

## **The Impact of Crossmodal Correspondences on Working Memory Performance**

Riccardo Brunetti\*, Allegra Indraco\*, Serena Mastroberardino°,

Charles Spence\*, & Valerio Santangelo^

\* Department of Human Sciences, Università Europea di Roma, Italy

° Neuroimaging Laboratory, Santa Lucia Foundation, Rome, Italy

• Department of Experimental Psychology, Oxford University, UK

^ Department of Philosophy, Social Sciences & Education, University of Perugia,  
Italy

Running head: Crossmodal Correspondences in Working Memory

Submitted as regular article to: *Journal of Experimental Psychology: Human,  
Perception & Performance*

Corresponding authors:

R.B., Department of Human Sciences, Università Europea di Roma, Via degli  
Aldobrandeschi 190, 00163 Roma, Italy, [riccardo.brunetti@gmail.com](mailto:riccardo.brunetti@gmail.com); V.S.,  
Department of Philosophy, Social Sciences & Education, University of Perugia,  
Piazza G. Ermini 1, 06123 Perugia, Italy; [valerio.santangelo@unipg.it](mailto:valerio.santangelo@unipg.it)

## ABSTRACT

Crossmodal correspondences influence perceptual performance in adults, infants, and even non-human primates across a variety of different sensory modalities and dimensions, [including in](#) tasks involving speeded detection, discrimination, and categorization. However, to date, it is still unclear whether and how correspondences could modulate post-perceptual processes, such as working memory (WM). Here we investigated this issue using a bimodal (audiovisual) 2-back task. In Experiment 1, three kinds of correspondences were used: audio/visual numerosity, pitch/shape, and pitch/elevation, each presented congruently (e.g., for numerosity: three auditory tones along with three visual shapes) or incongruently (3tones/2shapes). Participants attended to the visual, [or](#) auditory, ~~or both~~ modalities, [or both](#), simultaneously. The results revealed faster target-detection latencies following congruent as compared to incongruent stimulation, especially for numerosity congruence. In Experiment 2, we therefore focused on numerosity, varying the correspondence of the unattended modality, thus having correspondences at both sample (e.g., 3tones/3shapes) and target (e.g., 3tones/3shapes), only at sample (sample: 3tones/3shapes; target: 3tones/2shapes), only at target (sample: 3tones/2shapes; target: 3tones/3shapes), or never. [In order to](#) investigate the format of the encoded information we also included “symbolic” quantities (i.e., visually/auditorily-presented digits). The results confirmed the facilitation of correspondences on WM performance, highlighting that the effect arises when the correspondence [is](#) at the target display, thus affecting response selection. Moreover, the experiment revealed modal differences in the effect, showing how symbolic information affects the auditory and visual modalities differently. Overall, these findings highlight the impact of crossmodal

Commented [CS1]: operates

correspondences on WM processing, adding new light on the link between perceptual and post-perceptual stages of human information processing.

Keywords: Crossmodal correspondences; Working memory; Bimodal N-back; Multisensory integration; Crossmodal binding; Modal/Amodal representations.

## INTRODUCTION

The last decade has seen a rapid growth of interest in the study of crossmodal correspondences (CCs). CCs reflect the tendency to associate seemingly arbitrary features or dimensions of stimuli across the senses (e.g., Spence, 2011; Spence & Deroy, 2013). The existence of such effects have been known about for well over a century now (e.g., see Stumpf, 1883, for one early report), the most famous probably being the ubiquitous association between the nonsense words *maluma* and *takete* and rounded/pointy shapes, respectively (see Köhler, 1929, 1947; evident also cross-culturally, see Hinton, Nichols, & Ohala, 2006).<sup>1</sup> To date, numerous different CCs have been documented (see Spence, 2011, for a review), acknowledging their presence in adults, infants, and even [in](#) non-human primates (see Ludwig, Adachi, & Matzuzawa, 2011).

Over the years, researchers have investigated the influence of CCs on performance in a wide variety of behavioral tasks, including speeded detection, discrimination, and categorization tasks, with the mapping of auditory pitch onto visual dimensions including elevation (Bernstein & Edelstein, 1971; Evans & Treisman, 2010; Melara & O'Brien, 1987), visual size (Evans & Treisman, 2010; Parise & Spence, 2009), and even visual shape (Marks, 1987; Parise & Spence, 2012; see Marks, 2004, for a review). These kinds of CCs are *modal* (Spence, 2011), in that they are based on sensory-specific dimensions, in contrast to the so-called *amodal* correspondences, that are based on common stimulus properties, such as duration, temporal pattern, intensity, and numerosity (Féron, Gentaz, & Streri, 2006;

---

<sup>1</sup> More recent replications of this phenomenon have tended to use the words *bouba/kiki* instead (see Bremner, Caparos, Davidoff, de Fockert, Linnell, & Spence 2013; Ramachandran & Hubbard, 2001, 2003).

Lewkowicz & Turkewitz, 1980; Spence, 2011; see also Marks, Szczesiul, & Ohlott, 1986).<sup>2</sup>

Despite the robust evidence that has been collected to date concerning the perceptual effects of CCs (for reviews, see Spence, 2011; Spence & Deroy, 2013), the possible influences of CCs on post-perceptual information processing have not, as yet, been [explored](#) extensively—~~explored~~. The results of such studies, would, however, likely be informative when it comes to trying to clarify the mechanisms underlying the putatively different kinds of CC. For instance, Chiou and Rich (2012) demonstrated that CCs affect the allocation of exogenous spatial attention. Analogously, Klapetek, Ngo, and Spence (2012) have documented enhanced target detection performance when a visual target was paired with a spatially uninformative but crossmodally congruent sound, presented along with the visual search display. More recently, the spatial (exogenous) attentional shifts elicited by a CC (i.e., visual shape / sound type) have been shown to affect working memory (WM) performance<sup>3</sup>, but only when the interval between the presentation of the memory sample and probe was short (150 msec; Makovac, Kwok, & Gerbino, 2014). By contrast, with intervals in excess of 1150 msec, the presentation of the CC resulted in decreased WM performance, thus highlighting the need for further confirmation concerning the impact of CCs on WM performance.

Previous WM studies have shown that both intramodal and bimodal

---

<sup>2</sup> It should, though, be noted that the utility of the concept of *amodality* has been questioned by some researchers (cf. Spence, Deroy, & Bremner, 2013). The suggestion being that it is not altogether clear exactly under which criteria we may define a feature as amodal.

<sup>3</sup> Note that CCs have [also](#) been ~~also~~-related to memory by means of the Implicit Association Test (IAT: see Parise & Spence, 2012), which is based on existing (long-term memory) associations (Greenwald, McGhee, & Schwartz, 1998).

congruence might facilitate performance in the n-back task. A recent study on grapheme-color synesthesia revealed that when participant had to detect the grapheme presented earlier in a 2 or a 3-back task, the congruence between grapheme and color improved the performance of both synesthetes and non-synesthetes (Terhune, Wudarczyk, Kochuparampil, & Kadosh, 2013). A beneficial effect of bimodal congruence on WM has also been shown by Santangelo, Mastroberardino, Botta, Marucci, and Olivetti Belardinelli (2006). Their results highlighted an advantage for the bimodal presentation format as compared to the unimodal one, during a 2-back task (see also Mastroberardino, Santangelo, Botta, Marucci, & Olivetti Belardinelli, 2008).

The present study was designed to confirm and to further explore the notion that CCs can affect WM performance. We investigated for the first time the effect that different types of CCs can have on WM processing. Furthermore, we investigated whether this effect remains stable when attending to different sensory modalities (visual, auditory, or audiovisual). After investigating whether CCs do indeed affect performance in an n-back task, our second experiment was designed to assess whether the effects of CCs on WM could be attributed to a merely perceptual/attentional/response facilitation or to an effect on mnemonic enhancement. Experiment 2 also investigated the **format of CC information encoded into WM by including “symbolic” quantities (i.e., visually/auditorily-presented digits; see below).**

Moreover, we assessed the reliability of CCs in improving WM performance using a more complex and ecologically-valid scenario as compared to the previous literature, based on continuous WM performance, as measured by the popular n-back task (Kirchner, 1958). This investigation might be particularly useful when it comes to trying to account for the link between the effect of CCs at the perceptual and post-

perceptual stages of information processing. Assessing the impact of CCs on WM performance can potentially shed new light on the initial stages of memory representation and (long-term) consolidation of CCs, which, in turn, will impact the perceptual level.

Our hypothesis was that audiovisual CCs would enhance participants' performance in the n-back task, improving their ability and readiness to identify the stimuli (e.g., Makovac et al., 2014). On the basis of the literature (e.g., Bernstein & Edelstein, 1971; Evans & Treisman, 2010; Féron, Gentaz, & Streri, 2006; Marks, 1987; Melara & O'Brien, 1987; Parise & Spence, 2009, 2012), we focused on three audiovisual CCs that have been [studied](#) widely—~~studied~~: pitch/shape (e.g., high pitch/sharp shape –“kiki” or low pitch/round shape –“bouba”), pitch/elevation (e.g., high pitch/top elevation or low pitch/bottom elevation), and audio/visual numerosity (e.g., number of sinusoidal tones/number of visual shapes). The stimuli were divided into 8 different categories (each category was represented through 6 possible instances), according to the specific crossing of these three types of crossmodal matching (see Table 1 and Figure 1). The participants had to focus on one sensory modality (and thus the other modality was task-irrelevant) or to focus on both modalities at the same time. This aspect of the experimental design was implemented to verify whether the three CCs under study showed symmetrical facilitation effects according to the attended modality. As would be expected based on recent “perceptual” findings, our hypothesis was that auditory and visual modality would be symmetrically affected by CC (see, e.g., Parise & Spence, 2012; even if the degree of facilitation were to vary between the senses; see Evans & Treisman, 2010). Moreover, attending to both sensory modalities may be expected to affect performance differently (e.g., enhanced or attenuated facilitation) than when attention is directed to

a single sensory modality at a given time (Mozolic, Hugenschmidt, Peiffer, & Laurienti, 2008). This would offer evidence concerning the automaticity of such effects (e.g., it would provide evidence concerning their dependence on the specific intentions of the participant; see Spence & Deroy, 2013, on the goal-independence criterion of automaticity). According to the extant literature in the perceptual domain, a more general prediction would be related to the overall enhancement of WM performance following congruent vs. incongruent bimodal stimulation.

