Spatial Shifts of Audio-Visual Interactions by Perceptual Learning are Specific to the Trained Orientation and Eye [*]

Melissa A. Batson [1,2,**], **Anton L. Beer** [3], **Aaron R. Seitz** [4]
and Takeo Watanabe [1,2]

[1] Program in Neuroscience, Boston University, 24 Cummington Street, Boston, MA 02215, USA
[2] Department of Psychology, Boston University, 64 Cummington Street, Boston, MA 02215, USA
[3] Universität Regensburg, Institut für Psychologie, Universitätsstr. 31, 93053 Regensburg, Germany
[4] Department of Psychology, University of California – Riverside, 900 University Avenue, Riverside, CA 92521, USA

Received 13 January 2011; accepted 12 August 2011

Abstract

A large proportion of the human cortex is devoted to visual processing. Contrary to the traditional belief that multimodal integration takes place in multimodal processing areas separate from visual cortex, several studies have found that sounds may directly alter processing in visual brain areas. Furthermore, recent findings show that perceptual learning can change the perceptual mechanisms that relate auditory and visual senses. However, there is still a debate about the systems involved in cross-modal learning. Here, we investigated the specificity of audio-visual perceptual learning. Audio-visual cuing effects were tested on a Gabor orientation task and an object discrimination task in the presence of lateralised sound cues before and after eight-days of cross-modal task-irrelevant perceptual learning. During training, the sound cues were paired with visual stimuli that were misaligned at a proximal (trained) visual field location relative to the sound. Training was performed with one eye patched and with only one Gabor orientation. Consistent with previous findings we found that cross-modal perceptual training shifted the audio-visual cueing effect towards the trained retinotopic location. However, this shift in audio-visual tuning was only observed for the trained stimulus (Gabors), at the trained orientation, and in the trained eye. This specificity suggests that multimodal interactions resulting from cross-modal (audio-visual) task-irrelevant perceptual learning involves so-called unisensory visual processing areas in humans. Our findings provide further support for recent anatomical and physiological findings that suggest relatively early interactions in cross-modal processing.

[*] This article is part of the Multisensorial Perception collection, guest edited by S. Wuerger, D. Alais and M. Gondan.

[**] To whom correspondence should be addressed. E-mail: melbatson@gmail.com

DOI:10.1163/187847611X603738

Keywords

Audio-visual effects, cross-modal cuing effects, task irrelevant perceptual learning, audio-visual plasticity, multisensory processing, bottom-up integration

1. Introduction

In order to effectively perceive the external environment, inputs from multiple sensory systems (modalities) need to be combined. Many studies have shown that sounds affect visual perception (Beer and Watanabe, 2009; Eimer *et al.*, 2002). For instance, sounds presented briefly before a visual stimulus facilitate visual perception at visual field locations overlapping with the sound source (McDonald *et al.*, 2000; Spence and Driver, 1997). This cross-modal facilitation, however, reverses to cross-modal inhibition when the time between sound cue and the visual target exceeds about 300 ms (Spence and Driver, 1998) — a phenomenon also described as inhibition of return (IOR) (Klein, 2000).

Traditionally, it has been thought that these multisensory interactions involve multimodal areas of the parietal (e.g., intraparietal sulcus), temporal (e.g., superior temporal sulcus), and frontal cortex (e.g., Beauchamp, 2005; Calvert and Thesen, 2004). However, the restriction of multimodal effects to these multimodal processing areas has recently come under debate: brain regions previously believed to process strictly unimodal inputs such as regions of the thalamus and primary visual cortex (V1) respond differently to multimodal *versus* unimodal stimuli (Cappe and Barone, 2005; Wang *et al.*, 2008). Furthermore, direct structural connections between primary auditory cortex and V1 have been revealed in both humans (Beer *et al.*, 2011b) and non-human primates (Falchier *et al.*, 2002; Rockland and Ojima, 2003). These connections may form the basis of low-level interactions between so-called unisensory cortices (Eckert *et al.*, 2008; also see Driver and Noesselt, 2008; Foxe and Schroeder, 2005, for review).

Our present study builds upon this evidence of multisensory interactions at early stages of sensory processing to ask how cross-modal cueing effects may be altered through training and to shed light on what neural systems may be involved in this type of learning. We do this with an exploration of how cross-modal cuing effects can be altered through learning. Cross-modal cuing effects have a spatio-temporal dependence on cue-target validity and stimulus onset asynchrony (SOA) (Beer *et al.*, 2011a; Eimer *et al.*, 2002; Klein, 2000; McDonald *et al.*, 2000; Spence and Driver, 1997, 1998). Valid cues occur on the same side of space as the corresponding target, while invalid cues occur on the opposite side; and short SOAs result in enhanced visual performance on the valid over the invalid side of space, while long SOAs have the opposite effect with greater performance on the invalid compared to the valid side (IOR).

Recent research has shown that cross-modal interactions between auditory and visual perception can be modified through cross-modal perceptual learning (Alais and Cass, 2010; Beer *et al.*, 2011a; Kim *et al.*, 2008). Training with misaligned auditory-visual stimuli results in a shift of audio-visual cross-modal tuning curves. Beer *et al.* (2011a) showed that prior to training, short-term auditory cues facilitated visual perception only at aligned visual field locations; however, after training, the same sounds facilitated visual perception at neighbouring (proximal) retinal locations.

In this previous study (Beer *et al.*, 2011a), we adapted the task-irrelevant perceptual learning (TIPL) paradigm (Seitz and Watanabe, 2003, 2005, 2009; Watanabe *et al.*, 2001, 2002) for use with cross-modal (audio-visual) stimuli. With TIPL, observers learn stimulus configurations simply by being exposed to them, even without perceiving them (Watanabe *et al.*, 2001). The basic phenomenon of TIPL is that the stimulus features of a subject's task will be learned when they are consistently presented at times of reward or behavioural success (Seitz and Watanabe, 2009). For example, discrimination of motion stimuli improves after being paired, at subthreshold level, with the (relevant) targets of the rapid serial visual discrimination task (Seitz and Watanabe, 2003; Seitz *et al.*, 2005). TIPL has been shown to result in alterations of low-level perceptual processes (Seitz *et al.*, 2009; Watanabe *et al.*, 2002) and to result in plasticity of early visual cortex (Franko *et al.*, 2010). TIPL has been found for motion processing (Watanabe *et al.*, 2002), orientation processing (Nishina *et al.,* 2007), critical flicker fusion thresholds (Seitz *et al.*, 2005, 2006b), contour integration (Rosenthal and Humphreys, 2010), and auditory formant processing (Seitz *et al.*, 2010) and thus appears to be an effective and general mechanism of learning in the brain that spans levels of processing and sensory modalities. We previously found that audio-visual interactions were highly location specific (Beer *et al.*, 2011a). Location-specific visual learning has been suggested to involve brain regions such as V1 (Karni and Sagi, 1991; Schoups *et al.*, 1995; but see Mollon and Danilova, 1996). However, it remains unclear whether cross-modal learning effects are also specific to other attributes of low-level processing stages such as the trained orientation or the trained eye.

