Journal of Experimental Psychology:
Human Perception and Performance
2016, Vol. 42, No. 12, 2115–2124

© 2016 American Psychological Association
0096-1523/16/$12.00 http://dx.doi.org/10.1037/xhp0000268

The Capacity to Detect Synchronous Audiovisual Events Is Severely Limited: Evidence From Mixture Modeling

Christian N. L. Olivers
Vrije Universiteit Amsterdam

Edward Awh
University of Chicago

Erik Van der Burg
Vrije Universiteit Amsterdam and University of Sydney

Visual attention serves to select salient and relevant events from the visual input. Selective attention to a visual event can be driven by a synchronous sound. Interestingly, recent evidence suggests that a sound can only drive selection of 1 concurrent visual event, suggesting that attentional capacity is much lower for audiovisual events than for purely visual events. Here we corroborate and extend this finding using a mixture modeling technique that distinguishes between the probability and precision of perception. Observers were presented with displays of multiple continuously flickering objects, of which either 1 or 2 were coupled to a single sound. In 2 experiments, we found that the probability of correctly reporting an object was almost halved when the number of synchronized visual objects increased from 1 to 2. Precision, however, was not affected. This indicates that rather than attention being distributed across multiple simultaneous audiovisual events, just 1 of them is singled out for attentional selection. This was not due to a capacity limit for selecting the visual objects per se; pure visual cues elicited a much higher probability of report and in that case there were clear declines in precision at larger set sizes, indicating the concurrent selection of multiple items. The results point toward a dissociation in capacity for visually and aurally cued prioritization of visual objects.

Keywords: multisensory processing, audiovisual integration, audiovisual capacity, visual attention, cross-modal attention

The brain is highly sensitive to signals that coincide in time across different senses (e.g., Alais & Burr, 2004; Dalton & Spence, 2007; Jack & Thurlow, 1973; MacDonald & McGurk, 1978; Olivers & Van der Burg, 2008; Schroeder & Foxe, 2005; Senkowski, Talsma, Grigutsch, Herrmann, & Woldorff, 2007; Spence & Squire, 2003; Stein & Meredith, 1993; Sumby & Pollack, 1954; Vroomen & De Gelder, 2000). This is adaptive, because synchronous signals are likely to stem from the same event and thus increases reliability way beyond that provided by the separate signals in isolation. Moreover, each sense may contribute unique information on moment, location, and identity of the outside event—information that can then be used by the other sensory modality. In line with this, we have shown in a number of studies that a spatially uninformative sound makes a synchronous visual transient stand out from multiple competing visual transients (Olivers & Van der Burg, 2008; Van der Burg, Cass, & Alais, 2014; Van der Burg, Cass, Olivers, Theeuwes, & Alais, 2010; Van der Burg, Olivers, Bronkhorst, & Theeuwes, 2008a; Van der Burg, Olivers, & Theeuwes, 2012; Van der Burg, Talsma, Olivers, Hickey, & Theeuwes, 2011). Compared with conditions where there is no accompanying sound, or the sound is synchronized with a nontarget event, the sound increases the saliency of the visual target stimulus to the extent that it is prioritized for selection, as has become evident from visual search, but also from temporal order judgment tasks (Van der Burg, Olivers, Bronkhorst, & Theeuwes, 2008b). Consistent with this, Van der Burg, Talsma, Olivers, Hickey, and Theeuwes (2011) found that a sound-accompanied visual transient causes an enhanced response of the P1 component in the electro-encephalogram (EEG) signal, followed by an increased response of the N2pc component. These components are respectively taken as markers of increased visual saliency and visual selective attention (Luck, Woodman, & Vogel, 2000).

The fact that a sound can prioritize visual transients for selection raises the question of capacity. Given that attentional resources are limited, how many visually synchronized objects can a single sound deliver to downstream selection mechanisms? We know from visual attention studies that attention can select and hold on to about four visual objects simultaneously (Awh, Barton, & Vogel, 2007; Burkell & Pylyshyn, 1997; Cowan, 2010; Franconeri, Alvarez, & Enns, 2007; Luck & Vogel, 1997; Mandler &

Christian N. L. Olivers, Department of Experimental and Applied Psychology, Vrije Universiteit Amsterdam; Edward Awh, Department of Psychology, University of Chicago; Erik Van der Burg, Department of Experimental and Applied Psychology, Vrije Universiteit Amsterdam, and School of Psychology, University of Sydney.

This work was supported by ERC-CoG-615423 grant awarded to Christian N. L. Olivers, NWO Visitor Grant 040.11.268 for Edward Awh, and ARC-DE130101663 grant awarded to Erik Van der Burg.

Correspondence concerning this article should be addressed to Christian N. L. Olivers, Department of Experimental and Applied Psychology, Faculty of Behavioural and Movement Sciences Van der Boechorststraat 1, 1081 BT, Amsterdam, the Netherlands. E-mail: c.n.l.olivers@vu.nl

Shebo, 1982; Pashler, 1988; Pylyshyn & Storm, 1988; Trick & Pylyshyn, 1994; Yantis & Johnson, 1990). A priori one might therefore expect that an optimal system allows a sound to boost the same number of concurrent visual events, so that it makes maximal use of the capacity of downstream bottlenecks. Yet, recently we found evidence that the number of visual objects that can be boosted by a synchronous auditory signal is very limited, namely to at most one (Van der Burg, Awh, & Olivers, 2013). In that study, we presented 24 asynchronously flashing disks in a circular array. At a certain point, a short beep sounded in synchrony with one up to eight flashing disks. At the end of the trial, one of the disks was probed, and observers simply had to indicate whether it belonged to the synchronized set or not. Several experiments yielded estimates consistent with observers only effectively seeing one synchronized disk. In contrast, when target disks were not cued by a sound but by a brief color change, capacity rose to between three and four disks, in line with estimates of visual working memory capacity.

We concluded that a sound can only effectively bind to a single visual event. From a capacity point of view, this may appear puzzling. Why would multisensory mechanisms limit integration to single objects if more downstream mechanisms are happy to take on more. From an ecological perspective, however, it does make sense to bind only one visual object to a specific sound. In natural scenes, individual, object-related sounds come from a single source (Alais & Burr, 2004; Körding et al., 2007; Roseboom, Nishida, & Arnold, 2009; Roseboom, Nishida, Fujisakei, & Arnold, 2011; Soto-Faraco, Kingstone, & Spence, 2003; Van der Burg, Alais, & Cass, 2013). A perceptual system that is maximally tuned to its natural environment is unlikely to reserve a large capacity for multimodal events that it is unlikely to have, or gather, experience with. We will refer to this explanation as *the single source hypothesis*.

