BRILL

Multisensory Research 26 (2013) 307–316

MULTISENSORY RESEARCH

brill.com/msr

Cross-correlation between Auditory and Visual Signals Promotes Multisensory Integration

Cesare V. Parise [1,2,3,*], **Vanessa Harrar** [3], **Marc O. Ernst** [1,2] **and Charles Spence** [3]

[1] Max Planck Institute for Biological Cybernetics and Bernstein Center for Computational Neuroscience, Tübingen, Germany

[2] Cognitive Neuroscience Department and Cognitive Interaction Technology — Center of Excellence, Bielefeld University, Germany

[3] Department of Experimental Psychology, University of Oxford, UK

Received 19 October 2012; accepted 9 February 2013

Abstract

Humans are equipped with multiple sensory channels that provide both redundant and complementary information about the objects and events in the world around them. A primary challenge for the brain is therefore to solve the 'correspondence problem', that is, to bind those signals that likely originate from the same environmental source, while keeping separate those unisensory inputs that likely belong to different objects/events. Whether multiple signals have a common origin or not must, however, be inferred from the signals themselves through a causal inference process.

Recent studies have demonstrated that cross-correlation, that is, the similarity in temporal structure between unimodal signals, represents a powerful cue for solving the correspondence problem in humans. Here we provide further evidence for the role of the temporal correlation between auditory and visual signals in multisensory integration. Capitalizing on the well-known fact that sensitivity to crossmodal conflict is inversely related to the strength of coupling between the signals, we measured sensitivity to crossmodal spatial conflicts as a function of the cross-correlation between the temporal structures of the audiovisual signals. Observers' performance was systematically modulated by the cross-correlation, with lower sensitivity to crossmodal conflict being measured for correlated as compared to uncorrelated audiovisual signals. These results therefore provide support for the claim that cross-correlation promotes multisensory integration. A Bayesian framework is proposed to interpret the present results, whereby stimulus correlation is represented on the prior distribution of expected crossmodal co-occurrence.

This article is part of the 13th IMRF collection, guest edited by M. Murray, C. Spence, and L. R. Harris.

* To whom correspondence should be addressed. E-mail: cesare.parise@uni-bielefeld.de

DOI:10.1163/22134808-00002417

Keywords
Cross-correlation, multisensory integration, audition, vision

1. Introduction

Solving the crossmodal correspondence problem, that is, inferring which signals have a common underlying cause, and hence provide redundant information that should be integrated, is a crucial task for a perceptual system trying to make sense of multiple sensory inputs (Ernst and Bülthoff, 2004; Shams and Beierholm, 2010; Welch and Warren, 1980). Previous research has demonstrated that spatiotemporal cues and prior knowledge are exploited by the human brain in order to solve the correspondence problem (Bresciani *et al.*, 2005; Gepshtein *et al.*, 2005; Körding *et al.*, 2007; Parise and Spence, 2009; Vatakis and Spence, 2007). However, most of these studies used rather simple stimuli (such as isolated beeps, flashes, and taps), and the role of the temporal structure of the signals has remained relatively unexplored. Recently, it has been shown that the similarity in the fine-temporal structure of visual and auditory signals, that is their cross-correlation, also plays an important role in causal inference, resulting in a statistically optimal integration for correlated but not for uncorrelated signals (Parise *et al.*, 2012).

The cross-correlation between multiple sensory signals is an important cue for causal inference: signals originating from a single event normally share a tight temporal relation, due to their dependence on the same underlying event. Conversely, when multiple signals are generated by independent physical events, their temporal structures are normally unrelated. Here we further explore the role of similarity between auditory and visual signals (measured with cross-correlation, for simplicity in this paper often referred to as correlation) in causal inference by parametrically manipulating the cross-correlation of the signals. Specifically, we hypothesized that similar signals, i.e. signals with a similar fine-temporal structure, and thus a high cross-correlation, are more likely inferred to originate from a single underlying event and hence will be integrated more strongly.

