzed the effect of tone change magnitude in Experiment 2. All the "change" trials were sorted according to the frequency difference between target and test tones, ranging from 1 to 11 tones apart. Trials with frequency difference from 1 to 4 tones apart were combined as one condition ("small"), and the other trials were combined as another condition ("large"). A three-way repeated measure ANOVA of the hit rate data with set size (2, 3, 4, 5, and 6), articulation condition ("whisper" and "tap"), and tone change condition ("small" and "large") as within-subjects factors revealed significant main effects of set size, $F(4, 92) = 22.56$, $\eta_p^2 = .50$, $p < .001$, articulation condition, $F(1, 23) = 16.33$, $\eta_p^2 = .42$, $p < .001$, and tone change condition, $F(1, 23) = 22.56$, $\eta_p^2 = .50$, $p < .001$. The hit rate was 0.60 ± 0.05 for small tone differences and 0.66 ± 0.05 for large tone differences; the interval represents 95 % repeated measure confidence intervals (Hollands & Jarmasz, 2010). The main effect of tone change magnitude suggests that performance was better with large tone changes. However, no interaction effect was significant, including the interaction

between tone change magnitude and set size, $F(4, 92) = 1.03$, $p > .05$. Perhaps a larger tone change was always important because it changed the pitch contour in terms of the sequence of categorizations, which would be more easily observable given the full-list probe in this experiment.

Cross-experiment comparison In order to examine the effect of experimental method (single-tone probe versus full-list probe), we combined the data from Experiments 1 and 2 and conducted an ANOVA with accuracy rates as the dependent variable, the two experiment groups as a between-participants factor, and three within-participant factors: set size, articulation condition, and tone change magnitude. The only significant effects involving experiment were two-way interactions with set size, $F(4, 192) = 5.47$, $\eta_p^2 = .10$, $p < .001$, and tone change magnitude, $F(1, 48) = 4.43$, $\eta_p^2 = .08$, $p < .05$. In Experiment 1, for the five set sizes, the means were .73, .70, .64, .66, and .62, respectively. In Experiment 2, the means were .77, .67, .67, .53, and .51. Thus, they were substantially lower with large set sizes in Experiment 2. Moreover, in Experiment 1, for small and large tone changes, the means were .62 and .72, respectively. For Experiment 2, the means were .60 and .66. There was clearly less benefit of large tone changes in Experiment 2.

Items in WM

We again calculated the k value for each set size. The highest k value was 1.58 ± 0.41 tones at set size 6, even lower than the highest k value in Experiment 1 (Fig. 2, bottom middle). The k value at set size 4 seems higher than the trend would predict, probably because encoding of items into WM becomes less efficient when the capacity limit is surpassed (in this case, set size 5 and 6), which occurs for visual arrays (Cusack, Lehmann, Veldsman, & Mitchell, 2009).

A two-way repeated measure ANOVA with capacity (k) of set size 2 and 4 as a dependent variable and set size (2 and 4) and presentation method (PM1 and PM2) as factors was conducted to investigate the presentation method effect as we did in Experiment 1. We again found significant main effects of set size, $F(1, 23) = 5.91$, $\eta_p^2 = .20$, $p < .05$, and presentation method, $F(1, 23) = 7.98$, $\eta_p^2 = .26$, $p < .01$. The interaction effect was not significant, $F(1, 23) = 0.77$, $p > .05$. The average k value was 1.18 ± 0.11 (set size 2) and 1.53 ± 0.18 (set size 4) for PM2, as compared with 1.04 ± 0.12 and 1.27 ± 0.22 for PM1. Capacity improvement is very limited for PM2, considering capacity estimates of close to 2 and 4 at set size 2 and 4, respectively, in previous WM studies using categorical stimuli (Cowan, 2001). Neither Experiment 1 nor 2 found a large capacity improvement for PM2, suggesting that the low capacity estimates in these experiments were not simply due to the specific experimental procedure.

