Bayesian Modeling of Perceiving: A Guide to Basic Principles

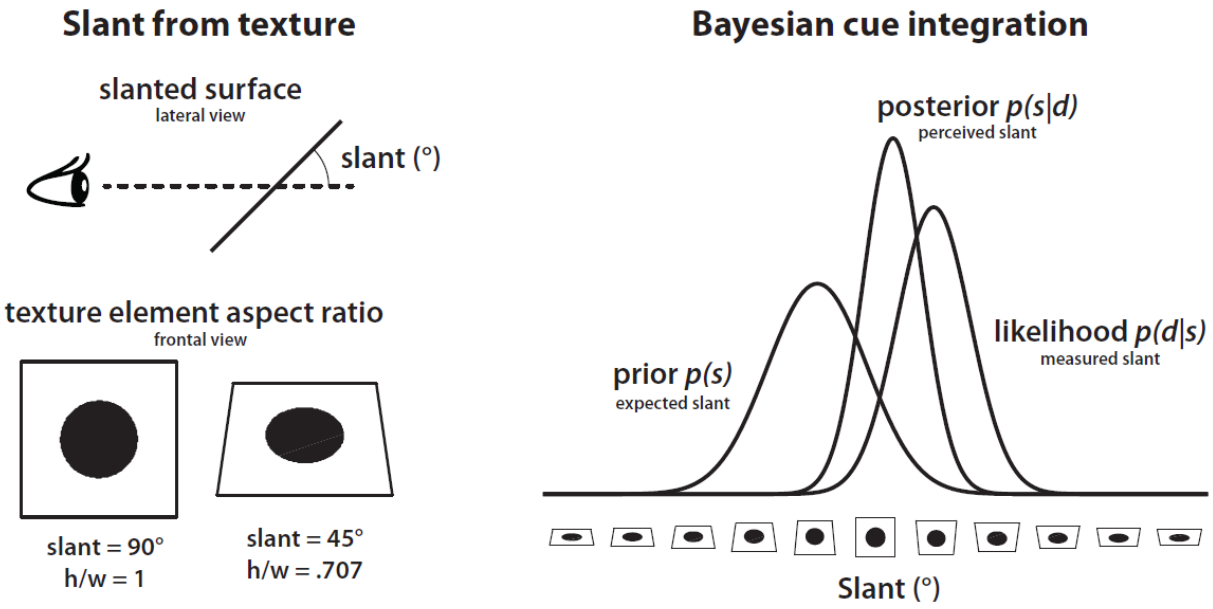
David J. Bennett (Brown), Julia Trommershäuser (NYU), Loes C. J. van Dam (Bielefeld)

Contents:

1. The single cue case. Example: perceiving slant from texture.

We begin with a simplified example concerning the perception of surface slant from sensitivity to texture information (compare Knill 2003, 2007). The basic, perception-science version of Bayes will fall out naturally and intuitively.

In our toy model, suppose that it is assumed that surface texture elements are circular (see Figure **1a**). Suppose a perceiver views a surface head on, looking straight at a circular texture element. We'll say that at 'upright' the slant of the surface is 90 degrees. That circular element will project to a circle on a flat projection plane, which approximates a sensory surface.
The height-to-width ratio of a projected circular texture element is called the 'aspect ratio' of the projected image. So, the aspect ratio of the image when looking straight-on at the (90 degree, upright) surface is 1, because the image is itself circular.



1a **1b**

Figure 1[1]

Now imagine the surface to be slanted back in depth. The projections of the same circular texture element will now be progressively 'thinner' ellipses, with progressively smaller aspect ratios (all fractions of 1, but ever smaller). There is a simple relation between such image aspect ratios and the slant of the surface:

Equation 1). A = sin (S).

('**A**' stands for the aspect ratio. '**S**' is the slant of the surface, measured from the, 'straight on', upright, where **S =90**. If the surface is viewed straight-on, the projection of a circular texture element will be a circle and the aspect ratio will be **1**. As the surface is slanted back in depth,

[1] Figure due to Cesare Parise.

away from upright—to 70 degrees, to 50 degrees, and so on--the projections will be successively 'thinner' ellipses, with smaller aspect ratios.)