## EXPERIMENT 1

### METHODS

#### PARTICIPANTS

Sixty-eight undergraduate students (38 females; mean age = 22.0 years; SD = 2.0; range = 19-26 years) took part in the study for course credit. All of the participants reported being right-handed, with normal or corrected-to-normal vision, and normal hearing. All of the participants were naïve as to the purpose of the study. This research has been approved by the ethics review board of Università Europea.

#### APPARATUS and MATERIALS

Stimulus presentation, conditions, pseudo-randomization, and the recording of responses were all controlled by a custom-made script in the MAX programming environment (Cycling'74 - ver. 6), running on a 15" 2.4 Ghz MacBook Pro laptop computer. The sounds were presented through AKG K171 headphones at a



comfortable listening level. Bimodal (i.e., auditory and visual) stimuli were presented in a 2-back task (Kirchner, 1958). The auditory component of the bimodal stimuli consisted of sequences of either one, two, or three sinusoidal tones (each tone 120 msec in duration, including two 5 msec fading ramps separated by 120 msec pauses in sequences of two or three tones), presented at either a low (300Hz) or high (4500Hz) frequency, thus giving rise to a total of 6 different sounds (that is, 3 sequences x 2 frequencies). The one, two, and three tone sequences had a total duration of 120, 360, and 600 msec, respectively. The visual component of the bimodal stimuli consisted of one to three rounded ("bouba") and angular ("kiki") shapes, placed at the top or at the bottom of the display (see Figure 1; the specific shapes were taken from Ramachandran & Hubbard, 2001). Hence, we used a total of 12 different visual stimuli (i.e., 2 shapes x 3 numerosities x 2 visual elevations): One, two, or three "bouba" shapes were placed at either the top or bottom of the screen; One, two, or three "kiki" shapes were also placed at either the top or at the bottom of the screen. Each visual stimulus had a size of  $3.95^\circ \times 4.62^\circ$  and was presented in grey (50%) against a white background. To summarize, two dimensions were manipulated in the auditory modality (i.e., pitch and numerosity) while three dimensions were manipulated in the visual modality (shape, numerosity, and visual elevation), resulting in 8 categories. Two of these categories involved the full matching or mismatching of the auditory and visual dimensions (i.e., auditory pitch and numerosity; visual elevation, shape and numerosity; see Figure 1A). They were used to assess the overall impact of CCs on performance in the n-back task, irrespective of the type of CC: full-match (PENS<sub>c</sub>– Pitch Elevation Numerosity Shape congruence), where all the visual dimensions were congruent with the auditory dimensions; full-mismatch (PENS<sub>i</sub>–

Pitch Elevation Numerosity Shape incongruence), where all the visual dimensions were incongruent with the auditory dimensions.

\*\*\*\*\* Insert Table 1 & Figure 1 about here \*\*\*\*\*

## DESIGN and PROCEDURE

A bimodal audio-visual 2-back task was designed in which the participants had to detect when the current stimulus matched the one that had been presented 2 steps earlier in the sequence (Kirchner, 1958). We used semi-randomized sequences in which we excluded repetitions at 1- or 3-back in order to avoid confusing the participants. The 2-back task was administered in 4 blocks, of 160 trials each, with a 30% target ratio (48 per block). Each block contained an equiprobable number of instances from each stimulus category (each category featured 6 possible instances, each repeated once per block, giving rise to a total of 4 repetitions for each instance). Each block started with the presentation of a fixation cross (1400 msec) and continued with a sequence of synchronized audiovisual stimulus pairs. Each visual stimulus had a duration of 600 msec, followed by a 1400 msec fixation cross, resulting in an inter-stimulus interval of 2000 msec. While the duration of the auditory stimuli varied, their onset was always synchronous with that of the paired visual stimuli. The targets were always a perfect bimodal repetition of the stimulus that had been presented 2 stimuli earlier in the sequence. While the trials were always presented bimodally, the participants were randomly divided into three groups. Each group received the instruction to perform a different task. The first group ( $n = 24$ ) was instructed to perform the 2-back task focusing only on the auditory dimension (thus meaning that the visual dimension was task-irrelevant); the second group ( $n = 22$ ) had to focus only

on visual information (thus meaning that the auditory dimension was task-irrelevant); the third group ( $n = 22$ ) was instructed to pay attention to both stimulus dimensions at the same time. After having received their instructions, the participants completed a training session of 40 trials, with error feedback. No error feedback was provided during the experimental trials. All three groups of participants were instructed to detect target stimuli (visually, auditorily, or bimodally according to the group that they had been assigned to), by pressing the spacebar on the computer keyboard, as rapidly and accurately as possible.

## RESULTS

For the purposes of this paper, only the results of categories PENS<sub>c</sub>, PENS<sub>i</sub>, PE<sub>c</sub>, N<sub>c</sub>, and PSc were considered for the analyses.

### *Reaction times*

The main aim of Experiment 1 was to assess the impact of CCs on participants' performance in the 2-back task. Accordingly, in our first analysis, we compared target categories where visual dimensions were overall congruent with the auditory dimensions (i.e., PENS<sub>c</sub> trials) with target categories in which they were overall incongruent (i.e., PENS<sub>i</sub> trials). A 3x2 mixed ANOVA was conducted with the between-participants factor of "Attended modality" (auditory, visual, or both) and the within-participants factor of full congruent / incongruent categories. Importantly, this analysis revealed a significant main effect of "Category" [ $F(1, 65) = 31.242$ ;  $p < .001$ ;  $\eta^2 = .325$ ], indicating that the participants detected the targets more rapidly when the stimuli were fully crossmodally congruent (725 ms) than when they were fully

crossmodally incongruent (773 ms). The analysis also revealed a significant main effect of “Attended modality” [ $F(2, 65) = 15.504$ ;  $p < .001$ ;  $\eta^2 = .323$ ], with participants responding significantly more rapidly when attending to the visual modality than when attending to the auditory modality. The performance of the audiovisual group fell in-between that of the two other groups. The ANOVA did not reveal any interaction between the two factors [ $F(2, 65) = 1.342$ ;  $p = .268$ ;  $\eta^2 = .040$ ].

Moreover, in order to understand the specific effect of these individual categories, we conducted a 3x4 mixed ANOVA with the between-participants factor of “Attended modality” (auditory, visual, or both) and the within-participants factor of “Category” (PEc, Nc, PSc, PENSi). The analysis revealed a significant main effect of “Attended modality” [ $F(2,65)=18.437$ ;  $p<.001$ ;  $\eta^2=.362$ ], a significant main effect of “Category” [ $F(3, 65)=4.934$ ;  $p=.003$ ;  $\eta^2=.071$ ], and a significant interaction between these two factors [ $F(6, 195)=2.797$ ;  $p=.012$ ;  $\eta^2=.079$ ]. LSD post-hoc analyses revealed that when attending to the auditory modality, the participants were significantly faster in the PEc and Nc categories than PENSi ( $p=.045$  and  $p=.001$ , respectively), and also with respect to PSc ( $p=.023$  and  $p<.001$ , respectively). When attending to the visual modality, the participants were faster in the PSc category than in PENSi ( $p=.046$ ).

\*\*\*\*\* Insert Table 2 & Figure 2 about here \*\*\*\*\*

#### *Proportion correct*

A similar analysis of the proportion of correct answers (PC) data compared target categories when individual dimensions were globally crossmodally congruent with target categories in which they were overall incongruent. We carried out a 3x2 mixed ANOVA with the between-participants factor of Attended modality (auditory,

Commented [CS2]: in the

Commented [CS3]: still not quite clear

Commented [CS4]: the

Commented [CS5]: category

visual, or both) and the within-participants factor of full congruent/incongruent categories. This analysis revealed a marginally significant main effect of “Category” [ $F(1, 65) = 3.549$ ;  $p = .064$ ;  $\eta^2 = .052$ ], with participants responding more accurately in the full crossmodally congruent condition (PC scores = .80) than in the full crossmodally incongruent condition (PC scores = .77). The ANOVA revealed neither a significant main effect of “Attended modality” [ $F(2, 65) = 2.484$ ;  $p = .091$ ;  $\eta^2 = .071$ ], nor any interaction between the two factors [ $F(2, 65) = .056$ ;  $p = .946$ ;  $\eta^2 = .002$ ].

A 3x4 mixed ANOVA with the between-participants factor of “Attended modality” (auditory, visual, or both) and the within-participants factor of “Category” (PEc, Nc, PSc, and PENSi). The analysis revealed a significant interaction between these two factors [ $F(6, 195) = 2.693$ ;  $p = .016$ ;  $\eta^2 = .077$ ]. LSD post-hoc analyses revealed that when participants attended to the auditory modality, they were significantly more accurate in the PEc category than the PSc ( $p = .011$ ), and marginally more accurate than in the PENSi category ( $p = .050$ ). Finally, participants were significantly more accurate in the Nc category than in the PENSi category ( $p < .001$ ), and there was also a significant difference between Nc and PSc ( $p < .001$ ).