Previous research has proven that when conflicting multisensory signals (such as auditory and visual signals coming from discrepant spatial positions) are integrated, observers lose sensitivity to such conflicts (Bresciani *et al.*, 2006; Ernst, 2005, 2007; Hillis *et al.*, 2002). Notably, this loss of sensitivity to intersensory conflict, which reflects the cost of sensory integration, is proportional to the strength of the coupling between the unisensory signals: the stronger the coupling, the lower the sensitivity. In the extreme case in which multisensory information is completely fused (strong coupling), observers have only access to the integrated percept. This implies that observers

cannot directly compare the information from the different senses, and, as a consequence, sensitivity to intersensory conflicts is fully lost. Conversely, when signals are not integrated, observers have complete access to the individual unisensory information, and hence they can easily report on the presence of intersensory conflicts.

In the present study, we set out to measure how such behavioural costs associated with multisensory integration are modulated by the correlation between multiple sensory signals presented in different modalities. Human observers were presented with simultaneous visual and auditory stimuli with different levels of correlation. The stimuli came from discrepant locations along the horizontal axis, and, using the same experimental paradigm used by Parise and Spence (2009), observers had to report the left/right location of the auditory stimulus with respect to the visual stimulus (see Fig. 1B). If stronger integration takes place between cross-correlated stimuli, one would expect left/right discrimination thresholds to increase (that is, sensitivity should decrease) with increasing cross-correlation.

2. Materials and Methods

Fifteen participants, with normal or corrected-to-normal vision and normal audition, made unspeeded forced choice judgments as to whether an auditory stimulus was presented to the left or right of a visual stimulus.

The visual stimuli comprised trains of white Gaussian blobs each projected for 16 ms against a black background on a sound-transparent screen (width: 107.7 cm; height: 80.8 cm; Fig. 1B). The standard deviation of the Gaussian luminance profile of the blobs subtended 0.8° of visual angle at a viewing distance of 150 cm (a chinrest was used to control the participant's head position). The auditory signals consisted of trains of white noise clicks played from one of eight loudspeakers (separated by 10.5 cm) hidden behind the projection screen. The intensity of the auditory stimuli was randomly jittered from trial to trial (between $\pm 1\%$ of the standard intensity) in order to avoid the possibility that participants might use any potential subtle differences in the intensities of the loudspeakers as auxiliary cues for sound localization.

Each train of visual and auditory stimuli comprised eight blobs or eight clicks, respectively, randomly scattered over a 2 s temporal interval (central frequency = 4 Hz). Each click and blob lasted for 16 ms. A new fine-temporal structure (i.e. train of flashes and/or clicks) was generated for each trial. In trials with correlated signals (cross-correlation = 1, lag 0), the same temporal structure defined the visual and auditory stimuli, whereas, in the remaining trials (cross-correlation = [0.5, 0], lag = 0), we iteratively generated pairs of random sequences and selected those with the desired level of cross-correlation