So, if the aspect ratio of a projection from the circular element was measured accurately, surface slant could be immediately and directly recovered, under the assumption that surface texture elements are circular. Thus, a measured aspect ratio of .707 would correspond to a surface slant of 45 degrees.

But suppose, reasonably, that there is noise in the measuring process, in recording aspect ratios. As a result, on one occasion a surface slanted at 45 degrees might give rise to an aspect ratio 'measurement reading' of .643—'best corresponding' to a slant of 40 degrees by application of **equation 1**. On another occasion that same surface (at 45 degrees) might lead to a measured aspect ratio of .719—'best corresponding' to a surface slant of 46 degrees by **equation 1**. And so on.

So if percipients just go by the measured aspect (possibly in error due to noise), they'll often be a bit off, under repeated viewing of the same, 45 degree, slanted surface. The mean or center of the 'aspect ratio' readings might correspond (by **equation 1**) to a 45 degree slant. But there will be a spread or distribution. The amount of 'bouncing' or variability will depend on how noisy the measuring process is. The greater the spread, the less *precise* or *reliable* the

measurement is said to be (note: *accuracy* is a different matter, determined by whether the mean of the distribution aligns with the actual quantity being measured).

Suppose, though, that one also had a decent 'prior guess or hypothesis' that surfaces *tended* to be slanted at 45 degrees in the surrounding, organism, environment.[2] A basic 'Bayesian' idea is that drawing on such a 'prior/working-assessment' of the distribution of slants in the world can help leaven the possibly detrimental effects of measurement noise; such a prior hypothesis might also counter error in the operative, working model of how the stimulus is generated or measured. See figure **1** for specific illustration. Intuitively, assuming such a prior distribution of slants will 'tug' upshot estimates towards slant values in line with the prior probability distribution.

With that as background, here is the perception-science version of Bayes, applied to the slant perception case:

Equation 2). P(s/d) = (P(d/s) * P(s))/ P(d)

[2] A perennial challenge is to account for the origin of 'prior knowledge' or 'assumptions' engaged in perceptual response (see Ernst and De Luca 2012, for extended discussion). One possibility is that this assumed distribution has been 'programmed into' the organism over evolutionary time, in encounters over eons with typical Earthian natural environments. Another is that the prior distributions, and/or the likelihoods (reflecting 'world models') are learned. The question of whether the prior information is innate or learned—and if learned, under what conditions—can often be approached through experiment (van Dam, Parise, and Ernst, this volume, Ernst 2012). The exact mechanisms of learning the relevant distributions are not well understood (as observed in both van Dam, Parise, and Ernst, this volume, and by Ernst 2012).

Here, think of the '**d**'s' as the measured value of an aspect ratio. So, on a given perceptual encounter there will be a specific value of '**d**'--a measured aspect ratio—plugged into **equation 2)**, in 'inferring' a perceptual estimate of what slant is present.

The '**P(s)**' is the *prior probability distribution*—with the basic meaning just discussed. So, here the '**s**' ranges over slant values, that are assumed to be distributed in the world in a certain way (see **figure 1b**). In our example, surface slants are assumed to be distributed around a mean of 45 degrees.

The '**P(d/s)**' is the *likelihood function* (see **figure 1b)**. This reflects a 'working model' of: i) how the world projects to a viewpoint or sensory surface, and ii) how those projections are measured—especially the noise/variability in the measuring process. In our case, the likelihoods reflect the geometry of projection of a circular texture element, given by **equation 1)**; for a given measured aspect ratio, **d**, the mean of the likelihood function will correspond to the slant given by plugging that measured value into **equation 1)**. In our case, the likelihoods will also reflect the model of noise or variability in the measuring process. This corresponds to the 'spread' or 'variance' of the likelihood function (i.e., how peaked it is, or how spread out).