In the present study we systematically evaluated the effects of cross-modal (audio-visual) TIPL using simple (orientation) and complex (object) stimuli (Fig. 1(a)). Subjects were tested on two separate visual discrimination tasks with lateralised sound cues before and after eight-days of cross-modal TIPL training. During training, observers were exposed to task-irrelevant sounds that were paired with spatially misaligned Gabor patches at multiple visual locations. One sound+Gabor pair was paired with a task-relevant target stimulus. Training involved only one Gabor orientation and only one eye. In test sessions, subjects were tested on how sounds affected orientation and face/house discrimination separately for each eye. We were primarily interested in how TIPL shifted cross-modal cuing effects. In particular, we wanted to know whether the TIPL-related realignment of cross-modal facilitation was specific to trained stimulus attributes including orientation and eye.

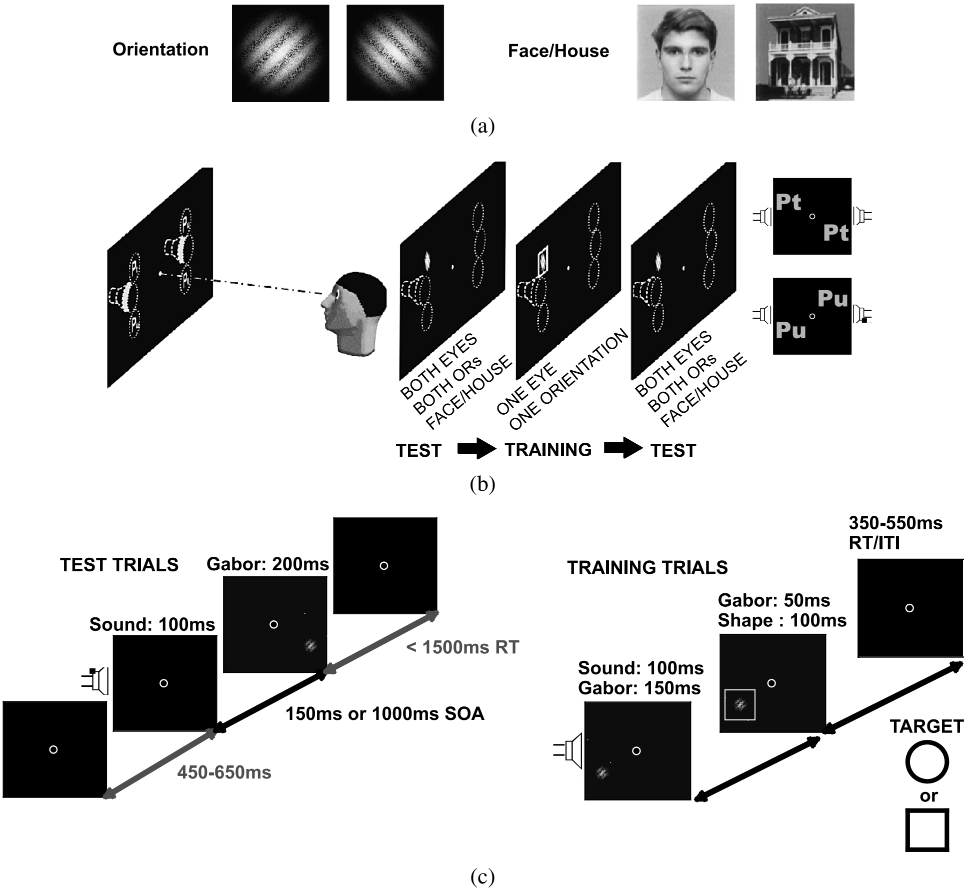


Figure 1. Experimental arrangement. (a) Stimuli: oriented Gabor patches or objects (faces/houses, adapted from Tong *et al.*, 1998), auditory stimuli were white noise bursts. See text for full stimulus details. All stimuli were presented at 16 d.o.v. from fixation. (b) Experiment apparatus and arrangement: subjects were seated facing the computer monitor. Speaker icons indicate the location of auditory cues. Visual stimuli appeared at proximal aperture locations (proximal trained — P_t or proximal untrained — P_u). One test session was conducted before and one after eight training sessions. In test sessions, subjects were tested on both orientation and object discrimination separately for each eye. During training sessions only one eye was exposed and only one of the test orientations was presented. (c) Stimulus timing: test — after a variable pre-trial period an auditory cue was presented (left or right) for 100 ms. The visual stimulus (Gabor or object) appeared at a proximal location (P_t or P_u); see (b) for 200 ms on either the valid or invalid side with a stimulus onset asynchrony (SOA) of 150 or 1000 ms. Training — subjects performed a shape detection task for eight training sessions. Each trial started with the presentation of a sound+Gabor pair (sound for 100 ms, Gabor for 200 ms). Gabors appeared at a proximal aperture location (P_u/P_t) on the left or right side. After 150 ms a circle or square (one being the target shape) encompassed the Gabor for 100 ms.

2. Materials and Methods

2.1. Subjects

Twenty-six paid volunteers with normal hearing and vision gave written informed consent. Seven participants quit prior to finishing all sessions. Five datasets were excluded because of technical problems during data acquisition. Three datasets were excluded due to ceiling performance (>96% correct) in either test session. Of the remaining eleven participants, age ranged from 18–24 years old, two were male, and all were right-handed. The Institutional Review Board of Boston University approved the study. Compensation for the experiment was $8 per completed session. No additional incentive was given for completing all sessions.