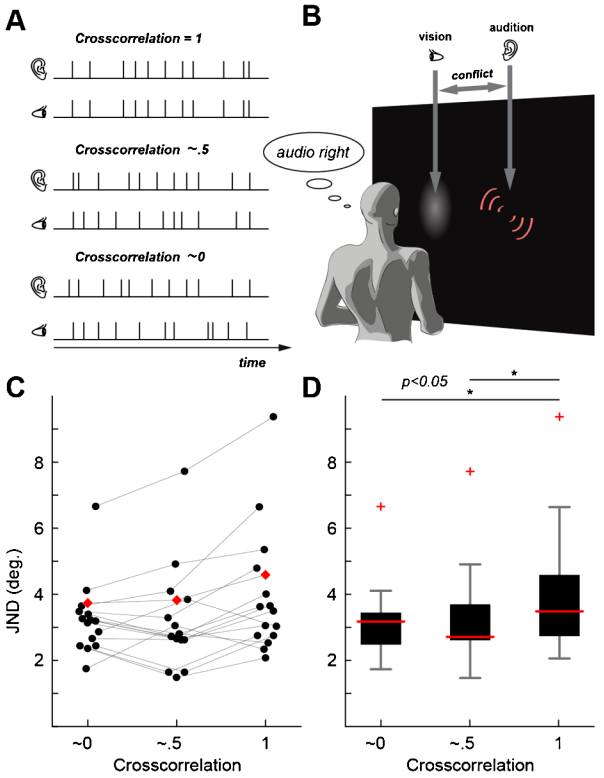

Figure 1. Stimuli, apparatus and results. (A) Examples of the intensity profiles of correlated and uncorrelated audiovisual stimulus pairs. Auditory stimuli consisted of trains of 8 white noise bursts, while visual stimuli consisted of trains of 8 Gaussian blobs presented for one frame (60 Hz projector). The overall duration of visual and auditory stimuli was 2 s. (B) Visual stimuli (blobs) were projected on a white sound-transparent screen placed in front of an array of 8 loudspeakers. This apparatus allowed for the introduction of physical conflicts between the spatial location of the visual and the auditory stimuli. (C) Individual JNDs to audiovisual conflicts in the three audiovisual cross-correlation conditions. The diamonds represent the 'aggregate observers' JNDs, obtained by pooling together the data from all participants. (D) Boxplot of the JNDs. This figure is published in colour in the online version.

(see Fig. 1A and Note 1). The participants were not informed as to the manipulation of the correlation between the signals.

The amount of spatial offset between the visual and auditory stimuli on each trial was determined by a custom adaptive psychophysical procedure based on the Quest (Watson and Pelli, 1983), and modified in order to fit a full psychometric function to the data. The absolute position of the visual and auditory stimuli on the screen was randomly selected on each trial. The participants' task was to judge whether the auditory stimuli were presented from the left or right of the visual stimulus stream. The presentation of the stimuli and the

collection of the responses were controlled by custom software based on the Psychtoolbox 3 (Kleiner *et al.*, 2007). Overall, the experiment consisted of 414 trials (138 trials per each of the 3 correlation conditions) and lasted about 1.5 h.

Before starting the experiment, participants performed a preliminary training session with the same stimuli and task as in the experiment proper. The training session consisted of 36 trials, and participants received feedback as to their performance on a trial-by-trial basis.

3. Results

Separate psychometric functions were calculated for each level of correlation and for each participant by fitting the ratios of 'right' responses plotted against spatial displacement (measured in degrees of visual angle) with a cumulative Gaussian distribution (see Wichmann and Hill, 2001, for full procedural details). The just noticeable differences (JNDs) were calculated from the psychometric curves by subtracting the level of audiovisual displacement at which participants made 75% 'right' responses from the displacement at which they made 25% 'right' responses and halving the result (see Fig. 1C–D). Temporal correlation significantly influenced the reliability of participants' estimates as revealed by the Friedman test ($\chi^2(2) = 8.93,\ p = 0.0115$), with larger discrimination thresholds reported for correlated trials (median = 3.5 deg) than for partially correlated trials (median = 2.7 deg), and uncorrelated trials (median = 3.2 deg), corresponding to a 28% increase with respect to the 0.5 cross-correlation condition and to a 10% increase with respect to the uncorrelated condition. *Post-hoc* comparisons using the Wilcoxon signed rank test ($\alpha = 0.0167$) demonstrate that precision is significantly lower in the correlated condition than both the partially correlated condition (Signed Rank = 12; $p = 0.004$) and the uncorrelated condition (Signed Rank = 14; $p = 0.006$), thus providing support for the claim that enhanced multisensory integration occurs for correlated pairs of audiovisual stimuli. No statistical difference was found between uncorrelated and partially correlated JNDs (Signed Rank = 57, $p = 0.9$).

4. Discussion

As hypothesized, the participants were less sensitive to spatial conflicts between vision and audition when the stimuli were highly correlated (cross-correlation = 1) than when they were less or un-correlated (see Fig. 1C–D). These results demonstrate that cross-correlation modulates multisensory integration, thus leading to behavioural costs normally associated with sensory integration. While previous research has already demonstrated the benefits as-

sociated with the integration of correlated signals (Parise *et al.*, 2012), the results of the present study demonstrate their costs. Putting together the previous and the present results describing the effects of cross-correlation between the fine temporal structure of signals presented to different modalities, both in terms of costs and benefits, we hereby provide strong converging evidence for the effect of cross-correlation in multisensory integration.

Recently, several implementations of Bayesian models have been proposed in order to describe the effects of causal inference in multisensory integration (Ernst, 2007; Ernst and Di Luca, 2011; Knill, 2007; Körding *et al.*, 2007). In the model proposed by Ernst and Di Luca (2011), humans operate as optimal integrators, combining multiple sensory signals with prior knowledge in order to derive more precise combined estimates. The mapping between multisensory signals can be modeled as Bayesian priors describing the expected joint distribution of those signals (Ernst, 2005, 2007). Such priors should be conditioned on the correlation between the signals: i.e. *a priori* signals are likely to be co-located or arising in close spatial proximity only when they arise from the same object, and when they are from the same object the signals are likely to be correlated as well. On the other hand, if the signals are not from the same object, the spatial relationship between any visual and auditory object is random, and so is their fine-temporal structure.