Overall, these findings highlight the fact that the current task was particularly sensitive to any changes in numerosity congruence when audition was involved.

\*\*\*\*\* Insert Figure 2 about here \*\*\*\*\*

## DISCUSSION

Experiment 1 was designed to assess the hypothesis that the CC between auditory and visual stimuli would influence participants’ performance in a WM task.

Our findings are certainly in agreement with this hypothesis: CCs significantly modulated performance in the 2-back task, as evidenced by faster target-detection latencies following fully crossmodally congruent as compared to fully crossmodally incongruent trials. The participants also responded marginally significantly more accurately in the congruent trials in terms of their hit rates (PC). The comparison between the single categories (Pitch/Elevation, audiovisual Numerosity, and Pitch/Shape) and the fully incongruent category demonstrated that each type of CC has a specific effect on improving the WM participants' performance. Specifically, PSc only had an effect only when participants focused on the visual modality, whereas PEc and Nc helped participants when they had to attend to the auditory modality. It might be thought that the effect of PEc is due to a similarity between pitch and elevation (since that they share the same verbal labels, e.g., high-low). However, the literature demonstrates that pitch/elevation congruence exists and has an effect even in preverbal infants (Dolscheid, Hunnius, Casasanto & Majid, 2012; Walker, Bremner, Mason, Spring, Mattock, Slater & Johnson, 2010), ruling out the hypothesis that it is the linguistic similarity that could account for the facilitation effect that we found.

This result appears to corroborate and expand previous studies showing that CCs affect higher order cognitive processes (here, WM), such as attention (Chiou & Rich, 2012; Klapetek et al., 2012; see also Spence & Deroy, 2013). However, before discussing these findings, we further investigated the nature of the effect reported here.

In Experiment 1, the targets always constituted a perfect repetition of the bimodal stimulus that had been presented 2 positions earlier in the sequence. As a result, it is difficult to know whether the effect of CCs on performance in the 2-back

Commented [CS6]: add also their 2015 follow-up paper?

task was attributable to an enhancement of the internal WM representation of that specific audiovisual stimulus display, or just a consequence of a “redundant” effect provided by the CC when the display was presented (e.g., as in the Redundant Target Effect, where a facilitation effect is linked to the presentation of [bimodal target dimensions](#); Gondan, Niederhaus, Rösler, & Röder, 2005). This latter hypothesis is also supported by [research by](#) Miller (1991), who demonstrated that responses to congruent bimodal targets (e.g., high pitched sounds coupled with visual shapes presented above fixation) are indeed faster than responses to unimodal targets. To distinguish between these two alternative hypotheses, we designed a further experiment in which the crossmodal audiovisual components included in the display were systematically varied. Even if we found that PSc has an effect on participants’ performance when they attended to the visual modality, and the PEc has an effect when focusing on the auditory modality, we chose to focus the investigation on the effect of audiovisual numerosity, since this represents the strongest result we [found obtained](#) (cf. Fig. 2 & 3). Moreover, we introduced 3-back repetition (lures; 5% of the trials) to make sure that the effect we found in Experiment 1 was not attributable only to stimulus familiarity but based on recollection processes<sup>4</sup>.

Commented [CS7]: a

Commented [CS8]: do we need dimensions?

<sup>4</sup>Familiarity is believed to be a process that merely relies on the identity and activation of a representation in memory, while recollection is assumed to be an analytic search process that involves the context in which an item was previously encountered. Applied to the n-back task, this means that an item that was presented shortly before will elicit a familiarity signal, but this familiarity signal will not allow [the observer to](#) differentiate whether this item is in the target n-back position or not. On mismatch trials, it is possible to react accurately on the basis of the familiarity signal only. However, on lure trials, the familiarity signal will fire (“the item has been encountered previously”) but the recollection process is needed to override the misleading activation from the familiarity process by providing contextual evidence (“but it is not in the correct n-back position”; see Szmalec, Verbruggen, Vandierendonck, & Kemps, 2011).

This design opens the way to a deeper exploration of the nature of the effect:

In fact, the facilitation could be attributable to either mnemonic or response selection processes. In Experiment 2, we compared a condition in which the target stimulus constituted a perfect repetition of the (bimodal) sample stimulus presented 2 positions earlier in the sequence (e.g., CC at sample: 3 shapes and 3 tones, and CC at target: 3 shapes and 3 tones; just as in Experiment 1), with those conditions in which the CC was delivered only at the sample stimulus (e.g., CC at sample: 3 shapes and 3 tones, but not CC at target: 3 shapes and 2 tones), at the target stimulus (e.g., no CC at sample: 3 shapes and 1 tone, but CC at target: 3 shapes and 3 tones), or in neither of these (e.g., no CC at sample: 3 shapes and 1 tone, and no CC at target: 3 shapes and 1 tone) (see Figure 3). Just as in Experiment 1, the participants had to focus their attention on one sensory modality, either auditory or visual (note that here we did not ask the participants to divide attention between the two modalities given that we varied systematically the correspondence between them).

If the congruent but unattended modality affects the mnemonic representation of the attended modality, then we would expect improved performance in the n-back task whenever the CC is presented at the sample level (i.e., during information encoding), irrespective of whether CC is also presented at the target level. By contrast, if improved n-back performance is found whenever the CC is presented at the target level, irrespective of whether the CC is also presented at the encoding stage, this would be more in line with the idea that the congruent but unattended modality affects the response selection stage, possibly by reducing crossmodal interferences. Improved n-back performance only when the CC was consistently presented at both the sample and target level (i.e., only when the audiovisual information was identical



between sample and target) would instead limit the chance to disentangle the specific stage of CC impact, mnemonic vs. response selection.

We ~~tested this~~ tested this in Experiment 2 by manipulating crossmodally corresponding auditory and visual numerosity, which was the specific CC that had given rise to the effective modulation of performance in the n-back task in Experiment 1. Moreover, in Experiment 1, the mix of different kinds of crossmodal and intramodal correspondences could have confounded participants' performance, reflecting a more general effect of congruence, and not a specific CC effect. The structure of Experiment 2 and the use of only one type of CC helped us to contain and avoid the possible effect of intramodal correspondences and look at the pure CC effect on WM.

Finally, Experiment 2 was designed to assess whether the impact of CC on WM performance might be obtained even with symbolic stimuli. Experiment 1 demonstrated that facilitation effects were primarily driven by congruent audiovisual numerosity correspondences. One might speculate that the same effect that was obtained with sounds and shapes could be obtained with actual numbers (i.e., visually- or auditorily-presented digits). If this is the case, this would demonstrate that the representation is in fact interacting directly with more conceptual information (see Martino & Marks, 1999), thus encouraging an *amodal* interpretation of the correspondence. Finally, there may be a differential effect of symbolic stimuli, according to the attended modality (Baddeley, 2012; Baddeley, Lewis, & Vallar, 1984).

**Commented [CS9]:** Not sure perhaps add

Walker, P. (2012). Cross-sensory correspondences and cross talk between dimensions of connotative meaning: Visual angularity is hard, high-pitched, and bright. *Attention, Perception, & Psychophysics*, **74**, 1792-1809.

## EXPERIMENT 2

## METHODS

### PARTICIPANTS

A new group of forty-seven undergraduate students (35 females; mean age = 20.4, years; SD = 1.3; range = 18-24 years) took part in the experiment for course credit. All of the participants were right-handed, with normal or corrected-to-normal vision, normal hearing, and were naïve as to the purpose of the study.

### APPARATUS and MATERIALS

The hardware and software were the same as in Experiment 1. The custom-made script was modified to match the material and procedure changes. As before, the 2-back procedure involved the presentation of bimodal audiovisual stimuli. The stimuli were of two different kinds: Quantities or digits. The auditory quantities were sequences of one to four sinusoidal tones (2400 Hz, 5 msec ramps, total duration = 50 msec). Pauses in the stimuli that consisted of multiple tones, were 100 msec in duration: Hence the one-tone stimulus was 50 msec long, while the four-tone stimulus lasted for 500 msec. The auditory digits were 4 recordings of a male voice saying "Uno", "Due", "Tre", and "Quattro" (Italian for "One", "Two", "Three", and "Four") in an emotionally-neutral tone. All stimuli were mono recordings, at a sample frequency of 44100 Hz and they all lasted 500 msec. These sounds were coupled with visual quantities consisting of one to four black discs (diameter of each disc =  $1.66^\circ$ ), horizontally and vertically aligned at the center of the display. The visual digits were the numbers from "1" to "4", presented visually at the center of the display. The audiovisual stimuli could either be congruent or incongruent in terms of audiovisual numerosity: Namely, the stimuli could feature the same or a different number expressed in quantity or with a digit.

## DESIGN and PROCEDURE

The participants' task was the same as in Experiment 1, except for two aspects: first, we introduced the lures to check whether the effects could be due to simple familiarity (Szmalec et al., 2011). Second, this time, the participants were randomly divided into two separate groups: The first group (22 participants) had to perform the task focusing only on sounds, while the second group (25 participants) had to focus only on visual information. Blocks, number of trials, and the ratio of targets were the same as for Experiment 1.