2.2. Apparatus and Stimuli

Participants were asked to fixate a bull's eye at the centre of a CRT monitor (40 × 30 cm, 1280 × 1024 pixels, 75 Hz) in a dark room. A chin rest supported the head at a viewing distance of 60 cm. Two small speakers were mounted to the left and right sides of the monitor, vertically aligned with fixation. Stimuli were presented with Psychophysics Toolbox (Brainard, 1997) version 3.0.8 and MatLab 7.1 (MathWorks, Natick, MA) with a 15″ Macintosh Power Book G4 computer (OS 10.3.9).

Visual stimuli were either oriented Gabor patches or visual objects. The Gabor patches (Fig. 1(a), left) were obliquely oriented (45° or 135°) sinusoidal gratings (maximum luminance 11 cd/m^2, spatial frequency 1.0 cycle/degree) faded to the black background (0.01 cd/m^2) by a two-dimensional Gaussian (standard deviation of 1.5° of visual angle (d.o.v)) and degraded by noise (60% of pixels randomly replaced by noise). Object stimuli (Fig. 1(a), right) consisted of faces and houses adapted from Tong *et al.* (1998) and degraded by noise (60% of pixels randomly replaced by noise). The luminance profile of the object stimuli was balanced with the Gabor patches. All visual stimuli covered approximately 6 d.o.v. and lasted 200 ms. Visual stimuli were presented at 16 d.o.v. from fixation either on the left or right and, on each side, at one of two vertical locations 6 d.o.v. from the visual field location that overlapped with the perceived sound location (Fig. 1(b)). We denoted these locations as proximal (P) because they were misaligned with, but close to the aligned locations. Proximal locations were chosen because reliable cross-modal learning effects were observed at these locations in a previous study (Beer *et al.*, 2011a). One of the proximal locations on either side (above the sound on one side and below the sound on the other side, to control for vertical bias) was the location at which new audio-visual associations were trained. This location was denoted the trained proximal location (P_t), and the other proximal location was untrained (P_u). Trained and untrained locations were counterbalanced across subjects. Trained and untrained locations were pooled across sides for analysis.

Auditory stimuli were white noise sounds presented *via* the two aforementioned speakers (KLH Audio System). Sound pressure level was about 80 dB as measured at ear position. The speaker centres were vertically aligned with fixation. Due to the monitor chassis, the speakers were horizontally displaced from the mid-vertical visual field location on the screen. Since close spatial overlap between auditory and visual stimuli is crucial for some cross-modal mechanisms (e.g., Meredith and Stein, 1986; Meyer *et al.*, 2005), sounds were horizontally aligned with the mid-vertical location by adjusting the inter-aural level differences according to the law of sines (Grantham, 1986).

2.3. Test Sessions

The experiment consisted of two test sessions and eight training sessions. All sessions were conducted on separate days. One test session was conducted before training (pre-training) and the other was conducted after training (post-training). In each test session subjects were asked to fixate a bull's eye in the centre of the screen. Following a binocular practice block, subjects were asked to cover one eye with an eye patch before the start of each block. Each trial started with the presentation of a sound for 100 ms (2 ms rise and fall time) aligned between the two proximal locations (P) on either the left or right side. After a stimulus onset asynchrony SOA of either 150 or 1000 ms, a visual stimulus appeared at one of the proximal locations (P_t or P_u) for 200 ms on either the same (*valid*) or opposite (*invalid*) side as the sound. Orientation and object discriminations were conducted in separate blocks. On trials with a long SOA (1000 ms) an additional central auditory reorienting event (Spence and Driver, 1998) consisting of a white noise sound (50 ms, 2 ms rise and fall time, equal amplitude level from each speaker) was presented 300 ms after the onset of the sound cue (see Fig. 1(c), left, for an illustration of the trial sequence). Subjects had to report the orientation (45° or 135°) of the Gabors or the object type (face or house) by pressing one of two keys within a 1500 ms response time window. A discrimination task on non-spatial features was used in order to avoid the possibility of response bias that may occur with a localisation task (Shinn-Cunningham, 2000; Spence and Driver, 1997; Zwiers *et al.*, 2003) inherently associated with a left *versus* right manual response to laterally cued stimuli. The next trial started after a variable inter-trial interval of 450–650 ms. Each test session consisted of four orientation discrimination blocks, two per eye, and two object discrimination blocks, one per eye.

2.4. Training Sessions

To investigate the specificity of cross-modal plasticity, subjects underwent eight sessions of audio-visual task-irrelevant perceptual learning (TIPL). The goal of these training sessions was to establish a new link between the sound source and one of the proximal (non-overlapping) visual field locations (P_t) — similar to a previous study (Beer *et al.*, 2011a) where a TIPL paradigm (Seitz and Watanabe, 2003) was adapted for use with audio-visual stimuli. In contrast to this previous study, subjects were trained on only one eye and one orientation in all sessions. Eye (left or right) and orientation were counterbalanced across subjects. During TIPL, stimulus configurations that occur together with a task-relevant target are learned. In each training session subjects performed a shape detection task, which provided this relevant target. Each trial started with a sound presented for 100 ms either on the left or right side (as in trials of the test session). At the onset of each sound a Gabor was presented for 200 ms at a proximal aperture location (P_u/P_t; Fig. 1(b), right) on the same (*valid*) or opposite (*invalid*) side as the sound. After a delay of 150 ms relative to the onset of the sound+Gabor pair a simple shape (circle or square) encompassed the Gabor for 50 ms and remained visible for another 50 ms after offset of the Gabor. Subjects had to detect one of these encompassing shapes, either the circle or the square (alternating across sessions), by pressing a button whenever the target shape appeared. Sounds and Gabors were irrelevant to this shape-detection task.

All sound–Gabor pairs were equally likely. However, target shapes were more likely to be paired with valid sound+Gabor pairs at the trained location (P_t). Target shapes therefore established a task-relative association with the task-irrelevant valid sound+Gabor stimulus pairs at the trained location (P_t: targets at trained locations were always cued validly). Target shapes at untrained locations (P_u) were equally often paired with valid or invalid sound+Gabor stimulus pairs. Note that subjects were exposed to only one Gabor orientation over all sessions (Fig. 1(b)). This trained orientation was counterbalanced across subjects. Further note that no object stimuli (faces or houses) were presented during training. Each training session consisted of six blocks of 448 trials each. Performance feedback (hits and false alarms) was provided after each block and subjects were informed that good performance is related to high hit and low false alarm percentages (see Fig. 1(c) for an illustration of test and training trials).