In order to model the effects of a Bayesian prior describing the statistical co-occurrence (here co-location) of two signals, the multisensory integration problem should be framed within a two-dimensional audio-visual space (see Fig. 2). The axes defining this two-dimensional space represent a property (here the spatial location of the stimuli along the horizontal axis) of visual and auditory signals. The prior distribution representing the expected mapping between the visual and auditory signals, the coupling prior, can be modeled as a bivariate Gaussian distribution with the axis aligned along the diagonals. Assuming that the probability of encountering the two signals is independent of the signals' mean value, the variance of the prior along the positive diagonal (i.e. the identity line, perfect co-location) will tend to infinity. The mapping uncertainty between the signals is instead encoded in the variance of the prior along the negative diagonal, as uncorrelated signals are more likely also not co-located: the more certain is the mapping, the lower the variance will be.

According to Bayesian decision theory, the sensory inputs (the likelihood function) and prior knowledge (the prior distribution) are combined to produce a posterior distribution, which determines the final percept. This implies that the strength of coupling scales with the variance of the coupling prior: the lower such variance, the more the posterior would be biased by the prior, and consequently the lower would be the access to intersensory conflict. In the extreme case, where such variance is zero, sensory signals are mandatorily fused and sensitivity to intersensory conflict is lost completely (Hillis *et al.*,

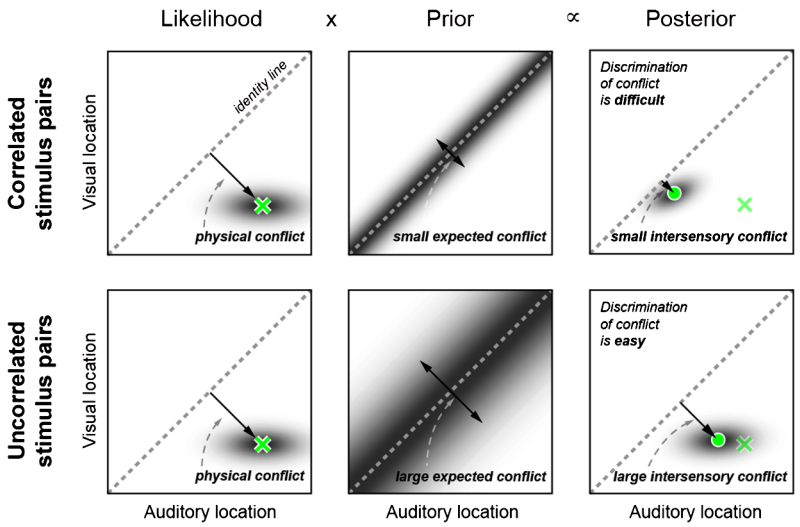

Figure 2. The cross-correlation between the signals is represented on the coupling prior distribution. Multisensory integration results from the combination of sensory inputs (likelihood distribution) and prior knowledge (prior distribution) through Bayes' rule (i.e. likelihood $\times$ prior $\propto$ posterior). The likelihood distribution (left panels) represents the distribution of physical stimuli inducing a given sensory response. The axes represent a property S (e.g. spatial location) of the visual and auditory stimuli. The prior (central panels) represents the expected joint distribution of S from the two sensory channels. A prior narrowly distributed along the identity line indicates that observers have a strong expectation that the two cues are identical and conflict free. In contrast, a flat prior indicates complete uncertainty about the mapping between the two cues (i.e. large and small crossmodal conflicts are equally likely). The integrated percept (the posterior, right panels) is the product of the prior and the likelihood distributions. A prior narrowly distributed along the identity line will bias the percept toward the identity line, thus silencing crossmodal conflicts (see upper panels). Conversely, a shallower prior would leave access to intersensory conflicts (see lower panels). This figure is published in colour in the online version.

2002), conversely, if the variance tends to infinity, the signals are treated as independent with no loss of sensitivity to intersensory conflict.