In a 2-back task, the targets are defined by a pair of stimuli: A sample stimulus appearing at some point in the sequence, and its repetition (i.e., the target stimulus), occurred 2 steps later. In order to discriminate between a "memory" and an "encoding" or "response selection" based account of the influence of CC on WM processes (cf. Experiment 1), the targets were divided into 4 categories that differed in terms of how these stimulus pairs were arranged. In the first category (see Figure 3a, both in Quantities and in Digits), the audiovisual numerosity congruence was present in both the sample and target stimuli (thus making the stimuli bimodally identical). In the second category, the audiovisual numerosity match was limited to the sample stimulus (i.e., at the encoding phase; Figure 3b), while the target featured audiovisual numerosity incongruence. The third category (see Figure 3c) featured an audiovisual numerosity congruence restricted to the target stimulus (i.e., at the retrieval phase), while the sample stimulus (the one that had been presented 2 positions earlier in the sequence) featured an audiovisual numerosity incongruence. Finally, the fourth category featured audiovisual numerosity incongruence in terms of both the sample and the target stimuli (see Figure 3d), keeping both stimuli bimodally identical as in the fully incongruent category of Experiment 1.

Moreover, the 4 blocks were divided in 2 “quantities” blocks (see upper Figure 3, Quantities), in which all numerical information was expressed with quantities (e.g., sounds and discs), and 2 “digits” blocks (see lower Figure 3, Digits) in which the attended modality was populated by quantities, while the unattended modality was populated ~~by~~with digits (e.g., in the auditory group we had two tones along with actual number “2”, while in visual condition we had a male voice saying “Due” along with two discs). Therefore, the type of stimuli (quantities or digits) in the different blocks was only manipulated in the task-irrelevant modality, while the task-relevant modality always used quantities (sounds or discs) in all blocks. Blocks order was counterbalanced across participants. Number of trials, and the ratio of targets were the same as for Experiment 1. All the blocks included 3-back lures (5% of the trials).

\*\*\*\*\* Insert Figure 3 about here \*\*\*\*\*

## RESULTS

The data inspection for outliers revealed that five participants got PC scores significantly higher or lower than average ( $> \pm 2SD$ ) and were excluded. Therefore, the analysis was carried out on 42 participants (33 females; mean age = 20.5, years;  $SD = 1.3$ ; range = 18-24 years), 18 of ~~whom~~whom ~~had to~~had to ~~attended to~~attended to the auditory modality.

### *Reaction times*

“CC at both stimuli” and “CC at neither stimulus” ~~always~~ represent ~~always~~ a perfect match between sample and target. Since this requisite is not present in the other two conditions we ran two separated ANOVA, one comparing “CC at both stimuli” and “CC at neither stimulus”, and the other comparing “CC at sample stimulus” ~~and with~~ “CC at target stimulus”. This manipulation might ~~also~~ help us ~~also~~ to disentangle ~~between the~~ perceptual/memory effects of CC.

A 2x2x2 mixed-design ANOVA was conducted on the RT ('Hits') data with the between factor of Attended modality (auditory and visual) and the within factors of Category (CC in both sample and target stimuli and CC in neither stimulus) and Type (quantities or digits). This analysis revealed a significant main effect of Attended modality [ $F(1, 40) = 20.064$ ;  $p < .001$ ;  $\eta^2 = .334$ ], and a significant main effect of Category [ $F(1, 40) = 6.135$ ;  $p = .018$ ;  $\eta^2 = .133$ ]. ~~As in Experiment 1, the~~ main effect of Attended modality demonstrates, ~~as in Experiment 1,~~ that RTs are significantly longer when attending to the auditory modality. The main effect of Category revealed that participants were faster with CC in both ~~the~~ sample and target stimulus than ~~when there was no~~ CC in ~~neither~~ stimulus. These results help us to note that there is a facilitation effect when CC is present in both sample and target stimuli, but they say us nothing about the specific effect that numerosity CC has on WM processes. To shed light on this ~~aspect~~result, we ~~run~~conducted a second ANOVA comparing the other two conditions, namely “CC at sample stimulus” and “CC at neither stimulus”.

A 2x2x2 mixed-design ANOVA was conducted on the RT ('Hits') data with the between factor of Attended modality (auditory and visual) and the within factors of Category (CC in the sample stimulus and CC in the target stimulus) and Type (quantities or digits). This analysis revealed a significant main effect of Attended

modality [ $F(1, 40) = 21.336$ ;  $p < .001$ ;  $\eta^2 = .348$ ], and a significant main effect of Category [ $F(1, 40) = 8.920$ ;  $p = .005$ ;  $\eta^2 = .182$ ]. The ~~main effect of Category~~ latter term revealed that participants were significantly faster with a CC at the target stimulus than with a CC at the sample stimulus. The interaction between these three factors was significant [ $F(1, 40) = 12.241$ ;  $p = .001$ ;  $\eta^2 = .234$ ]. LSD post-hoc revealed a different quantities/digits congruence effect according to modality (see Figure 5). Specifically, when participants focused on the auditory modality, ~~we found~~ effects were found only with quantities congruent stimuli (CC at target stimulus < CC at sample stimulus  $p=.003$ ). On the other hand, when participants attended to the visual modality ~~we found~~ effects were found only with task-irrelevant digits congruent stimuli (CC at target stimulus < CC at sample stimulus  $p=.002$ ).

#### *Proportion correct*

A similar 2x2x2 mixed-design ANOVA was performed on the PC of “CC at both stimuli” and “CC at neither stimulus”. We found a significant main effect of Attended modality [ $F(1, 40) = 18.216$ ;  $p < .001$ ;  $\eta^2 = .313$ ], confirming once again that the auditory version of the 2-back task is generally more difficult than its visual counterpart.

Finally, we ~~run~~ conducted a 2x2x2 mixed-design ANOVA on the PC of “CC at sample stimulus” and “CC at target stimulus”. We found a significant main effect of Attended modality [ $F(1, 40) = 15.382$ ;  $p < .001$ ;  $\eta^2 = .278$ ] and a significant interaction between Attended Modality and Type [ $F(1,40) = 4.448$ ;  $p=.041$ ;  $\eta^2=.100$ ]. LSD post-hoc tests revealed that in the task-irrelevant quantities block participants were more accurate when focused on visual modality than auditory ( $p < .001$ ). No others significant terms were found.

\*\*\*\*\* Insert Table 3, Figure 4 and 5 about here \*\*\*\*\*

## DISCUSSION

The main aim of Experiment 2 was to further investigate the impact of audiovisual numerosity CC on continuous WM performance (in an n-back task). Our findings showed that audiovisual numerosity correspondences were particularly effective in terms of enhancing n-back task performance whenever the CC was presented at the target level (“CC at both stimuli” and “CC at target stimulus” categories). These results suggest that the sensory information presented in the congruent but unattended modality (e.g., the number of discs when focusing on the amount of tones) successfully enhanced ~~response~~ selection and then target detection (see [the](#) General Discussion on this point).

Moreover, ~~the current~~[our](#) results also demonstrate that correspondences modulated performance in the n-back task either when participants focused on the visual or the auditory modalities. This significantly differs than what we saw in Experiment 1, where the effect on RTs was present only when participants attended to ~~the auditory modality~~. In Experiment 1 we found that n-back task performance was particularly sensitive to numerosity congruence as compared to other kinds of more purely *modal* CCs (namely, Pitch/Elevation and Pitch/Shape). Therefore, a second aim of Experiment 2 was to investigate the nature of the numerosity CC ~~as dependent on the type of information used (digits vs quantities)~~. However, the results of Experiment 2 point towards an explanation of the effect that accounts for the modality the participants are focusing on (auditory or visual).

The results of Experiment 2 revealed that when participants focused on [the](#) auditory modality, the numerosity effect was elicited primarily by task-irrelevant

visual quantities (e.g., number of discs) and not by visual digits (e.g., numbers). Interestingly, the reverse pattern was **true** when participants focused on the visual modality. Namely, in the case of attending visually ly-attending, auditory digits have an effect on participants' performance while auditory quantities (e.g., number of tones) did not. Finally, the pattern of facilitation effects that visual quantities have on auditory modality is comparable to the facilitation effect that auditory digits have on visual modality (see Figure 5). This pattern of results seems to indicate that numerosity is processed differently, in a modality-specific way (Campbell & Epp, 2004; see General Discussion below). This modality-related differential effect might contribute in explaining why Experiment 1 yielded significant results exclusively in the auditory modality. In Experiment 1 all stimuli were quantities, and Experiment 2 showed clearly that quantities (e.g., number of tones) do not affect performance when attending to visual information. Visual performance seems to be modulated by auditory digits alone: a condition absent in Experiment 1.

Commented [CS10]: Observed?

## GENERAL DISCUSSION

Across the two experiments reported here, we investigated whether different types of CCs (i.e., pitch/shape, pitch/elevation, and audio/visual numerosity) would modulate participants' performance in a continuous 2-back WM task. The results of Experiment 1 demonstrated an overall impact of the different CCs on performance in this task, with faster target-detection latencies being reported following congruent than incongruent audiovisual correspondences, along with a marginally significant improvement in hit rates (see also Makovac et al., 2014). The presence of lures (Exp. 2), along with the fact that PENSi and crossmodally incongruent stimuli were a



perfect bimodal repetition of samples (Exp. 1 & 2), allow us to exclude the possibility that participants simply rely on a sense of familiarity to achieve a correct response (see Harbison, Atkins, & Dougherty, 2011; Oberauer, 2005). Indeed, bimodally identical stimuli in sample and target displays did not produce any kind of facilitation without CC. Experiment 1 showed that, while all the three kinds of CC included have an effect on performance, the facilitation was primarily driven by congruent audiovisual numerosity correspondences, specifically when the participants had to attend to audition.