It is clear that presenting the tone list instead of a single tone did not improve people's performance in the simple-tone WM task. The discrepancy between our experiments and the Dewar et al. (1977) study was probably because the latter used tones of a chromatic scale. They found that recognition memory was more accurate under full-context conditions than under no-context conditions even for sequences of random tones. However, their random-tone sequences were always selected from 12 tones of a chromatic scale, as compared with musical sequences selected from seven notes in the same octave of a major scale. Note that even the random (atonal) sequences used by Dewar et al. included a majority of intervals from a major scale, so that sequences still might be encoded as melodies with some "wrong" notes. Certainly, the tones of a chromatic scale include far more musical and familiar intervals (Shepard, 1982) than do our tones, with frequencies that span nearly four octaves and notes without any consistent relationship between height and chroma. For example, the first five of our stimuli are closest to the musical notes G3, C4, F4, A4, and D5, with intervals that differ, on average, by 40 cents from any musical intervals. In that light, it is not surprising that context made so little difference to memory for our thoroughly nonmusical stimuli. On the basis of these results, the single-tone probe should be appropriate for this study, so we continued with the single-tone probe in Experiment 3.

Lastly, as in Experiment 1, we again examined maximum k value for any condition and found that it averaged 2.68 items ($SD = 0.81$), slightly lower than the mean capacities typically observed with categorical stimuli in previous studies (Cowan, 2001).

Experiment 3

In Experiment 1, we proposed two types of information maintained in WM: category and detail. We also proposed that the effect of tone change magnitude in Experiment 1 was due to the fact that at smaller target–test tone change magnitude, the categories that participants formed from target and test tones were the same, making it impossible to detect the change. If a stimulus set contains other discriminating information, participants might rely less on the continuous dimension of pitch to categorize the tones. In Experiment 3, we added timbres based on musical instruments to the tones we used in Experiments 1 and 2. We expected that, due to the added timbre information, performance would improve overall but the effect of pitch change magnitude would be less than that in Experiment 1 or 2.

This last experiment was conducted with a large sample size to yield a better picture of individual differences in maximum capacity, for comparison with the capacities obtained in other studies with more categorical stimuli.

Method

Participants

One hundred nine undergraduate students (60 male, 49 female) at the University of Missouri participated in the experiment to fulfill introductory psychology course requirements.

Apparatus and stimuli

The apparatus in Experiment 3 was the same as that used in Experiments 1 and 2. The only difference is that the stimuli we used in Experiment 3 had different timbres in addition to different frequencies. (These stimuli can be heard within the online supplementary materials at http://dx.doi.org/10.3758/s13414-012-0383-z.) We selected 12 sounds generated with GarageBand (Apple Inc., Cupertino, CA), a program in the Macintosh Operating System, each played by a distinct instrument (*Trumpet Section, Smooth Clav, Classic Rock Organ, Negril Bass, Tenor Sax, Space Harpsichord, Grand Piano, Live Pop Horns, Aurora Bell, Pop Flute, Hollywood Strings,* and *Clean Electric Guitar*). Then we varied the fundamental frequencies of these sound files to be the same as the frequencies that we used in Experiments 1 and 2, from lowest (200 Hz) to highest (3900 Hz) in the order shown.

Procedure

The procedure was the same as that in Experiment 1 (i.e., the single-tone probe method), except that we used the multidimensional sounds instead of pure tones. Additionally, just after the experiment, the participants answered a questionnaire to specify how many sounds they memorized by labeling them as objects such as instruments, instead of by purely acoustic properties. They also rated from 1 to 5 the extent to which they relied on the labels to remember the sounds, 1 being *mostly acoustic* and 5 being *mostly labeled*.

Results and discussion

Accuracy

We conducted a similar two-way repeated measure ANOVA on PM1 accuracy with set size (2–6) and articulation condition ("whisper" and "tap") as within-participants factors. The results (Fig. 2, top right) revealed significant main effects of set size, $F(4, 432) = 74.09$, $\eta_p^2 = .41$, $p < .01$, and articulation condition, $F(1, 108) = 23.69$, $\eta_p^2 = .18$, $p < .01$. The interaction was not significant, $F(4, 432) = 0.93$, $p > .05$.

Tone change magnitude effects We also analyzed the influence of tone change magnitude between target and test sounds on performance in Experiment 3. All the "change" trials were

sorted according to the magnitude of frequency change between target and test tones, ranging from 1 to 11 tones apart. Trials with frequency difference from 1 to 4 tones apart were combined as one condition ("small"), and the other trials were combined as another condition ("large"). A three-way repeated measure ANOVA of the hit rate data with set size (2, 3, 4, 5, and 6), articulation condition ("whisper" and "tap"), and tone change magnitude condition ("small" and "large") as within-subjects factors revealed significant main effects of set size, $F(4, 424) = 38.83$, $\eta_p^2 = .27$, $p < .001$, articulation condition, $F(1, 106) = 6.63$, $\eta_p^2 = .06$, $p < .01$, and tone change magnitude, $F(1, 106) = 6.39$, $\eta_p^2 = .06$, $p < .01$. The hit rate was 0.68 ± 0.03 for small tone differences and 0.70 ± 0.03 for large tone differences; the interval represents 95% repeated measure confidence intervals (Hollands & Jarmasz, 2010).