However, note that in our previous study (Van der Burg et al., 2013), observers were asked to indicate their perception with a discrete response (was this object synchronized or not). This leaves open the possibility that attentional resources—even though severely limited—were spread across multiple synchronized objects, rather than all being assigned to a single visual object. As we measured average performance, our study could not tell the difference between "one item's worth" of information distributed across multiple items and a case in which only one discrete item is selected at a time. Here we sought to distinguish between these possibilities with the use of a mixture modeling technique that since its introduction by Zhang and Luck (2008) has been widely adopted in the visual working memory literature. The procedure is outlined in Figure 1. As in Van der Burg et al. (2013), we presented one or two flashing disks either one of which could be synchronized with a single beep. The novel aspect concerned the task. Every disk in the display contained a randomly oriented line segment, which could be seen as a clock hand indicating a time. At the end of each trial, observers were probed on one of the target disks and had to indicate the time (which by then was no longer visible), on a continuous scale. By employing a continuous response scale, the mixture modeling technique allows for the dissociation of the *precision* of the orientation representation from the *probability* that the orientation was represented in the first place. It does this by estimating two underlying distributions: First, a Gaussian distribution of errors around the true value of the presented target stimulus. The standard deviation of this distribution provides a measure of the quality or precision of the mnemonic representation. Under the assumption that the more attentional or mnemonic resources are dedicated to a representation, the better its quality will be, this measure can thus provide an indication of how resources have been spread across different objects. The second component is a uniform distribution of errors across the response scale. This distribution represents the guesses people make when observers do not represent the stimulus at all. Its inverse thus represents the probability of observers having registered and remembered the stimulus in the first place. With this technique, it has generally been shown that both the probability of remembering and the precision of the representation suffer from increasing demands on visual working memory, suggesting a distribution of resources across multiple items (e.g., Bays & Husain, 2008; Huang, 2010; Wilken & Ma, 2004; Zhang & Luck, 2008).

Here we used the technique to test a number of scenarios with regard to detecting audiovisual events. Under the single source hypothesis, the audiovisual binding is limited to at most one visual

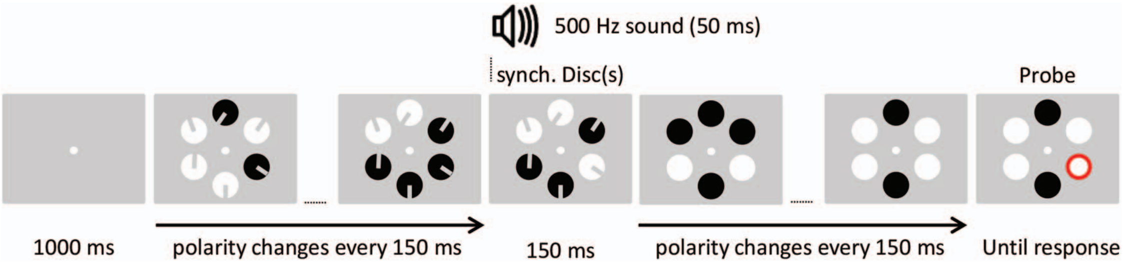

Figure 1. Illustration of the events in Experiment 1 (from left to right). Participants saw a total of 12 black and white disks on the screen (here only 6 for illustrative purposes). Every 150 ms, a randomly determined subset of one to six disks changed polarity, except for the target display in which only one or two disks changed. Each disk contained a randomly oriented line segment. The target display was always accompanied by a (spatially neutral) auditory signal. The task was to detect the synchronized disk(s) and to remember the line orientation within the disk(s). The line segments disappeared immediately following the target display, and another series of polarity changes followed. Eventually, the display became static, and a probe was presented on either one of the synchronized disks (valid probe), or on a nonsynchronized disk (invalid probe). Participants indicated the original orientation by using the mouse. See the online article for the color version of this figure.

item, and thus all resources fall on just one of the concurrently changing visual objects even if there are two.[1] That object should then be encoded with high precision—as high as in the case where there in fact is only one concurrently changing visual object. In other words, precision is predicted to remain constant with increasing set size. Instead, with two concurrently changing objects, we would predict a reduction in the probability of representation by half relative to one object, as one of the two objects is missed. The second scenario predicts the reverse. Under what we will refer to as the *distributed resources hypothesis,* multiple synchronized objects are detected, but limited processing resources are distributed more or less evenly across multiple synchronized disks. As a result of resources being spread more thinly, the resolution of representation will suffer. Thus, the precision with which the orientation inside the disk is represented will be affected, rather than the probability of detection. The third scenario then represents a mixture of the two. For example, limited resources may be spread so thinly across multiple visual objects that some line segments may not even reach the threshold of being perceived. In that case effects on both precision and probability of representation are predicted. This latter scenario would be most in line with what is known form the memory literature using visual stimuli. In sum, the result most indicative of a single item limit for audiovisual binding would be a drop in the probability of report together with constant precision across set sizes 1 and 2.

Experiment 1

Experiment 1 provided a first test of whether synchronizing multiple visual objects with a single sound leads to reduced precision or reduced probability of representing those objects. In an array of 12 disks, either one or two disks changed polarity concurrent with a sound. These polarity changes were embedded in a stream of random changes occurring every 150 ms. Each disk contained a line segment which immediately disappeared after the beep. At the end of the trial, a probe appeared on one disk, and participants were asked to indicate the probed orientation, by clicking on the rim of the disk. In this experiment, the probe was 50% valid—that is, on half the trials, a synchronized disk was probed, whereas on the other half a nonsynchronized disk was probed. The latter condition was included to check that the sound affected specifically the encoding of (and thus performance for) the individual synchronized disks, and not of the display in general. A mixture model (based on Zhang & Luck, 2008) then yields as most important parameters Pmem reflecting the probability of recalling the orientation, and the standard deviation (SD), reflecting the precision with which the orientation is represented. If the sound provides access to only one synchronized disk, as would be predicted by the single source hypothesis, then we should see Pmem drop when the number of synchronized objects increases from one to two. At the same time, precision for the accessed disk should remain constant. In contrast, under the distributed resource hypothesis, both objects may be accessed, but only through shared resources and thus reduced resolution.

Method

Participants. Twelve students (6 female; mean 23.7 years; ranging from 18–37 years) from the VU University participated in the experiment. Data from 2 participants were excluded as they were not able to do the task (*Pmem* was 0% for all four conditions). Participants were either paid (8 Euro per hour) or received course credits for their participation. All participants were naïve as to the purpose of the experiment. The protocol was approved by the local ethics committee.

Apparatus, stimuli, and procedure. The experiment was run in a dimly lit cubicle. Participants were seated approximately 80 cm from a 21-in. 120 Hz monitor and wore Sennheiser HD202 headphones. E-Prime software was used to program and run the experiment. Each trial began with a white fixation dot (0.08° visual angle, 95 cd m^{-2}) presented at the center of the screen for 1,000 ms. The background color was gray (10 cd m^{-2}) and kept constant during the experiment. Subsequently, participants viewed a changing visual display consisting of 12 black ($<$0.5 cd m^{-2}) and white (95 cd m^{-2}) disks (radius = 1.1° visual angle), randomly presented on an imaginary circle (radius = 6.5°) around the fixation dot. The initial polarity of the disks was randomly determined. Each disk contained a gray line segment (10 cd m^{-2}; length 1.1°) with a randomly determined orientation, thus resembling the hand of a clock. Its orientation remained fixed throughout the polarity changes.

The display changed every 150 ms, and the total number of display changes was randomly determined on a given trial. A trial first started with a sequence of 11–15 display changes, followed by the target display change, and three additional display changes (making a total of 15–19 display changes per trial). A nontarget display change consisted of a randomly determined number of disks (1–6) changing polarity (from white to black or vice versa). The target display consisted of either one or two disks changing polarity, which was always synchronized with the onset of a brief sound (500 Hz; 50 ms duration). All positions of the changes were chosen randomly (with replacement). Note that the short duration of each frame precludes any useful eye movements toward the target disks, and thus any performance differences cannot be due to differences in acuity.