The results reported here thus reveal important differences regarding the impact that the three different types of CC had on WM processing. Specifically, in Experiment 1, we observed a selective modulation of performance in the n-back task following audiovisual numerosity correspondences, but not following the two other kinds of CC that were investigated. While the *amodal* nature of numerosity is debated (Spence et al., 2013), our results point to the fact that numerosity congruence, as compared to other [CCs](#), has a measurably different effect on higher order processes (i.e., WM). It has been proposed that magnitude and numerical representations are closely linked (see Kadosh, Lammertyn, & Izard, 2008, for a review). This may be related to a common (multisensory) magnitude estimation system in the inferior parietal cortex (Walsh, 2003; see also Buetti & Walsh, 2009, for a review). It is thus possible that this magnitude/numerical estimation system plays a special role in the cognitive cascade, showing impact not only perceptually, but also for higher order processes, such as WM representation.

The relevance of numerosity CC was confirmed and expanded by the results of Experiment 2, where the results of Experiment 1 were replicated for both visual and auditory attending. Experiment 2 also helped us to clarify the specific nature of

Commented [CS11]: Types of CC

the impact of numerosity CC on WM performance. The facilitation effect found when numerosity congruence was delivered at the target stimulus (and when it was delivered at both stimuli, see discussion of Experiment 2) allow us to claim that the retrieval stage is the particular memory process that is affected by this CC. **Indeed, the enhancement of the participants' performance with the CC at the target stimulus demonstrates that the congruence effect is not due to some kind of encoding facilitation, e.g., some sort of perceptual reinforcement due to the perfect match within the sample stimulus, creating a single integrated memory trace (around the issue of integrated representations see Brunel, Carvalho & Goldstone, 2015; Del Gatto, Brunetti, & Delogu, 2016; Zmigrod & Hommel, 2010). In our results it is specifically the recognition stage that benefits from the congruence.** The facilitation of retrieval could be interpreted as a redundancy effect: **A** facilitation due to the fact that the target offers a crossmodal reinforcement of the information (Gondan et al., 2005; Miller, 1991). In perceptual paradigms (e.g., speeded categorization), this facilitation can be explained simply as an arousal effect (e.g., attention is exogenously captured by a congruent stimulus), and this mechanism could well be the source of the faster RTs we recorded when the target was crossmodally congruent. Alternatively, the crossmodally congruent stimulus might have contributed in making the attended stimulus at the target display more “salient” among the flow of audiovisual information continuously presented in our n-back task. Saliency have been recently shown to be an important factor in enhancing WM performance, by improving attention selection (e.g., Fine & Minnery, 2009; Melcher & Piazza, 2011; Pedale & Santangelo, 2015; Santangelo & Macaluso, 2013; Santangelo, Di Francesco, Mastroberardino, & Macaluso, 2015; see, [Santangelo, 2015](#), for a review, ~~Santangelo, 2015~~; for the long term effects, see Nardo, Brunetti, Cupellini, & Olivetti Belardinelli,

2009). Finally, congruent/unattended stimuli might provide less distraction than incongruent/unattended stimuli (e.g., see, for a review, Botvinick, Braver, Barch, Carter, & Cohen, 2001), thus freeing up resources to compare the currently attended information with that encoded at sample, thus improving target detection.

Although further research is needed in order to choose/discriminate between these alternative accounts, the key point here is that the mnemonic facilitation in the n-back performance due to CC presentation could be explained by being due mainly ~~due~~ to an attentional effect at the response selection stage, by the results of Experiment 2. This would confirm and expand previous studies that have already demonstrated an attentional effect of CCs (Chiou & Rich, 2012; Klapetek et al., 2012; see also Spence & Deroy, 2013). The crucial extension would be that the attentional effect evoked by CCs can also affect the retrieval stage of working memory performance.

The fact that numerosity relies on the concept of “number” may point towards a semantically-mediated interpretation (Martino & Marks, 1999), but other relevant findings seem to point in a different direction. The presence of numerosity matching effects in infants (Féron et al., 2006; Izard, Sann, Spelke, & Streri, 2009; Kobayashi et al., 2005), along with the evidence that links magnitude/numerosity estimation with specific brain areas (Kadosh, Lammertyn, & Izard, 2008)<sup>5</sup>, is likely to suggest its possibly *structural* nature (e.g., generated by the peculiarities of the neural systems that are involved in sensory information encoding: structural correspondences are believed to be hard-wired in the perceptual system; Marks, 1978; Stevens, 1957).

<sup>5</sup> ~~It is relevant to a~~ Notice that other CCs are not clearly linked with specific brain areas (see Spence, 2011).

**Commented [CS12]:** Though see

Sadaghiani, S., Maier, J. X., & Noppeney, U. (2009). Natural, metaphoric, and linguistic auditory direction signals have distinct influences on visual motion processing. *Journal of Neuroscience*, **29**, 6490-6499.

Bien, N., ten Oever, S., Goebel, R., & Sack, A. T. (2012). The sound of size: Crossmodal binding in pitch-size synesthesia: A combined TMS, EEG, and psychophysics study. *NeuroImage*, **59**, 663-672.

Moreover, the fact that most of these matching effects are found with small numbers (up to four, in our case), may point towards its basis in more perceptual, rather than semantic, processes. Of course, our results cannot exclude the influence of subsequent *statistical* learning effect<sup>6</sup> or *semantically mediated* influences<sup>7</sup> that may build on, and expand, these basic numerical effects.

Despite previous studies demonstrated a symmetrical effect of CCs on participants' performance in a wide range of perceptual tasks (e.g., Parise & Spence, 2012; see Spence, 2011, for a review), here we observed an asymmetrical effect of numerosity correspondences on WM performance, along the lines of Evans and Treisman (2010). Attending to auditory information was generally characterized by slower RTs as compared to when a visual focus was adopted. This effect might be interpreted as attributable to the numerosity of auditory quantities necessarily unfolding over time, while visual quantities is perceived much more rapidly (see the process of visual and auditory *subitizing*; Repp, 2007). Indeed, it ~~is mechanically~~ takes longer to estimate the numerosity with a sequential presentation as compared to an instantaneous one (e.g. in our case auditory and visual, respectively). Differences in RT are thus an unavoidable byproduct of the substantial differences between auditory and visual numerosity processing, as administered in this study. Since auditory stimuli

<sup>6</sup> Statistical correspondences (Spence, 2011) are due to learning processes linked to our brain's ability to extract regularities from the environment. These correspondences reflect natural correlations between stimulus attributes (Marks, 2000), but can as well be artificially produced in the laboratory (Conway & Christiansen, 2006; Ernst, 2007; Teramoto, Hidaka, & Sugita, 2010), and may determine functional modifications in brain activity (Zangenehpour & Zatorre, 2010).

<sup>7</sup> Semantically mediated correspondences (Spence, 2011) are those cases in which common linguistic terms are used to define different parameters in different modalities (e.g., "low" and "high" for both pitch and visual elevation; Stumpf, 1883). Martino and Marks (1999) proposed the semantic coding hypothesis to account for this kind of learned semantic correspondence.

are processed more slowly, they might become available too late to affect visual processing, thus making the visual responses less susceptible to enhancement following the concurrent presentation of the auditory component (i.e., the audiovisual numerosity correspondences).

Commented [CS13]: CC?

Our results in Experiment 2 seem to definitely point towards a modality-sensitive effect of numerosity CC. Such an asymmetry might be surprising given that we found an effect on an “amodal” dimension, which – by definition – ought to be similar across modalities. However, recent literature highlight that numerical representation is primarily non-abstract and is supported by specific neuronal populations (Kadosh & Walsh, 2009; Lyons, Ansari, & Beilock, 2015). According to the *encoding complex hypothesis*, separate modality-specific number codes exist. Therefore, number processing might be mediated by modality-specific processes (e.g., visual quantities, auditory digits) and not only by an abstract code (Campbell & Epp, 2004). **Although further research is clearly needed to confirm the current findings, our results ~~seem~~ would appear to confirm such an account:** numerosity has different effects according to where we focus our attention. This effect may be connected to the different ways words (or digits) are processed in WM and to the different ways neural signals are generated (Crottaz-Herbette, Anagnoson, & Menon, 2003; Lyons et al., 2015).

Moreover, some studies have shown that auditory presentation is superior to visual presentation for the short-term retention of verbal material (Craik, 1969; Penney, 1989). In effect, in Experiment 2 we found that spoken digits had an effect on visually congruent targets, while visual digits did not affect memory performance for auditory quantities. This finding could be straightforwardly interpreted as caused by the fact that spoken words entering in the phonological loop are processed more

directly than visual words (Baddeley et al., 1984). This fact alone would be crucial to explain the visual-auditory asymmetrical effect obtained in Experiment 2. The asymmetry could be due to crucial differences in the processing of words (digits in our case), that gain different access to the WM system (Baddeley, 2012; Baddeley et al., 1984; Dehaene & Akhavein, 1995). In this sense, the asymmetries observed in the processing of stimuli in the different modalities, could render numerosity congruence intrinsically dependent on the modality it is channelled through. Following along this explanation, its nature would be not purely abstract as some authors claim (Féron, et al. 2006), but instead numerosity would access an abstract representation differently, according to the specific modality it is channeled through.

Lastly, it is important to note that our participants might have relied on the strategy of estimating the magnitude of sounds or shapes according to their duration or extension (general visual size) and not properly their numerosity. This possibility does not, however, hold to explain the effect of digits and it would simply make the effect of magnitude estimation a related one – without changing the substance of the effect.

In summary, the results of the present study provide some evidence that CC effects also influence WM. This influence appeared to be attentionally-mediated and to be mainly related to the stage of response selection ~~stage~~ (e.g., in this memory task, the recognition stage). These findings extend current knowledge concerning the influence of CCs on post-perceptual information processing. Nevertheless, future researches are required to enhance our knowledge about the factors that affect these effects. Importantly, the impact of CCs in enhancing memory representations can shed new light on the link between the perceptual and post-perceptual, learning-mediated, stages of processing.