The interaction between set size and tone change magnitude was the only significant interaction effect, $F(4, 424) = 3.01$, $\eta_p^2 = .03$, $p < .01$. Post hoc Newman–Keuls tests revealed that the effect of tone change magnitude was significant ($p < .05$) at set size 2, but not at larger set sizes (see Fig. 3, right panel). This pattern is very different from the results in Experiment 1 (Fig. 3, left panel), in which performance was better with large frequency change at set size 2, 3, and 4. Clearly, magnitude of tone change no longer played as important a role as in Experiment 1, probably because participants were able to categorize the sounds on the basis of timbre information that is based on knowledge from long-term memory.

The magnitude of tone change still had an effect at the smallest set size. Since there were only two timbres to maintain at set size 2, it is likely that participants had spare mental resources to retain tonal information after timbre information entered into WM. Therefore, when frequency change was large, participants were able to make judgments on the basis of categorical information induced by frequencies, in addition to timbres, at set size 2.

The results of Experiment 3 show that when timbres were added to the stimulus set, the effect of frequency change became less pervasive. These results indicated the important role of categorical information in WM tasks.

Cross-experiment comparison In order to examine the effects of adding timbre in Experiment 3, we combined the results of Experiments 1 and 3 in an ANOVA with experiment as a between-subjects factor and with three within-subjects factors (set size, articulation condition, and the magnitude of tone change). The effects involving experiment as a factor that were significant include a two-way interaction of experiment group and tone change magnitude, $F(1, 131) = 16.99$, $\eta_p^2 = .11$, $p < .001$, as well as a three-way interaction of experiment group, tone change magnitude, and set size, $F(4, 524) = 2.84$, $\eta_p^2 = .02$, $p < .05$. As is shown in Fig. 3, the advantage of a small set size occurred when the tone change magnitude was large or when timbres were present (Experiment 3), but not when both of these advantages were absent (Experiment 1, small set sizes with small tone changes).[3]

Items in WM

We calculated the average k values for each set size and found the highest average k value of 2.27 ± 0.19 at set size 6 (Fig. 2, bottom right), somewhat lower than the usual k values of 3 to 5 in WM studies using categorical stimuli (e.g., Cowan, 2001; Rouder et al., 2008).

Again, we conducted a two-way ANOVA with capacity (k) of set size 2 and 4 as dependent variable and set size (2 and 4) and presentation method (PM1 and PM2) as within-participants factors. The results revealed a significant set size effect, $F(1, 108) = 143.53$, $\eta_p^2 = .57$, $p < .001$, but the main effect of presentation method was not significant, $F(1, 108) = 0.38$, $p > .05$. Different presentation methods had no effect on WM performance in this task.

The large sample of this experiment yielded an excellent view of individual differences in the maximum k value. As in Experiments 1 and 2, to explore the possibility that capacity was used inefficiently for tones, we estimated the maximum capacity for each individual participant, regardless of articulation conditions and set sizes. Figure 4 shows that the average maximum k value for individual participants was close to the average value of k from past studies of WM (e.g., Cowan, 2001). Indeed, the mean of the maximum k values was 3.28 ($SD = 0.94$), similar to the average k values in studies with more conventional stimuli, such as categorically different colors (e.g., 3.35 items in Rouder et al., 2008).

The set size at which the maximal k value was reached was not random, but principled. In fact, the set size at which the maximal k value occurred was correlated with both the maximal k value, $r = .34$, $p < .001$, and the average k value, $r = .48$, $p < .001$. Thus, participants with higher maximal or average capacities tended to reach their maximal capacities at larger set sizes. The fact that the overall k values do not reach a stable asymptote (Fig. 2) could be because the number of items that can be encoded into WM sometimes declines after the maximum is reached, a finding that occurs in some studies with visual arrays as well, among lower-span individuals (Cusack et al., 2009). It can occur because items beyond the capacity can cause interference during encoding or maintenance (Shipstead & Engle, 2012).