Participants were told that it paid to detect the one or two synchronized disk(s) and remember the line orientation, as synchronized disks were 50% likely to be probed (leaving 50% likelihood of either one of the other 10 or 11 disks being probed). After the presentation of the target display, the line segments inside the disks were removed, and after another three display changes, the display became static, and a probe was presented. The probe was a red circle (radius = 0.5°; 19.89 cd m^{-2}) and presented on either one of the synchronized disks (valid probe), or on a nonsynchronized disk (invalid probe). As alluded to earlier, the validity of the probe was 50%. The mouse pointer was shown at the center of the probe, and participants were required to use the

[1] The single source hypothesis is different from what is known as the “unity assumption” (Bedford, 2001; Vatakis & Spence, 2007; Welch, 1999; Welch & Warren, 1980). The unity assumption refers to the situation where there is one stimulus in one modality (e.g., visual) and one in another (e.g., auditory), and the system needs to decide whether they belong to the same event or to two separate events. That is, the unity assumption pertains to resolving ambiguity as to whether there should be integration or not. The single source hypothesis here refers to the situation where there are more events in one modality than in the other. It assumes integration, but this then creates the problem as to which of the majority events should be integrated.

mouse to indicate the original line orientation in the probed disk, by pointing it out on the rim of the probed disk, and to confirm this by pressing the left mouse button. After the response, participants received feedback, as the original orientation reappeared for 250 ms. Subsequently, a blank screen was presented for 500 ms, and the next trial was initiated.

Design. The dependent variable was the offset between the original target orientation and the reported orientation. Factors were the number of visual objects synchronized with the auditory signal (one, or two), and probe validity (valid, or invalid), which were randomly mixed within 16 experimental blocks of 24 trials each, yielding 96 trials per cell. Three practice blocks of 24 trials each were presented prior to the real experiment. The stimulus onset asynchrony (SOA) between the display changes in the first practice block was 200 ms, and 150 ms for the remaining practice blocks. Furthermore, the probe was always valid for the first two practice blocks in order to familiarize participants with the task of detecting synchrony.

Results and Discussion

The dependent variable was the offset between the original target orientation and the reported orientation, with a theoretical range of 0 (perfect answer) to ± 180 degrees (maximally imprecise answer). As has been shown previously, and as can also be seen from Figure 2, the pattern of errors is well described by a mixture of two distributions (Zhang & Luck, 2008): (1) A uniform distribution captures those trials in which the orientation was presumably not registered, resulting in a random response. (2) A nonuniform distribution in which responses are centered on the correct orientation, but with a degree of imprecision, which is best described by a von Mises distribution. To obtain an estimate of these two distributions, error distributions were fit using Markov Chain Monte Carlo (MCMC) by applying Zhang and Luck's (2008) standard mixture model using the memfit function of Memtoolbox developed by Suchow, Brady, Fougnie, and Alvarez (2013). The procedure repeatedly samples parameter values for the distributions in proportion to how well they describe the data and the prior (an uninformative Jeffreys prior) to obtain a Maximum a Posteriori (MAP) estimate of three parameters, *Pmem, SD,* and, if necessary, μ, together with 95% credibility intervals. Here, *Pmem* is the probability that the cued orientation is available for report, and is equivalent to one minus the proportion of guesses. *SD* represents the standard deviation of the von Mises distribution, and is the measure of precision with which the orientation is registered. Last, μ represents the mean of the von Mises distribution, which is nonzero when there is a systematic bias in the error distribution (either clockwise or counterclockwise).

Aggregate data. Given the limited number of trials per condition, the most important fit is on the aggregate data, combined across all participants. Here Memtoolbox provides parameter estimates as well as 95% credibility intervals (CrI), with there being 95% probability that the CrI contains the true parameter value for the sample. We will refer to parameters with nonoverlapping credibility intervals as significantly different. For the aggregate data, we fitted a three-parameter model, including the bias parameter μ. Figure 2 describes the aggregate data for (a) probe valid, and (b) probe invalid trials.

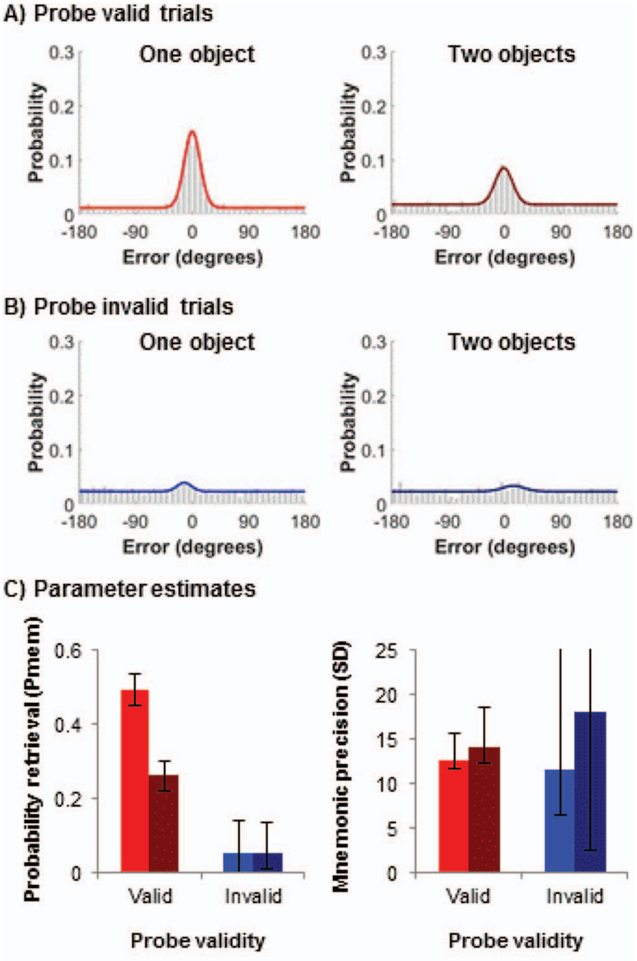

Figure 2. Results of Experiment 1. Mixture model fit to the aggregate data from all participants for probe valid trials (A). The left panel shows the performance when one visual object was synchronized with the auditory signal, while the right panel shows the performance when two visual objects were synchronized. The same, but now for probe invalid trials (B). Parameter estimates of probability of retrieval (*Pmem*) and precision (*SD*) as a function of probe validity and the number of synchronized objects (C) Lighter shades represent one synchronized object, darker shades two synchronized objects. Error bars represent Bayesian credibility intervals of the fits. See the online article for the color version of this figure.