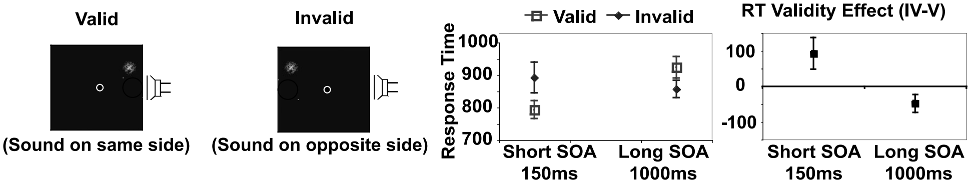
2.5. Data Analysis

Results were analyzed with regard to the validity effect (*VE*), that is, the difference in performance measures from validly (same side) *versus* invalidly (opposite side) cued trials (valid minus invalid for performance measures, invalid minus valid for response time). As the VE measures the *difference* between valid and invalid cuing effects for the same visual stimulus, performance differences across aperture locations are accounted for. Moreover, we were primarily interested in the change of the VE from the pre- to the post-training test. This measure also corrects for performance differences across tests and subjects (see Fig. 2(a) for more information on calculating the validity effect).

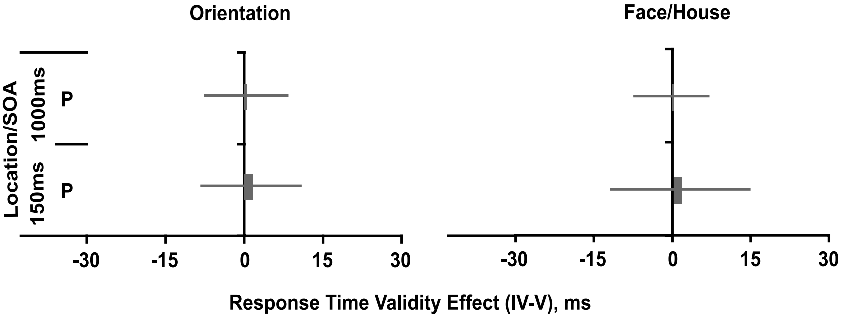
3. Results

For the first (pre-training) test, we examined the validity effect for both SOAs (150 and 100 ms) and both tasks (orientation discrimination and object discrimination). Note that no eye, orientation or location has been trained prior to this test. Therefore, we pooled across eyes, orientations and locations. No pooled VEs were significantly different from zero (Fig. 2(b)). This was expected since previous research has shown that cross-modal facilitation is observed at visual field locations overlapping with the sound cue and absent at neighbouring visual field locations (see Beer *et al.*, 2011a).

For the post-training test, characteristic cross-modal facilitation and inhibition were seen in response time VEs. However, these characteristic VEs were only ob-



(a)



(b)

Figure 2. Pre-training cross-modal validity effects. (a) Sounds appeared on the same side as the visual stimulus on valid trials and on the opposite side on invalid trials. Response time (RT) validity effects (VEs) were calculated by subtracting response times for valid trials from those for invalid trials. A positive VE means that responses were faster on valid trials than on invalid trails (see short SOA). A negative VE means that responses were slower on valid trials than on invalid trials (see long SOA). The decrease in valid *versus* invalid measures seen at long SOAs is called inhibition of return (IOR). The data shown here are for informational purposes and do not relate directly to this study; these data represent the natural VE at a visual location aligned with the sound cue, collected for a previous experiment. (b) Cross-modal response time VEs were not significant for either SOA or task (orientation (left) or object (right) discrimination) at any location prior to training. Note that no eye, orientation or location has been trained prior to this test. Therefore, these graphs represent data pooled across eyes, orientations and locations. Error bars represent the 95% confidence interval; $n = 11$.

served at the trained proximal location (P_t) and only on trials with the trained orientation in the trained eye: for the 150 ms SOA, responses were faster when preceded by valid sounds than invalid sounds ($\mu_{\mathrm{VE}} \pm CI_{\mathrm{VE}} = 8.17 \pm 14.24$; CI = confidence interval of the mean). This effect was significantly different from the pre-training VE ($t(10) = -3.72$, $p = 0.004$). For the 1000 ms SOA, responses were slower when preceded by valid sounds (-4.56 ± 29.36) and this effect was also significant from the pre-training VE ($t(10) = 2.30$, $p = 0.044$) (see Table A1 for VE averages (over subjects) from Tests 1 and 2, pre- and post-training, respectively).

Figure 3 illustrates subject averages of the training induced changes in response time VE for orientation discrimination (ΔVE = post-training VE minus pre-training VE; values from Fig. 3 can be calculated from the response time data in Table A1). Cross-modal TIPL resulted in changes that were specific to the trained location (P_t), trained orientation and trained eye (Fig. 3). Two-way within-subject ANOVAs showed significant interactions of eye by location for both the short (150 ms;