---

[3] Caution is advised in interpreting the ANOVA result because the sample sizes in Experiments 1 and 3 are very different, but Fig. 3 adds confidence to our interpretation.
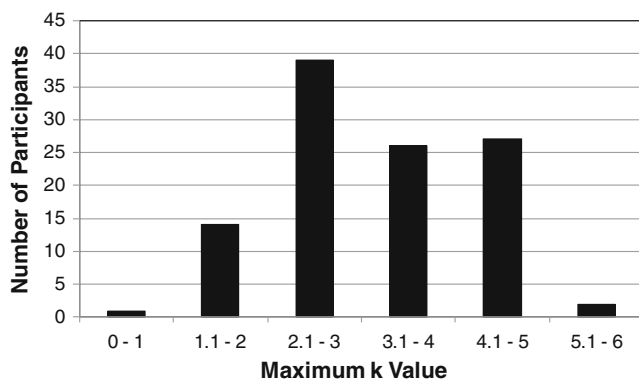
Fig. 4 Histogram of individual values of maximum k in Experiment 3

Questionnaire data

Although we wish to conclude that the capacity limit observed in this experiment is the limit in number of categorical acoustic items that can be held in WM, an alternative explanation might be that the participants labeled the sounds with certain instruments and memorized the sounds by their labels instead of their acoustic properties. This possibility can be examined, however, using the questionnaires that participants completed after the main procedure. They rated the number of sounds they were able to label, as well as the extent to which they relied on the labels, from 1 to 5. The mean number of sound labeled ($\pm SEM$) was 4.34 ± 0.57. We also calculated the weighted number of sounds for each individual as the number labeled multiplied by the rated reliance on labels divided by the maximum possible rating. For example, an individual who indicated that three sounds were labeled and that the reliance on those labels was 4 out of a possible 5 would receive a weighted score of $3 \times (4/5)$ $= 2.40$. The average weighted score was 3.23 ± 0.55, small compared with the maximum possible weighted number (12). Therefore, participants' usage of labeling seems to have been very limited in Experiment 3. Additionally, we also examined the correlation between the participants' overall accuracies and their ratings. No significant correlation was found between recognition accuracy and the number of sounds labeled, $r = -.11$, $p > .05$, or the weighted number, $r = -.15$, $p > .05$. These results show that labeling had little effect on WM performance in Experiment 3. (Presumably, that would have been the case also in Experiments 1 and 2, given that the stimuli in those experiments were pure tones.)

General discussion

Many previous studies have revealed a fixed memory capacity of three or four items or chunks when people were instructed to remember lists of simple items, such as auditory letters and visual colored squares (see, e.g., Cowan, 2001; Rouder et al., 2008). The most important evidence is

that the k value, which represents the number of items being kept in WM, increases with memory load, peaks at between three and four items in WM (or at about three items after a long retention interval), and then levels off. Such a pattern strongly indicates the presence of a fixed WM capacity. On the other hand, some studies have indicated that stimulus properties further limit WM capacity. Olsson and Poom (2005) showed that participants were able to remember only one item when the stimulus set contained little categorical information.

Few researches have studied WM capacity limit and its constraints in the domain of nonverbal auditory items. In the experiments above, we investigated the core auditory WM capacity limit by using different sets of auditory stimuli. Experiment 1 examined memory for tones in a list, using a single-tone probe, and found that capacity rose with increasing set size, but only to about two tones and not reaching a stable plateau even with a set size of six tones. Experiment 2 used a full-list probe in an attempt to create more tone context, but the outcome was even poorer performance. This finding is paralleled by poorer performance in visual array memory with a full-array probe, as compared with a single-item probe (see Wheeler & Treisman, 2002).

Finally, in Experiment 3, the tones were differentiated by the addition of timbre information from musical instruments. This manipulation did not uniformly improve performance. However, as is shown in Fig. 3, timbre made a difference in a particular circumstance. Specifically, an examination of trials according to the magnitude of tone change revealed that, at small set sizes, there was a notable advantage of information that could be used to form a small number of categories. This information could come in the form of large tone changes that might, for example, be used to distinguish between low, medium, and high tones (in all experiments), or it could come in the form of differentiable timbres (in Experiment 3 only) that might be used to distinguish between, say, percussion, strings, and woodwind or brass instruments.