## REFERENCES

- Baddeley, A. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology*, 63, 1-29.
- Baddeley, A., Lewis, V., & Vallar, G. (1984). Exploring the articulatory loop. *The Quarterly Journal of Experimental Psychology*, 36, 233-252.
- Bernstein, I. H., & Edelstein, B. A. (1971). Effects of some variations in auditory input upon visual choice reaction time. *Journal of Experimental Psychology*, 87, 241-247.
- Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, 108, 624-652.
- Bremner, A., Caparos, S., Davidoff, J., de Fockert, J., Linnell, K., & Spence, C. (2013). Bouba and Kiki in Namibia? A remote culture make similar shape-sound matches, but different shape-taste matches to Westerners. *Cognition*, 126, 165-172.
- Brunel, L., Carvalho, P. F., & Goldstone, R. L. (2015). It does belong together: Cross-modal correspondences influence cross-modal integration during perceptual learning. *Frontiers in Psychology*, 6, 358.
- Bueti, D., & Walsh, V. (2009). The parietal cortex and the representation of time, space, number and other magnitudes. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 364, 1831-1840.

Campbell, J. I., & Epp, L. J. (2004). An encoding-complex approach to numerical cognition in Chinese-English bilinguals. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, 58, 229.

Commented [CS14]: Really a 1 page article?

Chiou, R., & Rich, A. N. (2012). Cross-modality correspondence between pitch and spatial location modulates attentional orienting. *Perception*, 41, 339-353.

Conway, C. M., & Christiansen, M. H. (2006). Statistical learning within and between modalities pitting abstract against stimulus-specific representations. *Psychological Science*, 17, 905-912.

Craik, F. I. (1969). Modality effects in short-term storage. *Journal of Verbal Learning and Verbal Behavior*, 8, 658-664.

Crottaz-Herbette, S., Anagnoson, R. T., & Menon, V. (2004). Modality effects in verbal working memory: Differential prefrontal and parietal responses to auditory and visual stimuli. *NeuroImage*, 21, 340-351.

Dehaene, S., & Akhavein, R. (1995). Attention, automaticity, and levels of representation in number processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21, 314-326.

Del Gatto, C., Brunetti, R., & Delogu, F. (2016). Cross-modal and intra-modal binding between identity and location in spatial working memory: The identity of objects does not help recalling their locations. *Memory*, 24, 603-615.

Dolscheid, S., Hunnius, S., Casasanto, D., & Majid, A. (2012). The sound of thickness: Prelinguistic infants' associations of space and pitch. *In the 34th Annual Meeting of the Cognitive Science Society* (pp. 306-311).



Ernst, M. O. (2007). Learning to integrate arbitrary signals from vision and touch. *Journal of Vision*, 7, 1-7.

Evans, K. K., & Treisman, A. (2010). Natural cross-modal mappings between visual and auditory features. *Journal of Vision*, 10, 1-12.

Commented [CS15]: Aren't you missing article number?

Féron, J., Gentaz, E., & Streri, A. (2006). Evidence of amodal representation of small numbers across visuo-tactile modalities in 5-month-old infants. *Cognitive Development*, 21, 81-92.

Fine, M. S., & Minnery, B. S. (2009). Visual salience affects performance in a working memory task. *Journal of Neuroscience*, 29, 8016-8021.

Gondan, M., Niederhaus, B., Rösler, F., & Röder, B. (2005). Multisensory processing in the redundant-target effect: A behavioral and event-related potential study. *Perception & Psychophysics*, 67, 713-726.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74, 1464-1480.

Harbison, J. I., Atkins, S. M., & Dougherty, M. R. (2011). N-back training task performance: Analysis and model. In L. Carlson, L. C. Hoelscher, C., & T. Shiply, T. F. (Eds.), *Proceedings of the 33<sup>rd</sup> Annual Meeting of the Cognitive Science Society* (pp. 120-125), July 2011, Boston, MA.

Hinton, L., Nichols, J., & Ohala, J. J. (Eds.). (2006). *Sound symbolism*. Cambridge, UK: Cambridge University Press.

Izard, V., Sann, C., Spelke, E. S., & Streri, A. (2009). Newborn infants perceive abstract numbers. *Proceedings of the National Academy of Sciences of the USA*, 106, 10382-10385.

Kadosh, R. C., Lammertyn, J., & Izard, V. (2008). Are numbers special? An overview of chronometric, neuroimaging, developmental and comparative studies of magnitude representation. *Progress in Neurobiology*, 84, 132-147.

Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. *Journal of Experimental Psychology*, 55, 352-358.

Klapetek, A., Ngo, M. K., & Spence, C. (2012). Do crossmodal correspondences enhance the facilitatory effect of auditory cues on visual search? *Attention, Perception, and Psychophysics*, 74, 1154-1167.

Kobayashi, T., Hiraki, K., & Hasegawa, T. (2005). Auditory-visual intermodal matching of small numerosities in 6-month-old infants. *Developmental Science*, 8, 409-419.

Köhler, W. (1929). *Gestalt psychology*. New York, NY: Liveright.

Köhler, W. (1947). *Gestalt psychology: An introduction to new concepts in modern psychology*. New York, NY: Liveright.

Lewkowicz, D. J., & Turkewitz, G. (1980). Cross-modal equivalence in early infancy: Auditory-visual intensity matching. *Developmental Psychology*, 16, 597-607.

Ludwig, V. U., Adachi, I., & Matzuzawa, T. (2011). Visuoauditory mappings between high luminance and high pitch are shared by chimpanzees (*Pan troglodytes*)

and humans. *Proceedings of the National Academy of Sciences of the USA*, 108, 20661-20665.

Lyons, I. M., Ansari, D., & Beilock, S. L. (2015). Qualitatively different coding of symbolic and nonsymbolic numbers in the human brain. *Human Brain Mapping*, 36, 475-488.

Makovac, E., Kwok, S. C., & Gerbino, W. (2014). Attentional cueing by cross-modal congruency produces both facilitation and inhibition on short-term visual recognition. *Acta Psychologica*, 152, 75-83.

Marks, L. E. (1978). *The unity of the senses: Interrelations among the modalities*. New York, NY: Academic Press.

Marks, L. E. (1987). On cross-modal similarity: Auditory-visual interactions in speeded discrimination. *Journal of Experimental Psychology: Human Perception and Performance*, 13, 384-394.

Marks, L. E. (2004). Cross-modal interactions in speeded classification. In G. A. Calvert, C. Spence, & B. E. Stein (Eds.), *Handbook of multisensory processes* (pp. 85-105). Cambridge, MA: MIT Press.

Marks, L. E., Szczesiul, R., & Ohlott, P. (1986). On the cross-modal perception of intensity. *Journal of Experimental Psychology: Human Perception and Performance*, 12, 517-534.

Martino, G., & Marks, L. E. (1999). Perceptual and linguistic interactions in speeded classification: Tests of the semantic coding hypothesis. *Perception*, 28, 903-923.

Mastroberardino, S., Santangelo, V., Botta, F., Marucci, F. S., & Olivetti Belardinelli, M. (2008). How the bimodal format of presentation affects working memory: An overview. *Cognitive Processing*, 9, 69-76.

Melara, R. D. & O'Brien, T. P. (1987). Interactions between synesthetically corresponding dimensions. *Journal of Experimental Psychology: General*, 116, 323-336.

Melcher, D., & Piazza, M. (2011). The role of attentional priority and saliency in determining capacity limits in enumeration and visual working memory. *PLoS One*, 6, e29296.

Miller, J. O. (1991). Channel interaction and the redundant targets effect in bimodal divided attention. *Journal of Experimental Psychology: Human Perception and Performance*, 17, 160-169.

Mozolic, J. L., Hugenschmidt, C. E., Peiffer, A. M., & Laurienti, P. J. (2008). Modality-specific selective attention attenuates multisensory integration. *Experimental Brain Research*, 184, 39-52.

Nardo, D., Brunetti, R., Cupellini, E., & Belardinelli, M. O. (2009). The influence of melodic and rhythmic redundancies on recognition memory for unknown musical themes. *Musicae Scientiae*, 13, 337-355.

Oberauer, K. (2005). Binding and inhibition in working memory: Individual and age differences in short-term recognition. *Journal of Experimental Psychology: General*, 134, 368-387.

Parise, C., & Spence, C. (2009). 'When birds of a feather flock together': Synesthetic correspondences modulate audiovisual integration in non-synesthetes. *PLoS ONE*, 4, e5664.

Parise, C., & Spence, C. (2012). Audiovisual crossmodal correspondences and sound symbolism: An IAT study. *Experimental Brain Research*, 220, 319-333.

Pedale, T., & Santangelo, V. (2015). Perceptual salience affects the contents of working memory during free-recollection of objects from natural scenes. *Frontiers in Human Neurosciences*, 9, 1-8.

Commented [CS16]: You are missing article number no

Penney, C. G. (1989). Modality effects and the structure of short-term verbal memory. *Memory & Cognition*, 17, 398-422.

Ramachandran, V. S., & Hubbard, E. M. (2001). Synaesthesia – A window into perception, thought and language. *Journal of Consciousness Studies*, 8, 3-34.

Ramachandran, V. S., & Hubbard, E. M. (2003). Hearing colors, tasting shapes. *Scientific American*, 288(5), 52-59.

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

Santangelo, V. (2015). Forced to remember: When memory is biased by salient information. *Behavioural Brain Research*, 283, 1-10.