Therefore, the reduced sensitivity to spatial conflict between correlated audiovisual stimuli measured here suggests that observers have encoded the natural mapping between the sensory signals (e.g. the correlation between the signals) in the coupling prior distribution, and exploit this information in order to integrate them. That is, the expected probability of a large offset between correlated stimuli should be lower than for uncorrelated stimuli (see Fig. 2). Therefore, combining a coupling prior dependent on the correlation between the signals with offset multisensory input according to Bayes' rule, should lead to different results as a function of the temporal correlation of the signals: when they are correlated, the conflict is largely cancelled, whereas when

they are uncorrelated the conflict would still be accessible (see also Bresciani *et al.*, 2006; Ernst, 2005). This prediction is fully supported by the results of the present study; however, it is important to note that the loss of sensitivity for correlated signals reported here is rather small. This implies that, unlike sensory integration within modalities (Hillis *et al.*, 2002), mandatory fusion does not occur between auditory and visual signals, even when they are fully correlated.

A recent study by Raposo *et al.* (2012) comparing the effects of correlated and uncorrelated signals, has demonstrated that in a 'rate estimation' task, the correlation between the signals did not affect observers' (humans and rats) behavioural performance. That is, observers optimally integrated audiovisual rate estimates irrespective of the correlation between the signals (notably sometimes integration was super-optimal, with the behavioural benefits in terms of uncertainty reduction for combined signals exceeding the theoretical limit). Such results are at odds with the present findings, and those of Parise *et al.* (2012), where the correlation between the signals systematically modulated participants' responses. However, it should be noted that Raposo and colleagues' study differed from ours in a number of important ways. First of all, Raposo *et al.* (2012) always provided feedback with respect to their participants' performance, whereas we never provided feedback during the experiment. Providing feedback most likely led to improved performance, so that the lack of a difference between correlated and uncorrelated trials could be due to learning to ignore the fine-temporal structures of the signals as a cue for solving the rate discrimination task, thereby cancelling out any difference between conditions. On a related note, such a training coupled with very long experimental sessions (thousands of trials per observer), might have promoted sensory integration even for uncorrelated signals. Moreover, in Raposo *et al.*'s study, the rate of signal presentation was much higher than in our studies (i.e. between 8.3 and 16.6 Hz in Raposo *et al.*, 2012, *vs*. 4 Hz in the present study and 5 Hz in Parise *et al.*, 2012), and it is not clear whether at such high rates observers could detect the correlation between the signals (Fujisaki and Nishida, 2005, 2010). It will certainly be an interesting matter for future research to explore the role of signals' rate on the effect of temporal correlation in multisensory integration.

The present results highlight the role that temporal correlation plays in multisensory integration. However, it should be noted that the properties of the temporal structure of the signals modulates perceptual binding also within modalities. Lee and Blake (1999), for example, have demonstrated that dynamic visual elements (i.e. rotating disks) whose changes over time followed an identical temporal structure were systematically grouped into a single visual figure (see also Alais *et al.*, 1998).

Therefore, taken together, the effects of temporal correlation highlighted here should be considered as an instance of the manifold effects of temporal structure on sensory integration — both within and across modalities — where temporal correlation might indeed represent the key physical property underlying the Gestalt law of grouping by 'common fate' (Spence, in press; Wertheimer, 1923).

Acknowledgements

C.P. and M.E. were supported by the Bernstein Center for Computational Neuroscience, Tübingen, funded by the German Federal Ministry of Education and Research (BMBF; FKZ: 01GQ1002), and by the EU grant 248587 "THE".

Note

1. Given two continuous real valued functions, f and g, the cross-correlation $r(t)$ at time lag t is defined as:

$$r(t) = \int_{-\infty}^{\infty} f(\tau) g(t + \tau)\, \mathrm{d}\tau,$$

where t is the time-lag between the two signals. For simplicity, here we use the normalized cross-correlation, whereby functions f and g are normalized by subtracting their mean and dividing by their standard deviation and a scaling factor. For each value of t, the normalized cross-correlation returns a value between 1 and -1, where 1 signifies that the two normalized signals are identical while 0 that the two signals are completely different. In the present experiment, the cross-correlation time lag (t) was constrained within ± 100 ms.