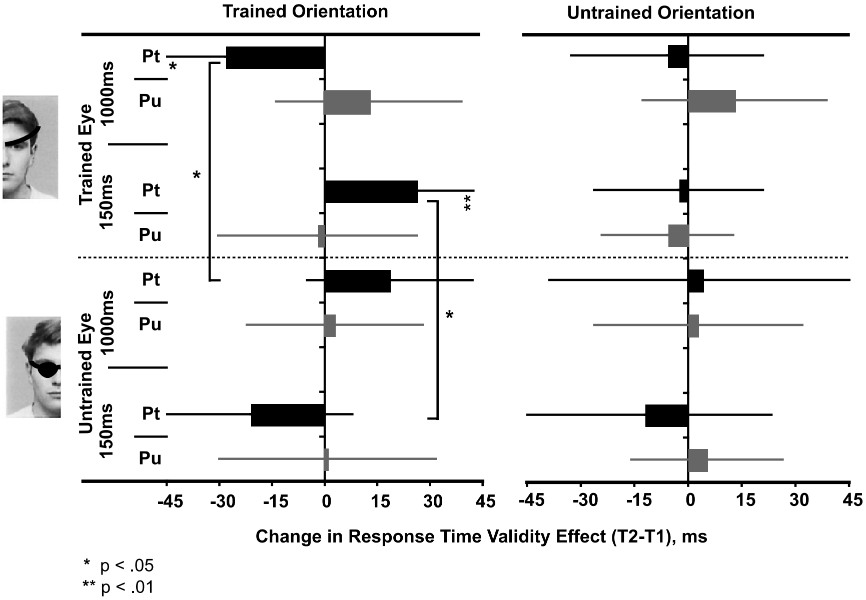


Figure 3. Changes in cross-modal validity effects (post-minus pre-training) for orientation discrimination. Changes in VE are displayed as the difference in response time VEs from pre- to post-training test sessions. Spatially specific realignment of the cross-modal facilitation effect was seen for the short SOA as a significant increase in response time VE at the trained location (P_t) only for the orientation and eye exposed during training sessions ($p < 0.005$). An opposite effect was observed in the untrained eye for the same (trained) orientation and location (opposition of effects seen in the trained *versus* the untrained eye: $p = 0.01$). Significant increases in cross-modal inhibition were seen for the long SOA at the trained location, specific to the trained orientation and eye ($p < 0.05$). An opposite effect was observed in the untrained eye for the same (trained) orientation and location (trained *versus* untrained eye: $p = 0.01$). Note that the trained eye was exposed, while the untrained eye was patched, during training sessions. Error bars represent the 95% confidence interval; $\alpha = 0.05$; $n = 11$.

($F(1, 80) = 4.58$, $p = 0.035$)) and long (1000 ms; ($F(1, 80) = 4.39$, $p = 0.039$)) SOA. For the short SOA (150 ms), cross-modal TIPL training resulted in an increase of cross-modal facilitation at the trained location (P_t) seen as significantly increased response time VE ($\Delta\text{VE} = 26.69 \pm 15.99$) for the orientation and eye exposed during training sessions (*versus* no change: $t(10) = 3.72$, $p = 0.004$). This increase differed significantly from the VE change ($\Delta\text{VE} = -20.89 \pm 29.11$) for the same (trained) orientation and location in the untrained eye (trained *versus* untrained eye: $t(10) = 3.00$, $p = 0.013$). A similar pattern emerged for the long (1000 ms) SOA. Training led to a significant decrease ($\Delta\text{VE} = -28.00 \pm 27.15$) of the VE (increase of cross-modal IOR) at the trained location (P_t) specific to the orientation and eye exposed during training sessions (*versus* no change: $t(10) = -2.30$, $p = 0.044$). This increased IOR differed significantly from the VE change ($\Delta\text{VE} = 18.91 \pm 23.56$) in the untrained eye for the same trained orientation and location (trained *versus* untrained eye: $t(10) = -3.10$, $p = 0.011$). There were no significant changes for untrained locations or the untrained orientation in either eye. No significant changes in pre- to post-training performance were observed on the object discrimination task (see Fig. A1); nor for accuracy measurements in any stimulus condition, eliminating concern of any speed-accuracy trade off.

Performance during the training sessions did not vary significantly across sessions. However, subjects tended to become more accurate and to respond faster from the first to the last training session (see Fig. A2).

4. Discussion

Training with task-irrelevant misaligned audio-visual stimuli changed cross-modal interactions. Prior to training, sounds had no effect on misaligned visual stimuli. After training, those sound–Gabor pairs that were tied with a task-relevant target during training (at P_t) showed enhanced cross-modal facilitation for brief SOAs and stronger inhibition for longer SOAs. Training had no effect on those sound–Gabor pairs that were not tied with a task-relevant target (at P_u). This finding replicates a previous study showing a similar shift of cross-modal interactions from aligned visual field locations with TIPL using misaligned sound–Gabor pairs (Beer *et al.*, 2011a).

The results further show a strong specificity for stimulus attributes involved in the cross-modal training. Perceptual learning resulted in enhanced cross-modal facilitation (and inhibition) only at the trained location (P_t), the trained orientation and the trained eye. Learning affected neither object discrimination nor the other Gabor orientations, nor discrimination at the untrained location (P_u). Learning effects on the untrained eye even tended to be in the opposite direction as expected (though not significant). This specificity for trained stimulus attributes is similar to previous findings from visual-only PL, where the learning does not transfer to a different orientation, location or eye from that exposed during the perceptual training (Ahissar and Hochstein, 1997; Dill, 2002; Fahle, 2005; Fiorentini and Berardi, 1980). The specificity of PL effects is often assumed to reflect plasticity in early sensory cortex (Karni and Sagi, 1991; Schoups *et al.*, 1995). For instance, cells in early visual processing areas are tuned to similar orientations and receive retinotopic inputs with small receptive fields from only one eye (Dill, 2002; Mishkin *et al.*, 1983). Along this line of reasoning, the shift of cross-modal effects observed in our study — which were highly specific to location, orientation, and eye — is consistent with the idea that neural circuits in early sensory cortex are involved in cross-modal perceptual learning.

Of note, Mollon and Danilova (1996) raised concerns about the suggestion that specificity implies the involvement of early sensory brain areas. They argue that specific stimulus attributes such as location or orientation could also be encoded in more central (higher-level) regions, that is, by stimulus-specific wiring. Indeed, some research and modelling highlights that specificity of visual perceptual learning can be accounted for without representation changes, e.g., some PL effects partially transfer across retinal locations (Law and Gold, 2008; Xiao *et al.*, 2008; see Petrov *et al.*, 2005, for discussion). Our data showed almost no transfer across location and orientation. However, it must be pointed out that V1 has cells that are not specific to orientation, some collicular cells have very large receptive fields, and visual areas later in the feedforward pathway, such as the inferotemporal area, have orientation specific cells and cells with small receptive fields (Sary *et al.*, 1995; Tanaka *et al.*, 1991). Interestingly, learning effects tested in the untrained eye tended to be even opposite to those in the trained eye (Fig. 3). To our knowledge, negative transfer across eyes has not been observed before and we can only speculate about it. It might result from lateral competition between monocular neurons in V1: i.e., representing the trained *versus* the untrained eye (Tong *et al.*, 2006). While we suggest that ocular specificity is difficult to account for in a read-out model (given that monocular cells are rare past V1), further experiments, using more direct methods, will be required to verify the exact locus of the learning effects identified in our study.