The '**P(s/d)**' is the *posterior probability distribution* (**figure 1b**). Intuitively[3]: this gives the probability of surface slants, given the measured value, **d**, of (here) an aspect ratio. This is what the organism 'works from' (the idea is).

The most likely slant, given the aspect ratio, is the point where the posterior distribution is maximal. Estimating the slant by choosing this point is called 'maximum a posteriori estimation' (MAP). If the distributions are Gaussian this will correspond to choosing the mean of the posterior. If the prior distribution is flat (and so, uninformative) the estimate of the slant will be based only on the likelihood. This is called 'maximum likelihood estimation' (MLE).

Summarizing: the upshot perceptual estimate of (here) slant will result from 'interpreting' the incoming stimulus in terms of two kinds of 'prior' assumptions about the world. Such 'prior assumptions' are reflected in the prior, and also in the likelihood functions. The latter (likelihood) assumptions may be embodied in a fairly complex model of stimulus/measurement generation, often called a "generative model".

2. Perceptual estimation: multiple sources of information.

So, we'll assume that one source of slant information is the visual slant from texture source, as described above. Suppose, on some occasion, such a visual-texture-based path yields an

[3] Strictly, for particular (real-valued) slants, the probability will be zero. So, if we are careful, we'd talk instead of areas under the 'probability density function' corresponding to the probability that a slant will fall within a certain range.

estimate of slant, '**vs**'. We will assume that the other perceptual route at work derives slant from haptic information, perhaps by gauging wrist flexion or angle. Suppose that, on this occasion, this haptic route leads to an estimate of slant, '**hs**'. One way to arrive at an upshot estimate of slant based on these two estimates, from different sources, would be to take a weighted average:

Equation 3). Slant = w1 *vs + w2*hs.

Here it is assumed that **w1 + w2 = 1** (i.e. the weights are fractions that sum to 1).[4]

Obvious question: how to choose the weights?

The basic idea of 'linear cue integration' is that sources are weighted proportionally less if more noisy/variable (that is, less 'precise' or less 'reliable'). Under some reasonable assumptions, the resulting, upshot estimates of (here) slant will be as precise—the least variable--as possible (cf. Landy, Banks and Knill, 2011, p. 7). That is: if the weights are chosen in this way to reflect variability/precision, averaging will minimize the variability in the upshot slant estimates, over

[4] Note: prior information or assumptions about the distribution of slants can be incorporated into the upshot estimate by adding a further 'estimate', '**pr**' (for 'prior'), with its own weight:

Equation 4). Slant = w1 *vs + w2*hs + w3*pr.

Here, again, it is assumed that the weights sum to 1.

other choices of weights. In this circumscribed sense, the slant estimates arrived in the way described are guaranteed to be 'optimal'.[5]

Empirically, it has often (though not invariably) been found that perceptual estimates that combine information from different sources are indeed 'optimal' in the sense described (cf., Ernst and Banks 2002; see Landy, Banks, and Knill, 2011, for an over-view). Adjustments of weights to align with source noisiness/variability can indeed be rapid, sometimes within a few trials (cf. Seydell, Knill, and Trommershäuser 2010, 2011).

It turns out that this sort of weighted averaging is a special case of a 'Bayesian' model of cue integration, that generalizes **equation 2)** above to the multi-cue case (Landy, Banks and Knill, 2011, pp. 8 – 10).[6] The case of estimation by **3)** would correspond, on the Bayesian formulation, to a form of MLE (see section 1 above), only here there are two likelihood functions, one for each source of information. If a non-flat prior distribution is used, then—on the Bayesian

[5] The upshot estimate is only guaranteed to be *unbiased* if the individual estimates are unbiased. Assuming Gaussian noise you can think of the assumption of lack of bias as the assumption that the distribution of measurements from a source—of slant, say, via texture—will center on the actual, worldly slant value present. Details are debated, but this is likely often an unrealistic assumption. Too much perceiving is biased or inaccurate. For sustained discussion of this 'challenge of bias' within the Bayesian modeling tradition, see Ernst and Di Luca (2011), and section IIIB below. The challenge was independently emphasized outside the Bayesian tradition, by Domini and Caudek (cf. Domini and Caudek 2011).