Santangelo, V., Di Francesco, S. A., Mastroberardino, S., & Macaluso, E. (2015). Parietal cortex integrates contextual and saliency signals during the encoding of natural scenes in working memory. *Human Brain Mapping*, 36, 5003-5017.

Santangelo, V., & Macaluso, E. (2013). Visual salience improves spatial working memory via enhanced parieto-temporal functional connectivity. *Journal of Neuroscience*, 33, 4110-4117.

Santangelo, V., Mastroberardino, S., Botta, F., Marucci, F. S., & Olivetti Belardinelli, M. (2006). On the influence of audio-visual interactions on working memory performance: A study with nonsemantic stimuli. *Cognitive Processing*, 7, 187.

Commented [CS17]: 1 page only?

Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics*, 73, 971-995.

Spence, C., & Deroy, O. (2013). How automatic are crossmodal correspondences? *Consciousness and Cognition*, 22, 245-260.

Spence, C., Deroy, O., & Bremner, A. (2013). Questioning the utility of the concept of amodality: Towards a revised framework for understanding crossmodal relations. *Multisensory Research*, 26, 57.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review*, 64, 153-181.

Stumpf, K. (1883). *Tonpsychologie I [Psychology of the tone]*. Leipzig: Hirzel.

Szmalec, A., Verbruggen, F., Vandierendonck, A., & Kemps, E. (2011). Control of interference during working memory updating. *Journal of Experimental Psychology: Human Perception and Performance*, 37, 137-151.

Teramoto, W., Hidaka, S., & Sugita, Y. (2010). Sounds move a static visual object. *PLoS ONE*, 5, e12255.

Terhune, D. B., Wudarczyk, O. A., Kochuparampil, P., & Kadosh, R. C. (2013). Enhanced dimension-specific visual working memory in grapheme-color synesthesia. *Cognition*, 129, 123-137.

Walsh, V. (2003). A theory of magnitude: Common cortical matrices of time, space and quality. *Trends in Cognitive Sciences*, 7, 483-488.

Walker, P., Bremner, J. G., Mason, U., Spring, J., Mattock, K., Slater, A., & Johnson, S. P. (2009). Preverbal infants' sensitivity to synaesthetic cross-modality correspondences. *Psychological Science*, 21, 21-25.

Zangenehpour, S., & Zatorre, R. J. (2010). Crossmodal recruitment of primary visual cortex following brief exposure to bimodal audiovisual stimuli. *Neuropsychologia*, 48, 591-600.

Zmigrod, S., & Hommel, B. (2010). Temporal dynamics of unimodal and multimodal feature binding. *Attention, Perception, & Psychophysics*, 72, 142-152.

Formatted: Font: Italic

**Table 1.** Categories of crossmodal correspondences used in Experiment 1. Examples are represented in Figure 1.

Congruent dimensions	Incongruent dimensions	Example
Pitch/Elevation, AV Numerosity, and Pitch/Shape - <b>PENSc</b>	-	Two “kiki shapes” placed at the top coupled with two high-pitched sounds
-	Pitch/Elevation, AV Numerosity, and Pitch/Shape - <b>PENSi</b>	One “bouba shape” placed at the bottom coupled with two high-pitched sounds
Pitch/Elevation – <b>Pec</b>	AV Numerosity and Pitch/Shape	Two “bouba shapes” located at the top accompanied by one high-pitched sounds
AV Numerosity* - <b>Nc</b>	Pitch/Elevation and Pitch/Shape	Three “bouba shapes” placed at the bottom coupled with three high-pitched sounds
Pitch/Shape - <b>PSc</b>	Pitch/Elevation and AV Numerosity	Two “bouba shapes” placed at the top coupled with a single low-pitched sound
Pitch/Elevation and AV Numerosity - <b>PENc</b>	Pitch/Shape	One “kiki shape” placed at the bottom of the screen, with one low-pitched sound
Pitch/Elevation and Pitch/Shape - <b>PESc</b>	AV Numerosity	Two “kiki shapes” placed at the top of the screen, with one high-pitched sound
AV Numerosity and Pitch/Shape - <b>PNSc</b>	Pitch/Elevation	Three “bouba shapes” at the top of the screen, coupled with three low-pitched sounds



**Table 2.** False alarms (FA), Accuracy (proportion correct; PC) and RTs (msec) in Experiment 1. Standard deviations in brackets (see Figure 1 and Table 1 for a detailed description of the categories used in Experiment 1).

Category	Auditory			Visual			Audiovisual		
	FA	PC	RT (Hit)	FA	PC	RT (Hit)	FA	PC	RT(Hit)
PENSi	0.08 (.06)	0.73 (.13)	885 (131)	0.05 (.03)	0.80 (.15)	660 (158)	0.04 (.04)	0.78 (.11)	763 (134)
PENSc	0.09 (.06)	0.75 (.13)	835 (131)	0.06 (.04)	0.83 (.11)	631 (149)	0.06 (.04)	0.82 (.14)	700 (110)
PEc	0.07 (.05)	0.77 (.09)	853 (141)	0.06 (.05)	0.81 (.12)	633 (127)	0.04 (.04)	0.76 (.13)	745 (133)
Nc	0.09 (.06)	0.80 (.10)	835 (120)	0.08 (.02)	0.79 (.10)	655 (151)	0.06 (.04)	0.78 (.11)	717 (99)
PSc	0.10 (.06)	0.72 (.13)	889 (140)	0.06 (.04)	0.80 (.14)	628 (125)	0.08 (.05)	0.76 (.13)	727 (127)