As can be seen from Figure 2, for probe valid trials, the probability of report (*Pmem*) dropped significantly from one to two synchronized objects, one object: 49.5%, CrI: −4.4, +4.1; two objects: 26.5%, CrI: −4.4, +3.7, a reduction by just over 46%. Importantly, the precision (*SD*) did not vary between one and two synchronized objects, one object: 12.7°, CrI: −1.0°, +3.0°; two objects: 14.1°, CrI: −1.8°, +4.5°. There was also no systematic bias (μ) between the two conditions (one object: 0.0°, CrI: −1.5°, +1.5°; two objects: −1.1°, CrI: −2.6°, +2.8°). In addition, we conducted a post hoc analysis to examine potential grouping effects when two synchronized objects shared the same polarity, relative to when they differed in polarity. There was a numerical but nonsignificant benefit in *Pmem* when the two objects were similar (28.6%, CrI: −6.6%, +5.8%) compared with when they were dissimilar (25.1%, CrI: −5.7%, +5.1%). Precision was also not significantly better for similar (15.0, CrI: −2.7%, +4.6%) versus dissimilar pairs (12.8, CrI: −2.1%, +4.1%). Since there may be a lack of power, we will return to this aspect in Experiment 2.

For probe invalid trials, the probability of report (*Pmem*) was much worse than for probe valid trials, regardless of whether one or two objects were synchronized (one object: 5.3%, CrI: −5.2, +8.9; two objects: 5.2%, CrI: −4.0, +8.6). In terms of precision (*SD*), there was no statistical difference between one and two synchronized objects (one object: 11.7°, CrI: −5.3°, +77.0°; two objects: 18.1°, CrI: −15.5°, +52.3°), and precision also did not significantly differ from that for valid trials. However, note the large CrIs, as the low *Pmem* made the precision estimates less reliable. There was a difference in bias (μ) between the two conditions (one object: −12.6°, CrI: −56.9°, +8.8°; two objects: 14.2°, CrI: −16.8°, +31.1°), but given that this too was based on a very low *Pmem,* we take this finding with a grain of salt.

Analyses on individual fits. In addition, we also report analyses for the individual fits. Note that because of low trial numbers, these fits are to be interpreted with caution. That said, the results are consistent with those of the aggregate data. On probe invalid trials, a substantial number of participants could not reliably retrieve the probed object, leading to Pmem values of close to zero, or to unrealistically wide nonuniform distributions with impossible or unrealistic precision estimates as a consequence. Hence, these trials were left out of the analyses. Furthermore, the aggregate data revealed no bias, while at the individual level including a bias parameter occasionally led to unrealistic fits for some participants who showed spurious noncentral peaks in their guesses (e.g., resulting in a large bias, $\mu = -100°$, in combination with extremely high precision, $SD = 2°$). We therefore decided to fit a two-parameter model without the bias parameter. A two-tailed *t* test on probability of report (*Pmem*) with number of visual objects as within-subject variable revealed a highly significant drop in probability of report from one (51.9%) to two synchronized objects, 28.1%, $t(9) = 7.6$, $p < .0001$, a reduction of 46%. This also held under an arcsine transformation of the data (since these are proportions), $t(9) = 11.53$, $p < .0001$. At the same time, there was no effect on precision (*SD*), one object: 17.9°, two objects: 14.8°, $t(9) = 0.5$, $p = .582$.

Overall then, observers were better at registering the orientation of the line segment in a disk that changed together with the sound than in disks that did not change together with the sound. Thus, observers used the sound to specifically select the concurrent visual objects, rather than that it merely caused some general alerting or temporal cueing effect. However, selection of these synchronized objects appeared very limited. The probability of representing a synchronized object was reduced by almost half when the number of objects increased from one to two. In contrast, the precision was unaffected by the set size, indicating that *if* observers registered a synchronized disk, they gained full access to it and could determine the line orientation as if it was the only synchronized disk in the display. In all then, the data do not provide any evidence for a distribution of attention across multiple synchronized events. Instead, consistent with the single source hypothesis, attention appears to lock on to a single event, while the other visual event fails to become registered.

Experiment 2

Although the results of Experiment 1 are consistent with a severe limitation on how many visual objects can simultaneously benefit from a sound, there are alternative explanations. One possibility is that in this type of task, for one reason or another, observers are only able to attend to one object at a time anyway, regardless of the audiovisual nature of the events. For example, observers may have been able to *attend* to multiple synchronized disks and their orientations, but found it hard to *remember* them. Alternatively, observers may have been capable of attending to multiple disks, but they chose not to do so because the likelihood of being tested on one of them was only 50%. Although 50% validity here still meant that the sound provided useful information (as it reduced the potential target set from 12 to either one or two on half the trials), observers may not have perceived it that way, and so they spent relatively little effort in trying to extract more than a single event.

In Experiment 2, we made the sound-target relationship 100% valid. That is, whenever there was a sound, observers would be tested on a synchronized disk. This increased the sound's utility to 100%. Furthermore, we directly compared the performance for audiovisual cues to that for purely visual cues by including a condition in which the target disks were indicated directly by a brief color change, instead of by a concurrent sound. In that condition, the rim of either one or two disks briefly turned green. Validity and thus potential utility of these cues were exactly the same as for the auditory cue condition. To demonstrate that either type of cue (auditory or visual) was used by the observer, we also included a no cue baseline, in which observers received no signal as to which of the disks would be probed.

Overall, given the literature, we expected visual capacity to exceed audiovisual capacity (see also Van der Burg et al., 2013). We know from visual cueing and visual memory studies that attention can be distributed across multiple items, but that this leads to a reduction in representational resolution, as well as in an increased chance of items being missed, at least for the number of items used here (Bays & Husain, 2008; Huang, 2010; Wilken & Ma, 2004; Zhang & Luck, 2008), so here we expected effects on *both* SD and Pmem. The most important prediction again concerned the audiovisual condition. On the basis of the single source hypothesis, we predicted the same pattern as in Experiment 1: A reduction in probability of report, but, in contrast to the visual cue condition, no effect on precision.

Method

The procedure was identical to Experiment 1, except for the following changes. Twenty-four students (16 female; mean 21.7 years; range 19–30 years) from the VU University participated. All participants were naïve as to the purpose of the experiment. Instead of validity, we now manipulated the nature of the target signal (cue). The cue was either auditory, visual, or absent. The auditory condition was as in Experiment 1. In the target frame of the visual cue condition, the outline of either one or two disks became green for the duration of the frame (150 ms). In the cue absent condition, there was no accompanying target signal, and in the experience of the observer, any of the disks could thus be probed. No sound was presented in the visual cue and cue absent conditions. In contrast to Experiment 1, a probed item was now always a cued item.

The factors in the design were number of synchronized visual objects (one vs. two; mixed within blocks) and cue condition (visual, auditory, and absent; blocked, with participants receiving instructions on what cue to expect within in each block). Cue condition was presented in counterbalanced order of 15 experimental blocks of 36 trials each. Three practice blocks of 36 trials each were presented prior to the experiment, one for each cue condition. During practice, the SOA between display changes was 250 ms, to make participants accustom to the task.