The fact that the pattern of k values indicating items in WM did not reach a stable asymptote may indicate that there was variability in performance functions across set sizes beyond what is typically observed with more easily perceived or categorized stimuli. To overcome some of the variability, we examined each individual's maximal performance regardless of the set size and found that the maximum was quite in line with previous studies of capacity for categorized items in the visual realm (e.g., Cowan, 2001; Rouder et al., 2008): In Experiments 1–3, the mean maxima were 3.11, 2.68, and 3.28 items, respectively.

It is unlikely that the low capacity estimates in this study are due to aspects of the experimental design, rather than to cognitive limits. First, two presentation methods were used

and produced similar capacity estimates in all three experiments, suggesting that the additional visual cue was not the cause for low capacity estimates. Second, a comparison of memory for color arrays and tone series in change detection tasks (Morey, Cowan, Morey, & Rouder, 2011) showed higher capacity for colors than for tones. On each trial, a list of tones and an array of colored squares were presented sequentially, and recognition tests were conducted on both materials after a short delay. Attention allocation between color and tone memory was manipulated by assigning different proportions of monetary rewards to these two tasks, with the total amount of monetary rewards being equal across trials. The results showed that with the same proportion of reward, capacity estimates for tone lists were consistently lower than those for color arrays. Even when all rewards were allocated to the tone WM task, capacity estimates were still less than three items for tones, as compared with over four items for colors when all rewards were allocated to the color WM task. This result confirms that WM for tones is indeed less effective than WM for categorical visual stimuli—in this case, colors.

Indeed, the low capacity estimates in auditory WM for simple tones as revealed in Experiments 1 and 2 are consistent with the results in some previous studies on memory for tone sequences (Prosser, 1995). Camos and Tillmann (2008) presented to participants a list of rapid auditory tones differing in frequency and instructed the participants to evaluate the number of the tones that they heard. A big discrepancy in terms of response time was found between the lists of two and three tones, suggesting that the participants were able to keep up to two simple tones in the focus of attention. The low capacity found in Experiment 3 is consistent with a recent study by Golubock and Janata (2012), which investigated auditory WM capacity for sounds with the same frequency but different timbres and found a capacity of up to 2.56 items even with a short 1-s delay and increased perceptual variability among list items. The highest capacity estimate was only 1.80 items with a long 6-s delay, which is the duration that we used in this study.[4] The capacity estimates in Golubock and Janata were even lower than those in Experiment 3 in this study, probably because their stimuli were more difficult to categorize than the stimuli we used in Experiment 3.

We did not directly compare WM capacity for nonverbal sounds and visual materials in this study. Therefore, WM capacity for stimuli with little categorical information across modalities requires further investigation. These stimuli may include colors varying in a single dimension (within-category hue detail, saturation, or brightness), lines along a length continuum, or tactile perception induced by varying

---

[4] These k values are recalculated using Pashler's formula modified for a single, central probe (Pashler, 1988). Golubock and Janata (2012) used Cowan's k formula, which does not fit the experiment design as well as Pashler's modified formula (see Cowan, Blume, and Saults, 2012). The values reported by Golubock and Janata were even smaller.

forces. Some previous studies indeed found low WM capacity estimates for visual stimuli (Alvarez & Cavanagh, 2004; Olsson & Poom, 2005). Olsson and Poom used visual stimulus sets with continuous features and found very low capacity estimates that are comparable to the capacity estimates found in this study. Therefore, it seems that WM capacity is more relevant to the amount of categorical information, instead of modalities in a stimulus set.

The mean capacity estimates in both our Experiment 3 and in Olsson and Poom's (2005) categorical condition were somewhat lower than the ordinary capacity limit of three or four items (Cowan, 2001) or about three after a long delay (Cowan et al., 2011). In Olsson and Poom's study, the participants needed to memorize the conjunction of shape and color, but the short stimulus presentation time might have prevented them from chunking the shape and color together into a unified object, leading to a lower capacity estimate. In our Experiment 3, the timbre might have been useful in order to classify sounds into some limited number of distinct categories (e.g., percussion, strings, brass, or woodwind) without having categories detailed enough to allow all 12 sounds to be classified.