Previous research has shown that cross-modal cues are able to 'boost' visual processing (Kim *et al.*, 2008; Seitz *et al.*, 2006a; see Shams and Seitz, 2008, for review) early in the visual hierarchy. Moreover, cross-modal perceptual learning also affects temporal processing (Alais and Cass, 2010). Training on audio-visual temporal order-judgments enhanced visual temporal perception, but purely visual training had no effect on audio-visual temporal processing. This finding is consistent with the notion that cross-modal interactions affect modality-specific processing. However, the authors also found that cross-modal (audio-visual) perceptual learning had no effect on auditory temporal processing suggesting that the nature of cross-modal interactions is complex.

Multimodal integration has traditionally been discussed as occurring in higher cortical processing areas, such as polysensory areas in temporal, parietal and frontal cortices (Beauchamp, 2005; Calvert and Thesen, 2004; Cappe and Barone, 2005). In addition to these defined polysensory areas, recent anatomical and physiological findings reveal the presence of multimodal interplay between unisensory cortical processing areas and subcortical structures. For instance, single-cell recordings from V1 of rhesus macaques revealed significantly reduced response latencies from these cells in response to visuo-auditory stimuli when compared to visual only stimuli (Wang *et al.*, 2008). Falchier *et al.* (2002) revealed direct projections from auditory cortex (including A1) and polysensory temporal lobe (STP) to peripheral V1 using retrograde tracers in cynomolgus monkeys. More recently, similar white matter connections between auditory cortex and the calcarine sulcus were shown in humans by means of diffusion-weighted magnetic resonance imaging (Beer *et al.*, 2011b). Multimodal interactions go deep into the brain where integration may also occur subcortically, based on differences in neuronal responses in deep layers of the superior colliculus (Meredith and Stein, 1986) and anatomical support for multisensory (audio-tactile) integration found in the thalamus of the macaque (Cappe and Barone, 2005). These low-level interactions between sensory areas provide a possible neural basis for the behavioural effects found in the present study.

We conclude that audio-visual cross-modal TIPL affects low level visual processing. The present findings support the presence of plastic multisensory interactions in unisensory areas. Our finding might pave the way for more specific audio-visual paradigms, e.g., cross-modal regimens that could drive bottom-up visual rehabilitation with the aid of auditory cues. Further research into the mechanism(s) responsible for cross-modal plasticity should include analysis of both known polysensory and primary sensory cortices. If changes in neuronal response and connectivity throughout the perceptual pathways were systematically compared and contrasted, we may begin to elucidate the origins of cross-modal plasticity.

Acknowledgements

We would like to thank Neha Rana and Priya Sekar for their assistance running subjects for this experiment. NIH R01 EY015980 and NIH RO1 AG031941 supported this work.

M.A.B. designed and ran the experiment, analyzed the data and wrote the paper, A.L.B. conceived the cross-modal training paradigm and edited the paper, A.R.S. assisted with experimental design and edited the paper, and T.W. conceived the training paradigm and supervised the project.

References

Ahissar, M. and Hochstein, S. (1997). Task difficulty and the specificity of perceptual learning, *Nature* **387**, 401–406.

Alais, D. and Cass, J. (2010). Multisensory perceptual learning of temporal order: audiovisual learning transfers to vision but not audition, *PLoS One* **5**, e11283.

Beauchamp, M. S. (2005). See me, hear me, touch me: multimodal integration in lateral occipital–temporal cortex, *Curr. Opin. Neurobiol.* **15**, 145–153.

Beer, A. L., Batson, M. A. and Watanabe, T. (2011a). Multisensory perceptual learning escapes both fast and slow mechanisms of cross-modal processing, *Cognit. Affect. Behav. Neurosci.* **11**, 1–12.

Beer, A. L., Plank, T. and Greenlee, M. W. (2011b). Diffusion tensor imaging shows white matter tracts between human auditory and visual cortex, *Exper. Brain Res.* **213**, 299–308.

Beer, A. L. and Watanabe, T. (2009). Specificity of auditory-guided visual perceptual learning suggests cross-modal plasticity in early visual cortex, *Exper. Brain Res.* **198**, 353–361.

Brainard, D. H. (1997). The psychophysics toolbox, *Spatial Vision* **10**, 433–436.

Calvert, G. A. and Thesen, T. (2004). Multisensory integration: methodological approaches and emerging principles in the human brain, *J. Physiol. (Paris)* **98**, 191–205.

Cappe, C. and Barone, P. (2005). Heteromodal connections supporting multisensory integration at low levels of cortical processing in the monkey, *Eur. J. Neurosci.* **22**, 2886–2902.

Dill, M. (2002). Specificity *versus* invariance of perceptual learning: the example of position, in: *Perceptual Learning*, M. Fahle and T. Poggio (Eds), Chapter 12. MIT Press, Cambridge, MA, USA.

Driver, J. and Noesselt, T. (2008). Multisensory interplay reveals cross-modal influences on ‘sensory-specific’ brain regions, neural responses, and judgments, *Neuron* **57**, 11–23.

Eckert, M. A., Kamdar, N. V., Change, C. E., Beckmann, C. F., Greicius, M. D. and Menon, V. (2008). A cross-modal system linking primary auditory and visual cortices: evidence from intrinsic fMRI connectivity analysis, *Hum. Brain Mapp.* **29**, 848–857.

Eimer, M., van Velzen, J. and Driver, J. (2002). Cross-modal interactions between audition, touch, and vision in endogenous spatial attention: ERP evidence on preparatory states and sensory modulations, *J. Cognit. Neurosci.* **14**, 1–18.

Fahle, M. (2005). Perceptual learning: specificity *versus* generalization, *Curr. Opin. Neurobiol.* **15**, 154–160.

Falchier, A., Clavagnier, S., Barone, P. and Kennedy, H. (2002). Anatomical evidence of multimodal integration in primate striate cortex, *J. Neurosci.* **22**, 5749–5759.

Fiorentini, A. and Berardi, N. (1980). Perceptual learning specific for orientation and spatial frequency, *Nature* **278**, 43–44.

Foxe, J. J. and Schroeder, C. E. (2005). The case for feedforward multisensory convergence during early cortical processing, *Neuroreport* **16**, 419–423.

Franko, E., Seitz, A. R. and Vogels, R. (2010). Dissociable neural effects of long term stimulus-reward pairing in macaque visual cortex, *J. Cognit. Neurosci.* **22**, 1425–1439.