[6] The basic idea is that the likelihood distributions—one for each source of information—are *multiplied* together, and with the prior. So, with two sources of information,

Equation 5). $P(s|vs,hs) \propto P(vs|s)*P(hs|s)*P(s)$.

Then a MAP estimate of the worldly property, like slant, is arrived at by taking the maximum of the posterior, $P(s|vs,hs)$.

Equation **3)** above is a special case of **5)** in the following way. First, there is no prior in **3)**. So it corresponds to the case where **5)** has a flat prior. The posterior, $P(s|vs,hs)$, in **5)** will in that case simply correspond to multiplying the two likelihoods, and thus to MLE estimation. Again, the maximum of the posterior *distribution* is the most likely slant given the two inputs, vs and hs. This should lead to the same perceived slant as equation **3)**, in the case where the weights in equation **3)** are optimally chosen to minimize variability.

formulation—the idea is that perceivers arrive at an estimate via a MAP estimate by taking the maximum of the upshot, posterior distribution.

3. Perceptual integration and fusion.

A challenge faced by perceptual systems is to determine when and how to combine information. Here we will describe a few of the puzzles involved in meeting such challenges. The first two subsections, **A** and **B**, concern 'decisions' confronting perceptual systems about whether to *combine* or *integrate* information in estimation. Subsection **C** concerns whether or when perceptual systems retain separate estimates from different sources (say, visual and haptic), after integrating those estimates in estimating a worldly property.

We do not here say much in about the details of the computational modeling designed to explain perceptual system behavior in these cases, which can be complex (especially for cases **B** and **C**). Our aim is mainly to isolate different perceptual system challenges. For a fuller summary of the modeling details, see van Dam, Parise, and Ernst (this volume)—who also provide numerous references to relevant recent work on these topics in the modeling literature.

A). In the example of slant perception developed in the preceding sections, the assumption made was that texture elements are circular. As a result, surface slant can be determined by equation **1)** above. But suppose the additional use of stereo to gauge slant yields a widely discrepant estimate of slant, consistent with the unusual (but possible) situation where the texture elements are ellipses and not circles. As a result, the most 'sensible' perceptual system

strategy may well be to veto the slant-from-texture assessment, and instead rely only on the stereo derived estimate of slant. However, the cue combination schemes described thus far would lead to an upshot estimate *between* these two estimates, thereby skewing the upshot estimate in the direction of the divergent texture based estimate.

On a scheme explored by Knill (2003, 2007), veto behavior is modeled using a 'mixture' likelihood. Such a 'mixture likelihood' is derived from a likelihood reflecting the assumption that texture elements are circular and from a likelihood reflecting the assumption that the texture elements are elliptical. Such a mixture likelihood has a long, flat, non-negligible 'tails', that, essentially, reflect the possibility that the texture elements are ellipses. This leads to 'veto-like' behavior when the estimates are widely discrepant (Knill 2003, 2007).[7]

Note that in this case it is granted as somehow known or assumed by the perceptual system that the separate (stereo and texture) estimates pertain to the same (here) slant property. The source of the discrepancy of estimates results from the falsity of the assumption that the texture elements are circular—which 'throws off' the texture-based estimate of slant. The long-tailed 'mixture-likelihood', in effect, allows the estimate that is discrepant, due to the failed assumption, to be vetoed or discarded.

B). Section **A** presents an example in which additional information from a second cue serves to rule out, as discrepant, the estimate resulting from a false assumption about the world. However, more typically a discrepancy between sensory information cannot so easily be 'blamed' on one information source or the other. Cue combination modeling should also

[7] Ernst (2012) contains a detailed discussion of the computational effects of positing such thick-tailed distributions, and of the motivations and justifications for doing so.

explain how perceptual systems make a 'reasonable' determination of whether sensory signals from different sources—say, haptic and stereo-vision—pertain to the same worldly property (slant, size, etc.).