Some previous research with relatively low WM capacities might be explained in terms of categorical information in the stimulus set (Alvarez & Cavanagh, 2004). For example, in Awh et al. (2007), when similarity between target and test was low, WM performance was equivalent for complex objects, such as Chinese characters, and simple objects, such as colored squares. The authors concluded with a two-factor model, which accounts for both the number and resolution of items in WM. This is a precursor to our hypothesis that, for a given stimulus, observers could store both categorical information (e.g., *this is a Chinese character*) and detailed information (e.g., *how this Chinese character looks*).

It is not the case that capacity can continue to grow by making stimuli more and more dissimilar from one another. Anderson et al. (2011) used a procedure in which the precision of the recollection of an item's orientation could be examined, and they found that with set sizes larger than three items, there were no further increases in the number of items recalled and no further loss in the precision of each item recalled. Therefore, we believe that for maximal WM storage, the stimuli must be dissimilar enough to allow clear categorization, but not necessarily any more dissimilar than that.

In this article, we discussed three experiments on core auditory WM capacity and its constraints. People were able to retain up to two tones, on average, and performance was better with large than with small target–test tone change at set size 2, 3, and 4, but not 5 and 6. This result suggested two types of information stored in WM: category and details. When timbre information was added to the stimulus set, most of the effect of tone change magnitude disappeared. The average capacity with the tones with timbre

was still slightly lower than the three to four items typically found for categorical stimuli, such as known characters or colors, after an extended retention interval. The low capacity may be due to the fact that categorical information for the timbres in this study was not as readily available in long-term memory as that for traditional visual and verbal stimuli. Thus, when each participant's maximal capacity was observed, the mean maximal capacity was found to be quite similar to what has been found for object arrays (slightly more than three tones, on average).

More work is needed to clarify the nature of the details in WM that fall outside of categorical information. Cowan (1984) noted that there is long-term memory for aspects of sound, and some of these aspects may reside in WM despite the presence of a mask. On the other hand, the detailed information could be stored in an auditory analogue of the visuospatial sketchpad hypothesized by Baddeley and Hitch (1974) and Baddeley (1986). Further research is needed to measure core auditory WM capacity with different stimulus sets and to study the relationship between core WM capacity and the distinction between categorical and detailed information.

References

Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by information load and by number of objects. *Psychological Science, 15,* 106–111.

Anderson, D.E., & Awh, E. (2012). The plateau in mnemonic resolution across large set sizes indicates discrete resource limits in visual working memory. *Attention, Perception, & Psychophysics, 74,* 891–910.

Anderson, D., Vogel, E. K., & Awh, E. (2011). Precision in visual working memory reaches a stable plateau when individual item limits are exceeded. *Journal of Neuroscience, 31,* 1128–1138.

Awh, E., Barton, B., & Vogel, E. K. (2007). Visual working memory represents a fixed number of items regardless of complexity. *Psychological Science, 18,* 622–628.

Baddeley, A. D. (1986). *Working memory.* Oxford, England: Clarendon Press.

Baddeley, A. & Hitch, G. J. (1974). Working memory. In: *Recent advances in learning and motivation, vol. 8,* ed. G. Bower. Academic Press.

Baddeley, A. D., Thomson, N., & Buchanan, M. (1975). Word length and the structure of short term memory. *Journal of Verbal Learning and Verbal Behavior, 14,* 575–589.

Bays, P. M., Catalau, R. F. G., & Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision, 9,* 1–11.

Bays, P. M., & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science, 321,* 851–854.

Boersma, P., & Weenink, D. (2009). Praat: Doing phonetics by computer (Version 5.2.02) [Computer program]. Retrieved Sept, 2010, from http://www.praat.org/

Burns, E. M., & Ward, W. D. (1999). Intervals, scales, and tuning. In D. Deutsch & D. Deutsch (Eds.), *The psychology of music* (2nd ed., pp. 215–264). San Diego, CA US: Academic Press.

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

Chen, Z., & Cowan, N. (2009). Core verbal working-memory capacity: The limit in words retained without covert articulation. *The Quarterly Journal of Experimental Psychology, 62,* 1420–1429.

Conrad, R. (1964). Acoustic confusion in immediate memory. *British Journal of Psychology, 55,* 75–84.

Conrad, R., & Hull, A. J. (1964). Information, acoustic confusion and memory span. *British Journal of Psychology, 55,* 429–432.