Results and Discussion

Aggregate data. Figure 3 shows the main results for the data aggregated across participants. For auditory cue trials, the probability of report (*Pmem*) dropped significantly from one to two synchronized objects, one object: 58.4%, CrI: −2.7, +2.2; two objects: 35.6%, CrI: −2.9, +2.7; a reduction by 39%. Importantly, the precision (*SD*) was again not significantly different between the two conditions, one object: 11.8°, CrI: −0.6°, +0.8°; two objects: *SD* = 13.0°, CrI: −1.1°, +1.3°. Neither was there a systematic difference in bias (μ) between the two conditions, one object: −0.7°, CrI: −0.8°, +0.8°; two objects: −0.4°, CrI: −1.3°, +1.1°. We again conducted a post hoc analysis to examine potential grouping effects when two synchronized objects shared the same polarity, relative to when they differed in polarity. Precision was not significantly better in the similar condition (12.5, CrI: −1.3%, +3.3%) than in the dissimilar condition (13.3, CrI: −1.5%, +3.4%). For *Pmem,* there was a numerical benefit for when the two objects were similar (37.6%, CrI: −3.6%, +7.4%) compared with when the two synchronized objects were dissimilar (33.8%, CrI: −3.8%, +7.2%), although this was not reliable. Since numerically the difference went in the same direction as in Experiment 1, we collapsed the data from the two experiments, resulting *Pmem* = 35.4% (CrI: −3.2%, +3.2%) for same polarity trials, and *Pmem* = 31.2% (CrI: −2.9%, +3.1%) for different polarity trials, which, given the considerable overlap in CrIs, remained unreliable. There was no effect of similarity on precision either when the data were collapsed across experiments (same polarity: *SD* = 13.5, CrI: −1.4, +1.7; different polarity: *SD* = 13.1, CrI: −1.3, +2.1).

For visual cue trials, the probability of report also dropped significantly from one to two synchronized objects, one object: 88.9%, CrI: −1.5, +1.5, two objects: *Pmem* = 66.0%, CrI: −2.5, +2.6; a reduction by 26%. Importantly, in contrast to the auditory cue condition, there was now a reliable effect on precision, as it too dropped significantly from the one object condition

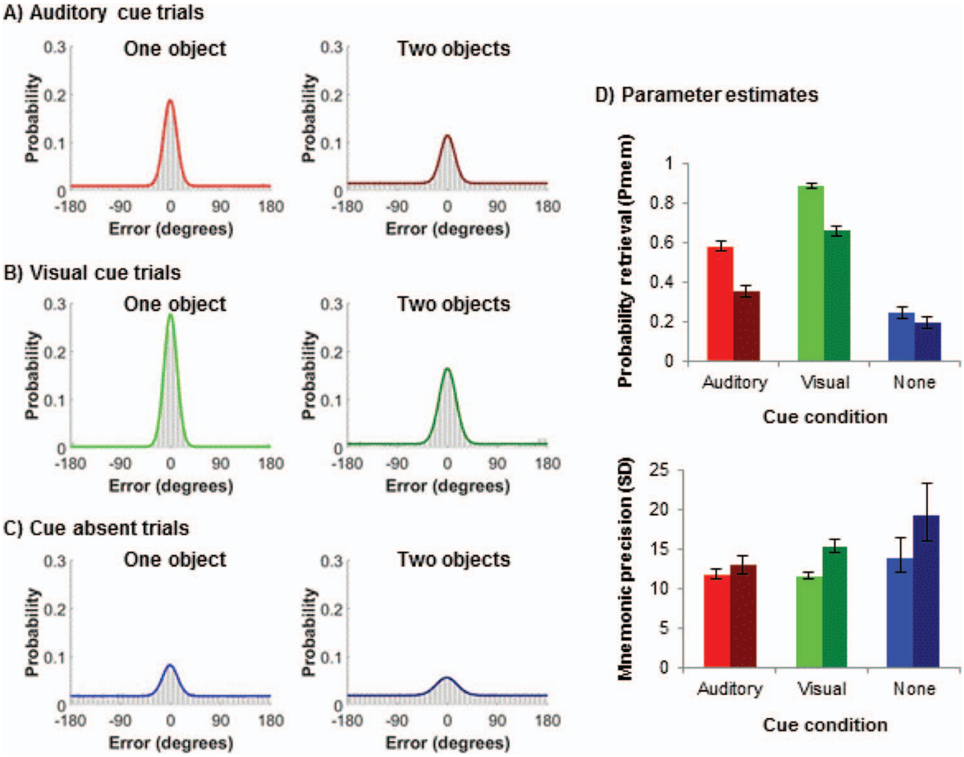

Figure 3. Results of Experiment 2. Mixture model fit to the aggregate data from all subjects for auditory cue trials (A), visual cue trials (B), and cue absent trials (C). The left panels show performance when one visual object was synchronized with the cue (if present), and the right panels show performance for two synchronized objects. Parameter estimates of probability of retrieval (Pmem) and mnemonic precision (*SD*) as a function of cue condition and the number of synchronized objects (D) Lighter shades represent one synchronized object, darker shades two synchronized objects. Error bars represent Bayesian credibility intervals of the fits. See the online article for the color version of this figure.

(11.7°,CrI: −0.4°, +0.5°) to the two objects condition (15.3°, CrI: −0.8°, +0.9°). Precision also became reliably worse than in any of the auditory cue conditions. There was no systematic bias, one object: −0.1°, CrI: −0.6°, +0.5°; two objects: −0.4°, CrI: −0.9°, +1.0°.

For the no cue trials, there were no effects on probability of report, one object: 24.5%, CrI: −2.9%, +2.8%; two objects: 19.7%, CrI: −3.1%, +2.9%, precision, one object: 13.8°, CrI: −1.7°, +2.6°; two objects: 19.3°, CrI: −3.3°, +3.9°, or bias, one object: −1.0°, CrI: −1.7°, +1.7°; two objects: −1.2°, CrI: −2.8°, +3.5°. As can be seen from Figure 3D, in terms of probability of report, performance was reliably worse than in the auditory and visual cues trials, with a trend in the same direction for precision.

Analyses on individual fits. The results for the individual fits were consistent with those for the aggregate data. An analysis of variance (ANOVA) on probability of report (*Pmem*), with cue condition (auditory, visual, and no cue) and number of visual objects (one vs. two) as within-subject variables revealed a main effect of cue condition, $F(2, 46) = 109.5$, $p < .0001$. Pairwise comparisons showed that on average, probability of report was best for the visual cue condition, worse for the auditory cue condition, and worst of all for the cue absent condition [for all comparisons, *t* values (23) > 5.5, *p* values < .0001]. Furthermore, probability of report was overall better for one cue than for two cues, $F(1, 23) = 90.4$, $p < .0001$. The Cue condition × Number of visual objects interaction was also significant, $F(2, 46) = 7.8$, $p < .002$, and further examined by separate two-tailed *t* tests for each cue condition. In the auditory cue condition, *Pmem* was significantly better when one disk (60.4%) was synchronized with the auditory signal than when two disks (36.4%) were synchronized, $t(23) = 7.7$, $p < .0001$. In the visual cue condition, *Pmem* was also better when one disk was cued (89.2%) than when two disks were cued (67.9%), $t(23) = 10.5$, $p < .0001$. In absolute terms, this reduction was similar for auditory and visual cue conditions, at 24 and 21 percentage points respectively, $t(23) = 0.79$, $p = .437$. In relative terms, auditory capacity was reduced by 37%, while visual capacity was reduced by 24%, a reliable difference, $t(23) = 2.81$, $p = .01$. No reliable effect of number of objects on *Pmem* was observed when no cue was presented, with *Pmem* being 29.8% for the one object condition and 23.5% in the two objects condition, $t(23)$ 1.4, $p = .169$. All these results also held when the analyses were performed on arcsine transformed data.