References

Alais, D., Blake, R. and Lee, S. H. (1998). Visual features that vary together over time group together over space, *Nat. Neurosci.* **1**, 160–164.

Bresciani, J. P., Dammeier, F. and Ernst, M. (2006). Vision and touch are automatically integrated for the perception of sequences of events, *J. Vision* **6**, 554–564.

Bresciani, J. P., Ernst, M. O., Drewing, K., Bouyer, G., Maury, V. and Kheddar, A. (2005). Feeling what you hear: auditory signals can modulate tactile tap perception, *Exp. Brain Res.* **162**, 172–180.

Ernst, M. O. (2005). A Bayesian view on multimodal cue integration, in: *Perception of the Human Body from the Inside Out*, G. Knoblich, I. Thornton, M. Grosejan and M. Shiffrar (Eds), pp. 105–131. Oxford University Press, New York, NY, USA.

Ernst, M. O. (2007). Learning to integrate arbitrary signals from vision and touch, *J. Vision* **7**, 1–14.

Ernst, M. O. and Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion, *Nature* **415**, 429–433.

Ernst, M. O. and Bülthoff, H. H. (2004). Merging the senses into a robust percept, *Trends Cogn. Sci.* **8**, 162–169.

Ernst, M. O. and Di Luca, M. (2011). Multisensory perception: from integration to remapping, in: *Sensory Cue Integration*, J. Trommershäuser, M. Landy and K. Körding (Eds), pp. 224–250. Oxford University Press, New York, NY, USA.

Fujisaki, W. and Nishida, S. (2005). Temporal frequency characteristics of synchrony–asynchrony discrimination of audio-visual signals, *Exp. Brain Res.* **166**, 455–464.

Fujisaki, W. and Nishida, S. (2010). A common perceptual temporal limit of binding synchronous inputs across different sensory attributes and modalities, *Proc. Royal Soc. B: Biol. Sci.* **277**, 2281–2290.

Gepshtein, S., Burge, J., Ernst, M. O. and Banks, M. S. (2005). The combination of vision and touch depends on spatial proximity, *J. Vision* **5**, 1013–1023.

Hillis, J., Ernst, M. O., Banks, M. and Landy, M. (2002). Combining sensory information: mandatory fusion within, but not between, senses, *Science* **298**, 1627–1630.

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R. and Broussard, C. (2007). What's new in Psychtoolbox-3? *Perception* **36**, 1.1–16.

Knill, D. C. (2007). Robust cue integration: a Bayesian model and evidence from cue-conflict studies with stereoscopic and figure cues to slant, *J. Vision* **7**, 1–24.

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B. and Shams, L. (2007). Causal inference in multisensory perception, *PLoS ONE* **2**, 943.

Lee, S. H. and Blake, R. (1999). Visual form created solely from temporal structure, *Science* **284**, 1165–1168.

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

Parise, C. V., Spence, C. and Ernst, M. O. (2012). When correlation implies causation in multisensory integration, *Curr. Biol.* **22**, 46–49.

Raposo, D., Sheppard, J. P., Schrater, P. R. and Churchland, A. K. (2012). Multisensory decision-making in rats and humans, *J. Neurosci.* **32**, 3726–3735.

Shams, L. and Beierholm, U. R. (2010). Causal inference in perception, *Trends Cogn. Sci.* **14**, 425–432.

Spence, C. (in press). Cross-modal perceptual organization, in: *The Oxford Handbook of Perceptual Organization*, J. Wagemans (Ed.). Oxford University Press, Oxford, UK.

Vatakis, A. and Spence, C. (2007). Crossmodal binding: evaluating the 'unity assumption' using audiovisual speech stimuli, *Percep. Psychophys.* **69**, 744–756.

Watson, A. and Pelli, D. (1983). QUEST — A Bayesian adaptive psychometric method, *Percept. Psychophys.* **33**, 113–120.

Welch, R. and Warren, D. (1980). Immediate perceptual response to intersensory discrepancy, *Psycholog. Bull.* **88**, 638–667.

Wertheimer, M. (1923). Untersuchungen zur Lehre von der Gestalt (Studies on the science of Gestalt). II, *Psychol. Res.* **4**, 301–350.

Wichmann, F. and Hill, N. (2001). The psychometric function: I. Fitting, sampling and goodness of fit, *Percept. Psychophys.* **63**, 1293–1313.