Cowan, N. (1984). On short and long auditory stores. *Psychological Bulletin, 96,* 341–370.

Cowan, N. (1995). *Attention and memory: An integrated framework.* Oxford Psychology Series (No. 26). New York: Oxford University Press

Cowan, N. (2001). The magical number four in short-term memory: A reconsideration of mental storage capacity. *The Behavioral and Brain Sciences, 24,* 87–114.

Cowan, N. (2005). *Working memory capacity.* Hove, East Sussex, UK: Psychology Press.

Cowan, N., Blume, C.L., & Saults, J.S. (2012). Attention to attributes and objects in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition.* doi:10.1037/a0029687.

Cowan, N., Johnson, T. D., & Saults, J. S. (2005). Capacity limits in list item recognition: Evidence from proactive interference. *Memory, 13,* 293–299.

Cowan, N., Li, D., Moffitt, A., Becker, T. M., Martin, E. A., Saults, J. S., & Christ, S. E. (2011). A neural region of abstract working memory. *Journal of Cognitive Neuroscience, 23,* 2852–2863.

Cusack, R., Lehmann, M., Veldsman, M., & Mitchell, D. J. (2009). Encoding strategy and not visual working memory capacity correlates with intelligence. *Psychonomic Bulletin & Review, 16,* 641–647.

Darwin, C. J., Turvey, M. T., & Crowder, R. G. (1972). An auditory analogue of the Sperling partial report procedure: Evidence for brief auditory storage. *Cognitive Psychology, 3,* 255–267.

Davies, J. (1979). Memory for melodies and tonal sequences: A theoretical note. *British Journal of Psychology, 70,* 205–210.

Deutsch, D. (1970). Tones and numbers: Specificity of interference in short-term memory. *Science, 168,* 1604–1605.

Dewar, K. M., Cuddy, L. L., & Mewhort, D. J. K. (1977). Recognition memory for single tones with and without context. *Journal of Experimental Psychology: Human Learning and Memory, 3,* 60–67.

Drake, C., & Palmer, C. (2000). Skill acquisition in music performance: Relations between planning and temporal control. *Cognition, 74,* 1–32.

Fujisaki, H., & Kawashima, T. (1970). *Some experiments on speech perception and a model for the perceptual mechanism* (Annual Report of the Engineering Research Institute, Faculty of Engineering). Tokyo: University of Tokyo.

Golubock, J. L., & Janata, P. (2012). Keeping timbre in mind: Working memory for complex sounds that can't be verbalized. *Journal of Experimental Psychology. Human Perception and Performance.* doi:10.1037/a0029720. Advance online publication.

Hickock, G., Buchsbaum, B., Humphries, C., & Muftuler, T. (2003). Auditory–motor interaction revealed by fMRI: Speech, music, and working memory in Area Spt. *Journal of Cognitive Neuroscience, 15,* 673–682.

Hollands, J. G., & Jarmasz, A. (2010). Revisiting confidence intervals for repeated measures designs. *Psychonomic Bulletin & Review, 17,* 135–138.

Idson, W. L., & Massaro, D. W. (1976). Cross-octave masking of single tones and musical sequences: The effects of structure on auditory recognition. *Perception & Psychophysics, 19,* 155–175.

Jones, D. M., Macken, W. J., & Harries, C. (1997). Disruption of short-term recognition memory for tones: Streaming or interference? *The Quarterly Journal of Experimental Psychology, 50,* 337–357.

Kidd, G. R., & Watson, C. S. (1992). The "proportion-of-the-total-duration rule" for the discrimination of auditory patterns. *Journal of the Acoustical Society of America, 92,* 3109–3118.

Lavie, N. (1995). Perceptual load as a necessary condition for selection attention. *Journal of Experimental Psychology. Human Perception and Performance, 21,* 451–468.

Lehnert, G., & Zimmer, H. (2006). Auditory and visual spatial working memory. *Memory & Cognition, 34,* 1080–1090.

Logie, R. H., Della Sala, S., Wynn, V., & Baddeley, A. D. (2000). Visual similarity effects in immediate verbal serial recall. *The Quarterly Journal of Experimental Psychology, 53A,* 626–646.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature, 390,* 279–281.

Massaro, D. W. (1975). Backward recognition masking. *Journal of the Acoustical Society of America, 58,* 1059–1065.