For the precision analyses, we only included the auditory cue and visual cue conditions, as for many participants overall probability of report was too low in the cue absent condition. An ANOVA with cue condition (auditory vs. visual) and number of visual objects as within-subject variables revealed a borderline significant main effect of number of objects, $F(1, 23) = 4.2$, $p = .050$. Most importantly, there was a reliable two-way interaction, $F(1, 23) = 9.9$, $p = .005$, which was further examined by separate two-tailed *t* tests for each cueing condition. In the auditory cue condition, there was no reliable effect of the number of visual objects on precision (*SD* = 14.8° when one visual object was synchronized with the sound, and *SD* = 14.3° when two visual objects were synchronized), $t(23) = .3$, $p = .738$. In contrast, in the visual condition, there was a highly reliable reduction in precision from one (*SD* = 11.8°) to two (*SD* = 16.9°) visually cued objects, $t(23) = 4.2$, $p < .0005$. The results of Experiment 2 replicate the main finding of Experiment 1.

Increasing the number of visual objects that were synchronized with the single auditory event (from one to two) resulted in a reduction of the probability of representing these visual objects, albeit by less than half (37% to 39%). At the same time, precision was maintained for items that were registered. Consistent with the single source hypothesis, there was thus no sign of attention being distributed across multiple events. In the visual cue condition, there was also an effect on probability, but in addition, there now was also a clear effect on precision. Despite the fact that probability of report was overall better for the visual cue condition than for the auditory cue condition, precision clearly deteriorated from one cued item to two. Consistent with previous work on visual working memory, in the visual cue condition attentional resources appear to be distributed across multiple items, such that even though multiple items can be encoded, this will occur at the expense of resolution. The observed decrease in probability of representing an item may also be a consequence of the distribution of resources, which are being spread so thinly that an item fails to reach the threshold of detection (although in the visual working memory literature this is still a matter of debate; e.g., Bays, Catalao, & Husain, 2009; Bays & Husain, 2008; Huang, 2010; Wilken & Ma, 2004; Zhang & Luck, 2008). Also noteworthy here is that the precision for single visual and single auditory cues was very much comparable. Thus, whether accessed through a single sound or through a single visual signal, the same representational resolution is the result.

General Discussion

The visual system is capable of using sound to prioritize synchronized visual objects in environments of strong visual competition (Ngo & Spence, 2010; Olivers & Van der Burg, 2008; Van den Brink et al., 2014; Van der Burg et al., 2008a, 2010, 2011, 2012, 2014; Vroomen & De Gelder, 2000; Zannoli, Cass, Mamassian, & Alais, 2012; Zou, Müller, & Shi, 2012), and the current data provide further evidence for this. Observers were able to select relevant flashing disks from among irrelevant flashing disks on the basis of a concurrent sound, as indicated by an increased probability of report. Selection improved relative to nonsynchronized objects (Experiment 1), and relative to when no sound was present (Experiment 2).

The important novel finding is that the *probability* of being able to report on a sound-synchronized visual object is reduced by about 40% when the number of visual objects increased from one to two, while at the same time, the *precision* of report was not at all affected. Taken together, this pattern of results goes against a mechanism by which attention is evenly distributed across multiple audiovisual events. Instead, it suggests that when one item is represented, it is registered in full (as indicated by intact precision), while the other item is not represented at all. By contrast, visual cues signaling which items to attend to resulted not only in decreased probability of report, but also in a reduction in precision, which is consistent with a splitting or spreading of attention across multiple simultaneous events, consistent with the literature on visual processing capacity (see also). The findings thus further dissociate between visual capacity and audiovisual capacity, as they show a different probability versus precision profile. They also further support the single source hypothesis of audiovisual integration (Van der Burg et al., 2013), which states that in normal environments, single sounds typically stem from single events, and hence the cognitive system may have little experience with, and no functional reason for, binding multiple visual objects to a sound.

Although a reduction in the probability of report by about 40% when doubling the number of synchronized objects from one to two fits with an approximate capacity of one item (which predicts 50%), we point out that overall the probability of representing the orientation of the line segment was quite low in the first place, at around 0.53 when a single visual object was synchronized with the sound in Experiment 1, and 0.63 for the same condition in Experiment 2. Also in the visual cue condition of Experiment 2, the probability of report was considerably below 1, at 0.87, dropping further to 0.64 for two cues. Although one would need to include larger set sizes to be able to measure the effective capacity, these values do suggest that, on average, even one object was often missed in these conditions. This limits the direct comparison between the audiovisual and the visual conditions in the current study, plus also the comparison to our previous study (Van der Burg et al., 2013). In that study, we used virtually the same displays and timing, but we only measured overall detection accuracy. Observers detected close to 90% of the single audiovisual events, and virtually 100% of the single visual events. A remaining question is why these values deviate so much from probability of report in the current study. Probably the most important difference is that in the current study, people had the additional task of registering and remembering the line segment inside the synchronized disk, whereas in our previous study, observers only had to localize the disk. Both perceiving and subsequent remembering of the line orientation may have suffered from the continuous changes in the display that occurred before and after the target frame. Some of these changes may themselves draw attention, and thus draw away resources from encoding and maintaining the target orientation in both the audiovisual and the purely visual conditions. In any case, since the mixture model treats the probability and precision of representation as independent, we do not see how this overall low probability of report could explain the differential effects on precision between the audiovisual and purely visual conditions. Nevertheless, future studies would probably do wise to try and equate the base rate of performance in audiovisual and visual conditions.

Remaining Questions and Limitations

The current findings raise important new questions for the future. If a single sound can only bind to a single visual event, one relevant issue is then what exactly counts as a single visual event. Our study is based on an intuitive notion that the disks were each regarded as individual objects, and their accompanying transients thus as separate events. This definition may be too limited. Multiple visual stimuli may group on the basis of properties such as proximity, color, shape, or common dynamics. Could such grouped events count as a single event for mechanisms of audiovisual integration? In fact, one might a priori have assumed the current target stimuli to group already on the basis of their common transient (Donk & Theeuwes, 2001; Jiang, Chun, & Marks, 2002; Olivers & Humphreys, 2004; Pinto, Olivers, & Theeuwes, 2008). The reason why such grouping by transients was apparently not so strong here may be that the two crucial events were embedded in a continuous stream of random events, with one transient camouflaging the other (Cass & Van der Burg, 2014). In fact, we explicitly designed the stimuli this way, so that the visual transients would draw little attention by themselves, without the sound. For the same reason grouping by polarity may also have been weak to absent, as we found in our analyses. Nevertheless, the possibility remains that when there is no such competition from surrounding transients, multiple simultaneous visual events can integrate with the sound. A recent report by Kawachi, Grove, Sakurai (2014) appears consistent with this. As a starting point, they took the stream/bounce phenomenon in which two moving stimuli that cross paths appear to be bouncing off each other if the moment of crossing is accompanied by a sound (Sekuler, Sekuler, & Lau, 1997). They found that participants reported seeing a bounce also for *two* sets of crossing stimuli presented at or around the same time. There are many differences between their type of experiment and ours, but one of them is that the crosses/bounces were the only events happening in the display in which case they may either have grouped more easily, or both events may already have been attended prior to the bounce. In another set of studies, we are currently investigating whether performance increases when two synchronized events can be grouped on the basis of similarity, specifically whether they change polarity in the same direction (i.e., from black to white or from white to black) or not. As far as we can tell from the results we have, there is little to no benefit for same polarity changes, suggesting that the audiovisual integration we are measuring here is not sensitive to grouping by visual similarity. Proximity may be a more promising candidate, such that two immediately neighboring visual events may both benefit from a sound—but this remains to be seen.