Figure 1

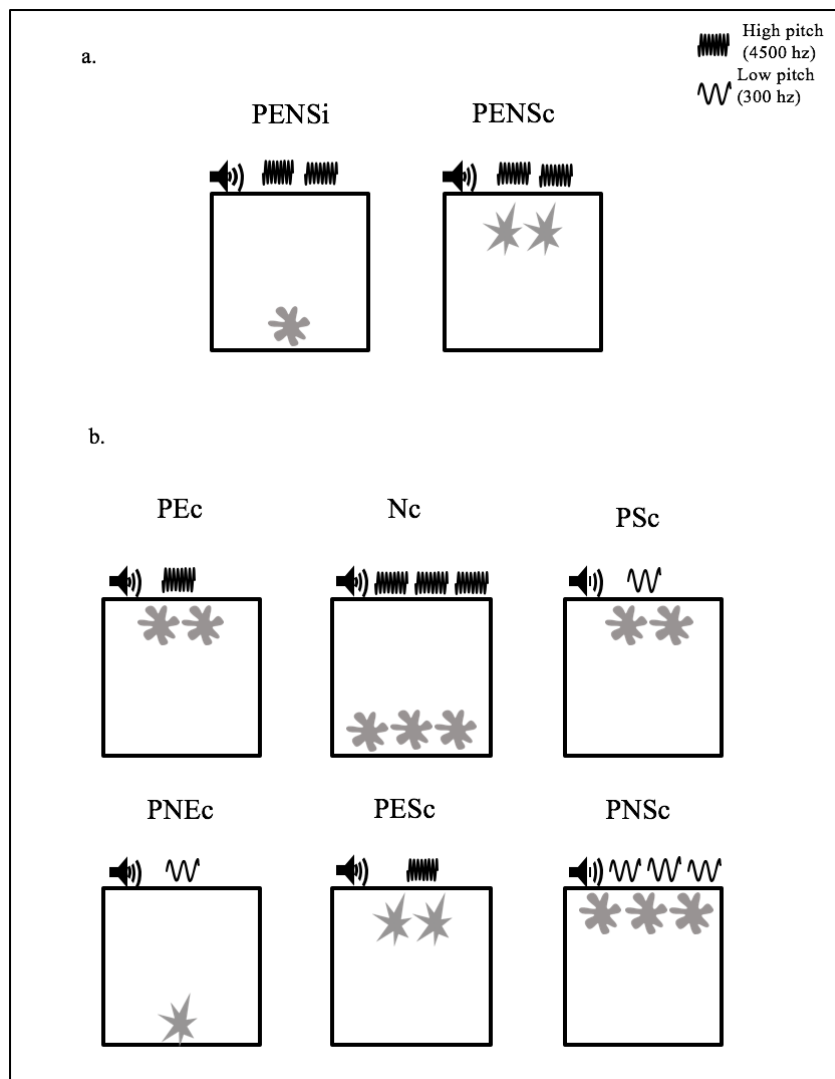


Figure 2

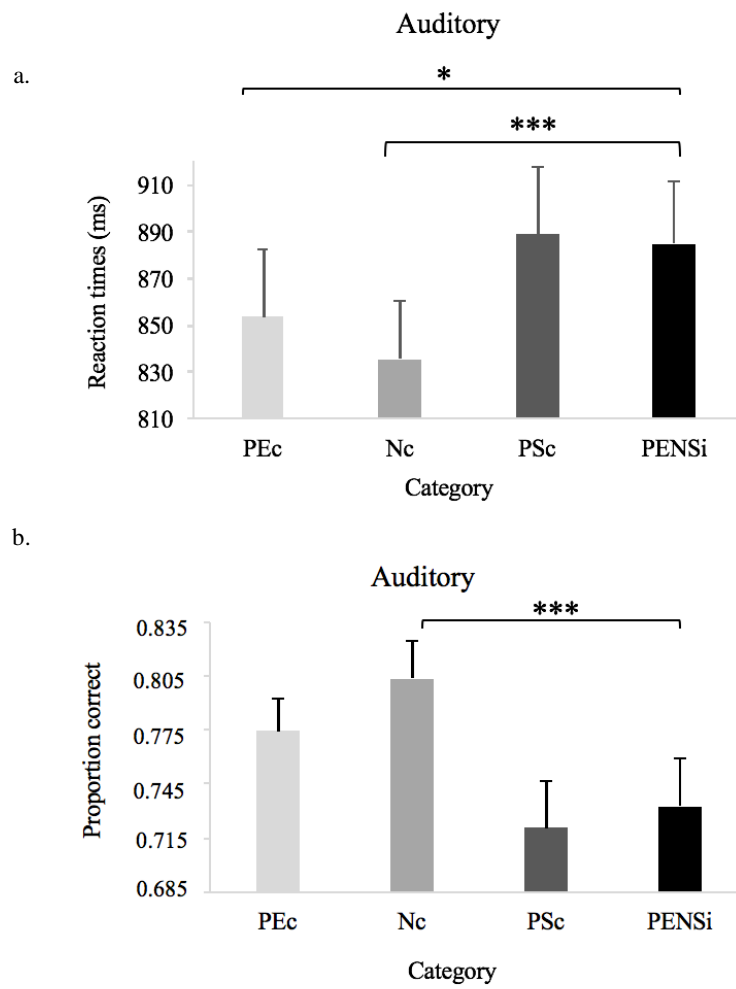
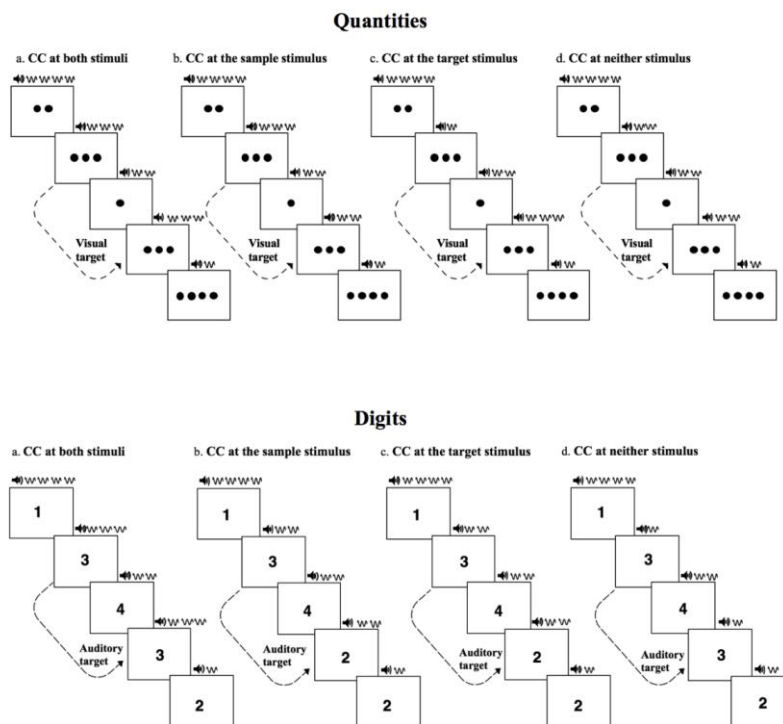


Figure 3

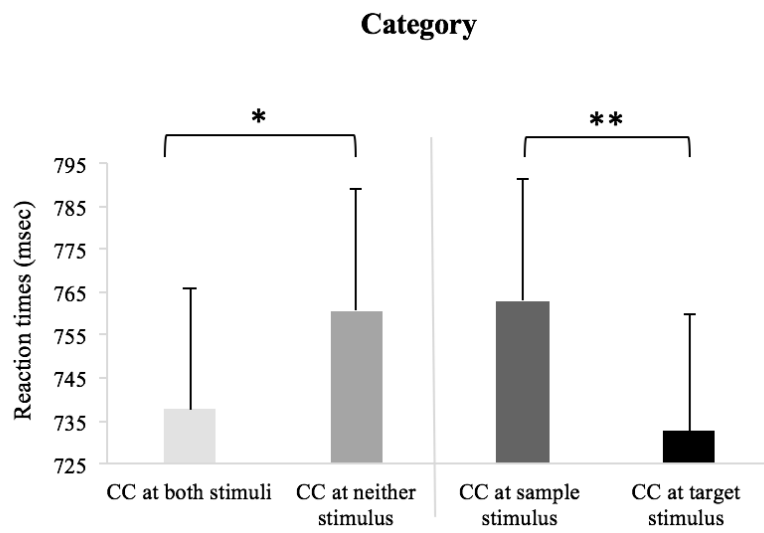


**Table 3.** Auditory and visual False alarms (FA), Accuracy (proportion correct; PC) and RTs (msec) for quantities and digits in Experiment 2. Standard deviations in brackets. CC stands for “crossmodal correspondences”.

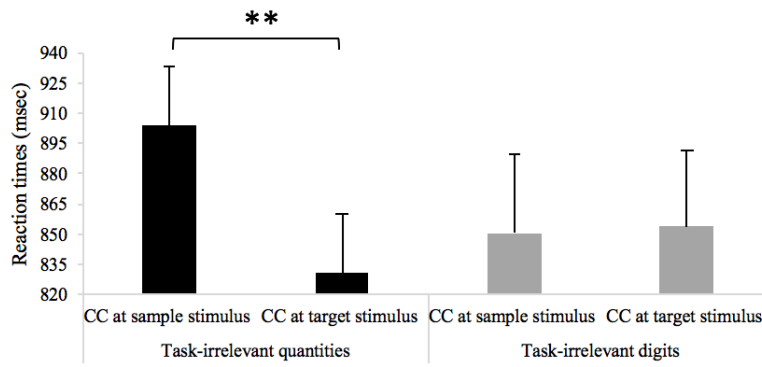
	Auditory					
	Quantities			Digits		
Category	FA	PC	RT (Hit)	FA	PC	RT (Hit)
CC at both stimuli	0.21 (.10)	0.63 (.12)	843 (156)	0.21 (.09)	0.58 (.11)	854 (182)
CC at neither stimulus		0.59 (.11)	870 (151)		0.60 (.12)	885 (175)
CC at sample stimulus		0.60 (.18)	904 (125)		0.65 (.12)	851 (163)
CC at target stimulus		0.56 (.12)	831 (124)		0.60 (.11)	854 (161)
		Visual				
	Quantities			Digits		
Category	FA	PC	RT (Hit)	FA	PC	RT (Hit)
CC at both stimuli	0.11 (.09)	0.75 (.15)	665 (141)	0.09 (.05)	0.74 (.14)	643 (154)
CC at neither stimulus		0.75 (.17)	687 (169)		0.73 (.20)	658 (126)
CC at sample stimulus		0.75 (.12)	671 (162)		0.69 (.17)	683 (173)
CC at target stimulus		0.74 (.16)	672 (137)		0.69 (.14)	627 (163)

Formatted: Underline

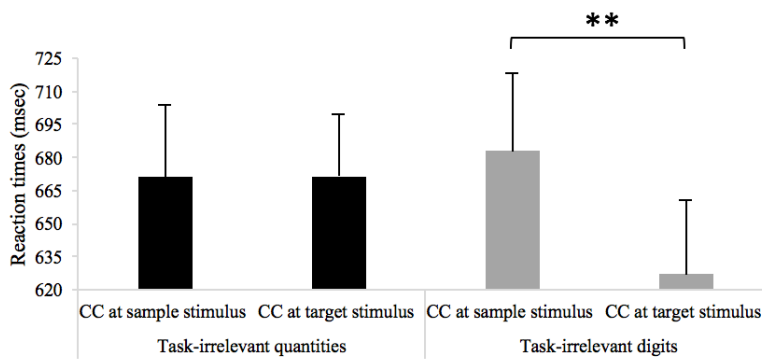
Commented [CS18]: Thought convention was to put all tables together and then all figures together or vice versa, not to mix them up?

**Figure 4**

**Figure 5**  
**Auditory**



**Visual**



## FIGURE CAPTIONS

**Commented [CS19]:** And do figure legends go before figures?

**Figure 1:** Types of crossmodal correspondences: (a.) matching/un-matching between all audiovisual dimensions; (b.) matching between 2 audiovisual dimensions and matching between 3 audiovisual dimensions. See Table 1 for a full description of the conditions.

**Figure 2:** Auditory RTs (a.) and accuracy (b.) in the interaction between singles categories (Pitch/Elevation congruence - PEc, audio/visual Numerosity congruence - Nc, and Pitch/Shape congruence - PSc) and fully incongruent category (Pitch Elevation Numerosity Shape incongruence - PENSi). Error bars represent standard errors of the means. Brackets highlight main results (\*  $p < .05$ ; \*\*\*  $p < .001$ ), please see text for full comparisons.

**Commented [CS20]:** Figure legends should spell out which expt the data are referring to ....

Add in Experiment 1 somewhere

**Figure 3:** Upper - Examples of the stimulus categories used in the “quantities” blocks of Experiment 2 for those participants instructed to focus on visual information. Please see the text for a detailed description of the four conditions. In Experiment 2, half of the participants attended to the auditory information while ignoring the visual information: For these participants, the sequences were built with the same categories, but with auditory targets. Lower - Examples of the stimulus categories used in the “digits” blocks of Experiment 2 for those participants instructed to focus on auditory information. Please see the text for a detailed description of the four conditions. In Experiment 2, half of the participants attended to the visual information while



ignoring the auditory information: For these participants, the sequences were built with the same categories, but with visual targets.

**Figure 4:** Category RTs in Experiment 2. Error bars represent standard errors of the means (\*  $p < .05$ , \*\*  $p < .01$ ). “CC” stands for crossmodal correspondences.

**Figure 5:** RTs in the interaction between “Attended Modality”, “Category” and “Type” in Experiment 2. Error bars represent standard errors of the means (\*\*  $p < .01$ ). “CC” stands for crossmodal correspondences.