Merat, N., & Groeger, J. A. (2003). Working-memory and auditory localization: Demand for central resources impairs performance. *The Quarterly Journal of Experimental Psychology, 56A,* 531–549.

Miller, G. A. (1956). The magical number seven plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63,* 81–97.

Morey, C. C., Cowan, N., Morey, R. D., & Rouder, J. N. (2011). Flexible attention allocation to visual and auditory working memory tasks: Manipulating reward induces a tradeoff. *Attention, Perception, & Psychophysics, 73,* 458–472.

Nairne, J. S. (1990). A feature model of immediate memory. *Memory & Cognition, 18,* 251–269.

Olsson, H., & Poom, L. (2005). Visual memory needs categories. *Proceedings of the National Academy of Sciences, 102,* 8776–8780.

Palmer, C. (2005). Sequence memory in music performance. *Current Directions in Psychological Science, 14,* 247–250.

Palmer, C., & Pfordresher, P. Q. (2003). Incremental planning in sequence production. *Psychological Review, 110,* 683–712.

Parmentier, F. B. R., & Jones, D. M. (2000). Functional characteristics of auditory temporal-spatial short-term memory: Evidence from serial order errors. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 222–238.

Pashler, H. (1988). Familiarity and visual change detection. *Perception & Psychophysics, 44,* 369–378.

Prosser, S. (1995). Aspects of short-term auditory memory as revealed by a recognition task on multi-tone sequences. *Scandinavian Audiology, 24,* 247–253.

Ricker, T. J., Cowan, N., & Morey, C. C. (2010). Visual working memory is disrupted by covert verbal retrieval. *Psychonomic Bulletin & Review, 17,* 516–521.

Rouder, J. N., Morey, R. D., Cowan, N., Zwilling, C. E., Morey, C. C., & Pratte, M. S. (2008). An assessment of fixed-capacity models of visual working memory. *Proceedings of the National Academy of Sciences (PNAS), 105,* 5975–5979.

Rouder, J. N., Morey, R. D., Morey, C. C., & Cowan, N. (2011). How to measure working-memory capacity in the change-detection paradigm. *Psychonomic Bulletin & Review, 18,* 324–330.

Saults, J. S., & Cowan, N. (2007). A central capacity limit to the simultaneous storage of visual and auditory arrays in working memory. *Journal of Experimental Psychology. General, 136,* 663–684.

Schendel, Z. A., & Palmer, C. (2007). Suppression effects on musical and verbal memory. *Memory & Cognition, 35,* 640–650.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime reference guide*. Pittsburgh, PA: Psychology Software Tools.

Sek, A., & Moore, B. J. (1995). Frequency discrimination as a function of frequency, measured in several ways. *Journal of the Acoustical Society of America, 97,* 2479–2486.

Shepard, R. N. (1982). Geometrical approximations to the structure of musical pitch. *Psychological Review, 89,* 305–333.

Shipstead, Z., & Engle, R.W. (2012). Interference within the focus of attention: Working memory tasks reflect more than temporary maintenance. *Journal of Experimental Psychology: Learning, Memory, and Cognition.* doi:10.1037/a0028467.

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs: General and Applied, 74,* 1–29.

Surprenant, A. M. (2001). Distinctiveness and serial position effects in tonal sequences. *Perception & Psychophysics, 63,* 737–745.

Surprenant, A. M., & Neath, I. (1996). The relation between discriminability and memory for vowels, consonants, and silent-center vowels. *Memory & Cognition, 24,* 356–366.

Thiele, J. E., Pratte, M. S., & Rouder, J. N. (2011). On perfect working-memory performance with large numbers of items. *Psychonomic Bulletin & Review, 18,* 958–963.

Warren, R. M., & Obusek, C. (1972). Identification of temporal order within auditory sequences. *Perception & Psychophysia, 12,* 86–90.

Watson, C. S., Foyle, D. C., & Kidd, G. R. (1990). Limits of auditory pattern discrimination for patterns with various durations and numbers of components. *Journal of the Acoustical Society of America, 88,* 2631–2638.

Wheeler, M. E., & Treisman, A. M. (2002). Binding in short term visual memory. *Journal of Experimental Psychology. General, 131,* 48–64.

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature, 453,* 233–235.

Zhang, W., & Luck, S. J. (2011). The number and quality of representations in working memory. *Psychological Science, 22,* 1434–1441.