If, under conditions of competition, a single sound can indeed only point to just one of multiple visual events, then this raises another question, namely as to which visual event is preferred. We currently lack a mechanistic explanation, but there appear to be a number of possibilities. One is that the sound binds to whichever at that moment is the most salient visual event. Relative salience may be determined by display properties, or by stochastic changes in the neural system. For example, a concurrent sound may increase the gain on visual processing, affecting the saliency landscape of visual representations (Itti & Koch, 2000) in such a way that the strongest representation becomes even stronger, in a winner take all manner. Another possibility is that the sound chooses to bind to the visual event that is closest to the current focus of attention (which will often coincide with the current fixation location). Our current experiments are limited in that they cannot decide between these or any other alternatives. One initial piece of evidence that the current attentional focus may determine which item is allowed to bind to the sound comes from Van der Burg et al. (2012). Using a visual search paradigm, they found that the extent to which a sound boosts the detection of a concurrent visual event is modulated by how widely distributed attention is. Detection of audiovisual targets improved when preceded by a task that required distributed spatial attention (i.e., identifying a very large background letter encompassing the entire search display) compared with when they were preceded by a task that required focused attention (i.e., identifying a small letter at central fixation). Thus, the audiovisual integration benefited from the visual information being attended, in line with earlier work suggesting that at least some forms of audiovisual integration depend on attention (Busse, Roberts, Crist, Weissman, & Woldorff, 2005; Talsma & Woldorff, 2005; see Talsma, Senkowski, Soto-Faraco, & Woldorff, 2010, for a review).

Furthermore, the finding that a single sound cannot boost multiple visual events begs the question as to whether *multiple* sounds can do so. In principle, the single source hypothesis leaves open the possibility that multiple concurrent sounds can boost multiple concurrent visual events, simply because multiple sounds may each have their own source. In this respect, our experiments were limited, and future studies would be needed to investigate the effect of multiple sounds. However, technically this presents some challenges, specifically the fact that the common onset of multiple sounds in itself provides a nonaccidental signal that one is likely dealing with just a single event. Perceptually speaking, concurrent sounds are therefore typically grouped into a single rich percept. We are currently developing ways to tear these mechanisms apart, to see if audiovisual capacity can be taken beyond a single event.

References

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology, 14,* 257–262. http://dx.doi.org/10.1016/j.cub.2004.01.029

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. http://dx.doi.org/10.1111/j.1467-9280.2007.01949.x

Bays, P. M., Catalao, R. F., & Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision, 9,* 1–11. http://dx.doi.org/10.1167/9.10.7

Bays, P. M., & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science, 321,* 851–854. http://dx.doi.org/10.1126/science.1158023

Bedford, F. (2001). Towards a general law of numerical/object identity. *Current Psychology of Cognition, 20,* 113–176.

Burkell, J. A., & Pylyshyn, Z. W. (1997). Searching through subsets: A test of the visual indexing hypothesis. *Spatial Vision, 11,* 225–258. http://dx.doi.org/10.1163/156856897X00203

Busse, L., Roberts, K. C., Crist, R. E., Weissman, D. H., & Woldorff, M. G. (2005). The spread of attention across modalities and space in a multisensory object. *Proceedings of the National Academy of Sciences of the United States of America, 102,* 18751–18756. http://dx.doi.org/10.1073/pnas.0507704102

Cass, J., & Van der Burg, E. (2014). Remote temporal camouflage: Contextual flicker disrupts perception of time's arrow. *Vision Research, 103,* 92–100.

Cowan, N. (2010). The magical mystery four: How is working memory capacity limited, and why? *Current Directions in Psychological Science, 19,* 51–57. http://dx.doi.org/10.1177/0963721409359277

Dalton, P., & Spence, C. (2007). Attentional capture in serial audiovisual search tasks. *Perception & Psychophysics, 69,* 422–438. http://dx.doi.org/10.3758/BF03193763

Donk, M., & Theeuwes, J. (2001). Visual marking beside the mark: Prioritizing selection by abrupt onsets. *Perception & Psychophysics, 63,* 891–900. http://dx.doi.org/10.3758/BF03194445

Franconeri, S. L., Alvarez, G. A., & Enns, J. T. (2007). How many locations can be selected at once? *Journal of Experimental Psychology: Human Perception and Performance, 33,* 1003–1012. http://dx.doi.org/10.1037/0096-1523.33.5.1003

Huang, L. (2010). Visual working memory is better characterized as a distributed resource rather than discrete slots. *Journal of Vision, 10,* 8. http://dx.doi.org/10.1167/10.14.8

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research, 40,* 1489–1506. http://dx.doi.org/10.1016/S0042-6989(99)00163-7

Jack, C. E., & Thurlow, W. R. (1973). Effects of degree of visual association and angle of displacement on the "ventriloquism" effect. *Perceptual and Motor Skills, 37,* 967–979. http://dx.doi.org/10.2466/pms.1973.37.3.967

Jiang, Y., Chun, M. M., & Marks, L. E. (2002). Visual marking: Selective attention to asynchronous temporal groups. *Journal of Experimental Psychology: Human Perception and Performance, 28,* 717–730.

Kawachi, Y., Grove, P. M., & Sakurai, K. (2014). A single auditory tone alters the perception of multiple visual events. *Journal of Vision, 14,* 16. http://dx.doi.org/10.1167/14.8.16

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal inference in multisensory perception. *PLoS ONE, 2,* e943. http://dx.doi.org/10.1371/journal.pone.0000943

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature, 390,* 279–281. http://dx.doi.org/10.1038/36846

Luck, S. J., Woodman, G. F., & Vogel, E. K. (2000). Event-related potential studies of attention. *Trends in Cognitive Sciences, 4,* 432–440.

MacDonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. *Perception & Psychophysics, 24,* 253–257. http://dx.doi.org/10.3758/BF03206096

Mandler, G., & Shebo, B. J. (1982). Subitizing: An analysis of its component processes. *Journal of Experimental Psychology: General, 111,* 1–22. http://dx.doi.org/10.1037/0096-3445.111.1.1

Ngo, M. K., & Spence, C. (2010). Auditory, tactile, and multisensory cues facilitate search for dynamic visual stimuli. *Attention, Perception, & Psychophysics, 72,* 1654–1665. http://dx.doi.org/10.3758/APP.72.6.1654

Olivers, C. N., & Humphreys, G. W. (2004). Spatiotemporal segregation in visual search: Evidence from parietal lesions. *Journal of Experimental Psychology: Human Perception and Performance, 30,* 667–688. http://dx.doi.org/10.1037/0096-1523.30.4.667

Olivers, C. N., & Van der Burg, E. (2008). Bleeping you out of the blink: Sound saves vision from oblivion. *Brain Research, 1242,* 191–199. http://dx.doi.org/10.1016/j.brainres.2008.01.070

Pashler, H. (1988). Familiarity and visual change detection. *Perception & Psychophysics, 44,* 369–378. http://dx.doi.org/10.3758/BF03210419