Grantham, D. W. (1986). Detection and discrimination of simulated motion of auditory targets in the horizontal plane, *J. Acoust. Soc. Amer.* **79**, 1939–1949.

Karni, A. and Sagi, D. (1991). Where practice makes perfect: evidence for primary visual cortex plasticity, *Proc. Natl. Acad. Sci. USA* **88**, 4966–4970.

Kim, R. S., Seitz, A. R. and Shams, L. (2008). Benefits of stimulus congruency for multisensory facilitation of visual learning, *PLoS ONE* **3**, e1532.

Klein, R. M. (2000). Inhibition of return, *Trends Cognit. Sci.* **4**, 138–146.

Law, C. and Gold, J. I. (2008). Neural correlates of perceptual learning in a sensory-motor but not a sensory cortical area, *Nature Neurosci.* **11**, 505–513.

McDonald, J. J., Teder-Sälejärvi, W. A. and Hillyard, S. A. (2000). Involuntary orienting to sound improves visual perception, *Nature* **407**, 906–908.

Meredith, M. A. and Stein, B. E. (1986). Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration, *J. Neurophysiol.* **56**, 640–662.

Meyer, G. F., Wuerger, S. M., Rohrbein, F. and Zetzsche, C. (2005). Low-level integration of auditory and visual motion signals requires spatial co-localisation, *Exper. Brain Res.* **166**, 538–547.

Mishkin, M., Ungerleider, L. G. and Macko, K. A. (1983). Object visions and spatial visions: two cortical pathways, *Trends Neurosci.* **6**, 414–417.

Mollon, J. D. and Danilova, M. V. (1996). Three remarks on perceptual learning, *Spatial Vision* **10**, 51–58.

Nishina, S., Seitz, A. R., Kawato, M. and Watanabe, T. (2007). Effect of spatial distance to the task stimulus on task-irrelevant perceptual learning of static Gabors, *J. Vision* **7**, 1–10.

Petrov, A. A., Dosher, B. A. and Lu, Z. L. (2005). The dynamics of perceptual learning: an incremental reweighting model, *Psychol. Rev.* **112**, 715–743.

Rockland, K. S. and Ojima, H. (2003). Multisensory convergence in calcarine visual areas in macaque monkey, *Int. J. Psychophysiol.* **50**, 19–26.

Rosenthal, O. and Humphreys, G. W. (2010). Perceptual organization without perception: the subliminal learning of global contour, *Psychol. Sci.* **21**, 1751–1758.

Sary, G., Vogels, R., Kovacs, G. and Orban, G. A. (1995). Responses of monkey inferior temporal neurons to luminance-, motion- and texture-defined gratings, *J. Neurophysiol.* **73**, 1341–1354.

Schoups, A. A., Vogels, R. and Orban, G. (1995). Human perceptual learning in identifying the oblique orientation: retinotopy, orientation specificity and monocularity, *J. Physiol.* **483**, 797–810.

Seitz, A. R., Kim, R. and Shams, L. (2006a). Sound facilitates visual learning, *Curr. Biol.* **16**, 1422–1427.

Seitz, A. R., Kim, D. and Watanabe, T. (2009). Rewards evoke learning of unconsciously processed visual stimuli in adult humans, *Neuron* **12**, 700–707.

Seitz, A. R., Nanez, J. E., Holloway, S. R. and Watanabe, T. (2005). Visual experience can substantially alter critical flicker fusion thresholds, *Hum. Psychopharmacol.* **20**, 55–60.

Seitz, A. R., Nanez, J. E., Holloway, S. R. and Watanabe, T. (2006b). Perceptual learning of motion leads to faster flicker perception, *PLoS ONE* **1**, e28.

Seitz, A. R., Protopapas, A., Tsushima, Y., Vlahou, E. L., Gori, S., Grossberg, S. and Watanabe, T. (2010). Unattended exposure to components of speech sounds yields same benefits as explicit auditory training, *Cognition* **115**, 435–443.

Seitz, A. R. and Watanabe, T. (2003). Is subliminal learning really passive?, *Nature* **422**, 36.

Seitz, A. R. and Watanabe, T. (2005). A unified model for perceptual learning, *Trends Cognit. Sci.* **9**, 329–334.

Seitz, A. R. and Watanabe, T. (2009). The phenomenon of task-irrelevant perceptual learning, *Vision Res.* **49**, 2604–2610.

Shams, L. and Seitz, A. R. (2008). Benefits of multisensory learning, *Trend Cognit. Sci.* **12**, 411–417.

Shinn-Cunningham, B. (2000). Adapting to remapped auditory localization cues: a decision theory model, *Percept. Psychophys.* **62**, 33–47.

Spence, C. and Driver, J. (1997). Audio-visual links in exogenous covert spatial orienting, *Percept. Psychophys.* **59**, 1–22.

Spence, C. and Driver, J. (1998). Auditory and audio-visual inhibition of return, *Percept. Psychophys.* **60**, 125–139.

Tanaka, K., Saito, H., Fukada, Y. and Moriya, M. (1991). Coding visual images of objects in the inferotemporal cortex of the Macaque monkey, *J. Neurophysiol.* **66**, 170–189.

Thesen, T., Vibell, J. F., Calvert, G. A. and Österbauer, R. A. (2004). Neuroimaging of multisensory processing in vision, audition, touch, and olfaction, *Cognit. Process* **5**, 84–93.

Tong, F., Meng, M. and Blake, R. (2006). Neural basis of binocular rivalry, *Trends Cognit. Sci.* **10**, 502–511.

Tong, F., Nakayama, K., Vaughan, J. T. and Kanwisher, N. (1998). Binocular rivalry and visual awareness in human extrastriate cortex, *Neuron* **21**, 753–759.

Wang, Y., Celebrini, S., Trotter, Y. and Barone, P. (2008). Visuo-auditory interactions in the primary visual cortex of the behaving monkey: electrophysiological evidence, *BMC Neurosci.* **9**, 79.

Watanabe, T., Náñez, J. E., Koyama, S., Mukai, I., Liederman, J. and Sasaki, Y. (2002). Greater plasticity in lower-level than higher-level visual motion processing in a passive perceptual learning task, *Nat. Neurosci.* **5**, 1003–1009.