Pinto, Y., Olivers, C. N., & Theeuwes, J. (2008). The detection of temporally defined objects does not require focused attention. *The Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 61,* 1134–1142. http://dx.doi.org/10.1080/17470210701851198

Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision, 3,* 179–197. http://dx.doi.org/10.1163/156856888X00122

Roseboom, W., Nishida, S., & Arnold, D. H. (2009). The sliding window of audio-visual simultaneity. *Journal of Vision, 9,* 4. http://dx.doi.org/10.1167/9.12.4

Roseboom, W., Nishida, S., Fujisaki, W., & Arnold, D. H. (2011). Audio-visual speech timing sensitivity is enhanced in cluttered conditions. *PLoS ONE, 6,* e18309. http://dx.doi.org/10.1371/journal.pone.0018309

Schroeder, C. E., & Foxe, J. (2005). Multisensory contributions to low-level, "unisensory" processing. *Current Opinion in Neurobiology, 15,* 454–458. http://dx.doi.org/10.1016/j.conb.2005.06.008

Sekuler, R., Sekuler, A. B., & Lau, R. (1997). Sound alters visual motion perception. *Nature, 385,* 308–308. http://dx.doi.org/10.1038/385308a0

Senkowski, D., Talsma, D., Grigutsch, M., Herrmann, C. S., & Woldorff, M. G. (2007). Good times for multisensory integration: Effects of the precision of temporal synchrony as revealed by gamma-band oscillations. *Neuropsychologia, 45,* 561–571. http://dx.doi.org/10.1016/j.neuropsychologia.2006.01.013

Soto-Faraco, S., Kingstone, A., & Spence, C. (2003). Multisensory contributions to the perception of motion. *Neuropsychologia, 41,* 1847–1862. http://dx.doi.org/10.1016/S0028-3932(03)00185-4

Spence, C., & Squire, S. (2003). Multisensory integration: Maintaining the perception of synchrony. *Current Biology, 13,* R519–R521. http://dx.doi.org/10.1016/S0960-9822(03)00445-7

Stein, B. E., & Meredith, M. A. (1993). *The merging of the senses.* Cambridge, MA: MIT Press.

Suchow, J. W., Brady, T. F., Fougnie, D., & Alvarez, G. A. (2013). Modeling visual working memory with the MemToolbox. *Journal of Vision, 13,* 9. http://dx.doi.org/10.1167/13.10.9

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26,* 212–215. http://dx.doi.org/10.1121/1.1907309

Talsma, D., Senkowski, D., Soto-Faraco, S., & Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. *Trends in Cognitive Sciences, 14,* 400–410. http://dx.doi.org/10.1016/j.tics.2010.06.008

Talsma, D., & Woldorff, M. G. (2005). Selective attention and multisensory integration: Multiple phases of effects on the evoked brain activity. *Journal of Cognitive Neuroscience, 17,* 1098–1114. http://dx.doi.org/10.1162/0898929054475172

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. http://dx.doi.org/10.1037/0033-295X.101.1.80

van den Brink, R. L., Cohen, M. X., van der Burg, E., Talsma, D., Vissers, M. E., & Slagter, H. A. (2014). Subcortical, modality-specific pathways contribute to multisensory processing in humans. *Cerebral Cortex, 24,* 2169–2177. http://dx.doi.org/10.1093/cercor/bht069

Van der Burg, E., Alais, D., & Cass, J. (2013). Rapid recalibration to audiovisual asynchrony. *The Journal of Neuroscience, 33,* 14633–14637. http://dx.doi.org/10.1523/JNEUROSCI.1182-13.2013

Van der Burg, E., Awh, E., & Olivers, C. N. (2013). The capacity of audiovisual integration is limited to one item. *Psychological Science, 24,* 345–351. http://dx.doi.org/10.1177/0956797612452865

Van der Burg, E., Cass, J., & Alais, D. (2014). Window of audio-visual simultaneity is unaffected by spatio-temporal visual clutter. *Scientific Reports, 4,* 5098. http://dx.doi.org/10.1038/srep05098

Van der Burg, E., Cass, J., Olivers, C. N., Theeuwes, J., & Alais, D. (2010). Efficient visual search from synchronized auditory signals requires transient audiovisual events. *PLoS ONE, 5,* e10664. http://dx.doi.org/10.1371/journal.pone.0010664

Van der Burg, E., Olivers, C. N., Bronkhorst, A. W., & Theeuwes, J. (2008a). Audiovisual events capture attention: Evidence from temporal order judgments. *Journal of Vision, 8,* 2–10. http://dx.doi.org/10.1167/8.5.2

Van der Burg, E., Olivers, C. N., Bronkhorst, A. W., & Theeuwes, J. (2008b). Pip and pop: Nonspatial auditory signals improve spatial visual search. *Journal of Experimental Psychology: Human Perception and Performance, 34,* 1053–1065. http://dx.doi.org/10.1037/0096-1523.34.5.1053

Van der Burg, E., Olivers, C. N., & Theeuwes, J. (2012). The attentional window modulates capture by audiovisual events. *PLoS ONE, 7,* e39137. http://dx.doi.org/10.1371/journal.pone.0039137

Van der Burg, E., Talsma, D., Olivers, C. N., Hickey, C., & Theeuwes, J. (2011). Early multisensory interactions affect the competition among multiple visual objects. *NeuroImage, 55,* 1208–1218. http://dx.doi.org/10.1016/j.neuroimage.2010.12.068

Vatakis, A., & Spence, C. (2007). Crossmodal binding: Evaluating the “unity assumption” using audiovisual speech stimuli. *Perception & Psychophysics, 69,* 744–756.

Vroomen, J., & de Gelder, B. (2000). Sound enhances visual perception: Cross-modal effects of auditory organization on vision. *Journal of Experimental Psychology: Human Perception and Performance, 26,* 1583–1590. http://dx.doi.org/10.1037/0096-1523.26.5.1583

Welch, R. B. (1999). Meaning, attention, and the “unity assumption” in the intersensory bias of spatial and temporal perceptions. In G. Aschersleben, T. Bachmann, & J. Müsseler (Eds.), *Cognitive contributions to the perception of spatial and temporal events* (pp. 371–387). Amsterdam, the Netherlands: Elsevier.

Welch, R. B., & Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin, 88,* 638–667.

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision, 4,* 1120–1135. http://dx.doi.org/10.1167/4.12.11

Yantis, S., & Johnson, D. N. (1990). Mechanisms of attentional priority. *Journal of Experimental Psychology: Human Perception and Performance, 16,* 812–825. http://dx.doi.org/10.1037/0096-1523.16.4.812

Zannoli, M., Cass, J., Mamassian, P., & Alais, D. (2012). Synchronized audio-visual transients drive efficient visual search for motion-in-depth. *PLoS ONE, 7,* e37190. http://dx.doi.org/10.1371/journal.pone.0037190

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature, 453,* 233–235. http://dx.doi.org/10.1038/nature06860

Zou, H., Müller, H. J., & Shi, Z. (2012). Non-spatial sounds regulate eye movements and enhance visual search. *Journal of Vision, 12,* 2–18. http://dx.doi.org/10.1167/12.5.2

Received April 19, 2015
Revision received April 11, 2016
Accepted May 11, 2016 ■