Watanabe, T., Náñez, J. and Sasaki, Y. (2001). Perceptual learning without perception, *Nature* **413**, 844–848.

Xiao, L. Q., Zhang, J. Y., Wang, R., Klein, S. A., Levi, D. M. and Yu, C. (2008). Complete transfer of perceptual learning across retinal locations enabled by double training, *Curr. Biol.* **18**, 1922–1926.

Zwiers, M. P., Van Opstal, A. J. and Paige, G. D. (2003). Plasticity in human sound localization induced by compressed spatial vision, *Nature Neurosci.* **6**, 175–181.

Appendix

Table A1.
Accuracy and response time validity effect measures from Test 1 and Test 2

	Trained orientation				Untrained orientation			
	150 ms		1000 ms		150 ms		1000 ms	
	Pu	Pt	Pu	Pt	Pu	Pt	Pu	Pt
Trained eye validity effect — Test 1								
Accuracy	0.0195	0.0442	−0.0483	0.0050	0.0201	0.0280	−0.0127	−0.0558
RT	9.4	−18.5*	−3.9	23.4	−2.5	10.6	−14.4	5.5
Trained eye validity effect — Test 2								
Accuracy	0.0184	−0.0079	0.0228	−0.0413	−0.0088	0.0059	−0.0219	0.0341
RT	7.7	8.2**	9.0	−4.6*	−8.1	8.0	−1.3	−0.2
Untrained eye validity effect — Test 1								
Accuracy	0.0297	0.0084	−0.0156	−0.0135	−0.0115	0.0035	−0.0103	−0.0053
RT	2.5	0.9	2.4	−3.2	2.3	7.7	0.3	−13.4
Untrained eye validity effect — Test 2								
Accuracy	0.0161	0.0308	0.0122	0.0215	−0.0113	0.0161	0.0127	−0.0006
RT	3.7	−20.0	5.6	15.7	7.7	−4.2	3.2	−9.2

* $p < 0.05$; ** $p < 0.01$.

The table shows the mean validity effect (VE) on proportion correct responses (accuracy) and response times (RT, in ms) from Test 1 (pre-training) and Test 2 (post-training); separated by eye (trained or untrained). Note that the trained eye was exposed, while the untrained eye was patched, during training sessions. Asterisks indicate that the pre-training (Test-1) VE differed significantly from zero (i.e., no VE, $p < 0.01$) or that the post-training (Test-2) VE differed significantly from the pre-training VE for the matching condition in a paired t-test comparing the data from each test ($p < 0.05$). The Test 2 significance seen here for the trained conditions is mirrored in the difference data graphed in Fig. 3, which displays the values obtained by subtracting Test 1 VEs from Test 2 VEs; and, as follows, in both instances the difference is also significant from zero (i.e., from no change in VE). See the results section for more information. Also note that no condition has been trained prior to Test 1, and data at this stage cannot reflect trained *versus* untrained. For this reason, in the experimental analysis all conditions were averaged for conclusions from Test 1 (see Fig. 2b for averaged values), $\alpha = 0.05$; $n = 11$.

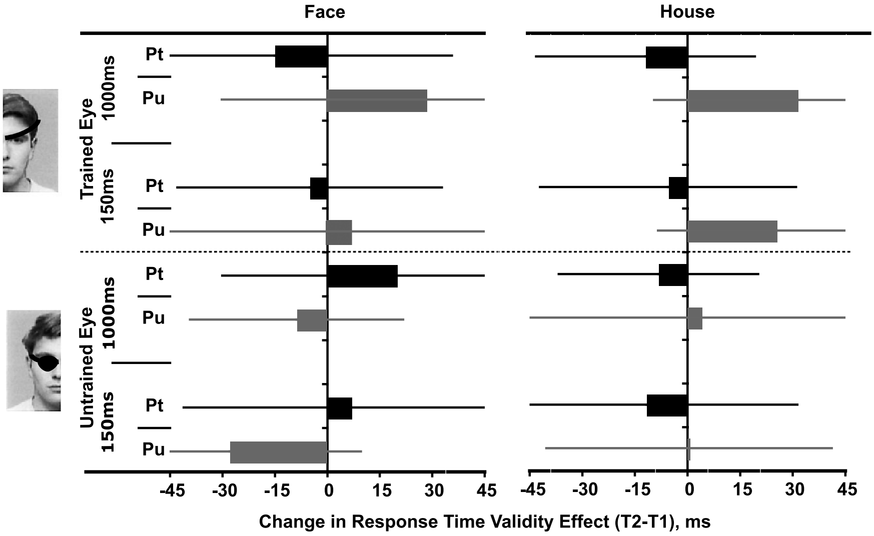


Figure A1. Changes in cross-modal VEs (post-minus pre-training) for object discrimination. Object stimuli (faces and houses) were not presented during training sessions. There are no significant changes for any untrained locations or stimuli. Note that the trained eye was exposed, while the untrained eye was patched, during training sessions. Error bars represent the 95% confidence interval; $n = 11$.

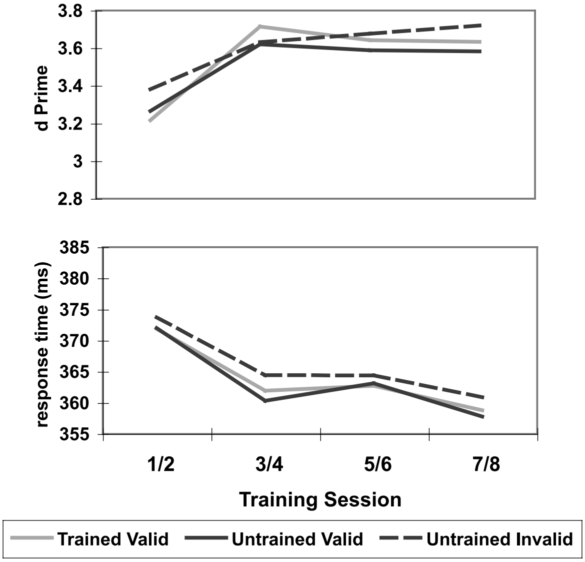


Figure A2. Performance during training sessions. Training data was pooled across two consecutive sessions as the target shape (circle and square) alternated across sessions. There were no significant effects on target detection across training sessions. However, subjects tended to show increased discrimination performance (d′) and decreased response time from session 1/2 to session 7/8; $n = 11$.