Consider, for instance, observers who are determining object size or width by both gripping the object and determining size visually via stereo. Due to measurement noise, there will inevitably be some discrepancy between the haptic size signal or estimate and the visual-stereo size signal, even when the estimates derive from the same width or expanse. The perceptual-system challenge is to 'decide' whether the estimates are indeed of the same size or expanse, and so should be integrated or combined, in arriving at an estimate of size. This "correspondence problem" (or problem of "causal inference") is an important challenge facing perceptual systems, that has only recently begun to draw the sustained attention of modelers.

The modeling details in, for instance, the work by Ernst and collaborators are somewhat complex (see Ernst and Di Luca 2011; Ernst 2012; for a summaries of the Ernst work, see van Dam, Parise, and Ernst, this volume; for a survey of related work, see Shams and Beierholm, 2010). But the basic idea is that the perceptual system determines the extent to which any discrepancy in (here) haptic and visual estimates of size is likely due to: i) noise: ii) a bias in the measuring of the worldly properties; or iii) a difference in the worldly properties measured. To do this, prior knowledge is needed specifying how likely it is that the combination of the relevant kinds of sensory estimates—haptic and stereo, say--co-occur together, or yield estimates that align with each other. This is captured in what is called the "coupling prior" encoding the assumption of how tightly the two signals are generally linked. An extremely tight

linkage would mean that the system assumes that these sensory sources are always providing information about the same world property; a flat coupling prior means that their co-occurrence is governed by chance. Using this prior the separate sensory estimates are each adjusted according to the estimate obtained from the other source, and the strength of the assumed association. If after this step a discrepancy between the updated estimates is still sensed, this means it is likely that there is either a measurement bias in one or both of the senses or a real difference in the world properties measured with each. In this case, if it is determined as very likely that the visual and the haptic estimates derive from different worldly size properties, then the estimates would not be further integrated or combined.

C). Suppose, once more, that subjects are determining size or width by both looking at an object and gripping it (cf Ernst and Di Luca 2011, Figure 12.3). Ernst and Banks (2002) found that haptic and visual size information is combined, in reaching an upshot estimate of size, in a way that maximizes precision. But it seems quite possible that even though perceivers combine the visual and haptic sensory information in this way, they still retain access to the visual and haptic signals or estimates. If so, there would not be complete perceptual "fusion"—in the sense that the initial visual and haptic estimates can still be accessed and used. This is just what is found for touch and sight via the 'oddity' task used in Hillis et al. (2002), and described in van Dam, Parise, and Ernst (this volume). By contrast, Hillis et al. (2002) found that for purely visual assessment of slant, as gauged by stereo and by texture information, subjects could *not* draw upon the individual (visual) stereo and texture signals or estimates, or at least not to the same extent as when each is presented in isolation. Thus, at least for this sort of purely visual slant

perception, there was perceptual fusion, more or less automatically leading to optimal integration or combination of the estimates.

In modeling when and why fusion occurs Ernst and collaborators make use of the 'coupling prior' distributions noted in **B** above (see, van Dam, Parise, and Ernst, this volume, and Ernst and De Luca 2011). Such a coupling prior reflects assumptions about how likely it is that the different sorts of sensory signals occur together. These might be haptic and stereo-vision signals; or the signals might both be visual, like texture and stereo. In effect, the strength of the association between (say) haptic and the stereo-vision sensory signals is encoded in the width or spread of this, coupling prior, distribution. So, a 'tight' or 'sharp ridged' coupling prior 'says' that haptic size signals and stereo (vision) size signals tend to vary together. With this sort of prior information on hand, it is 'sensible' for perceptual systems to treat any discrepancy between haptic and stereo signals or estimates as due to 'measurement noise', and to be guided, instead, by the (strongly peaked) coupling prior--discarding the individual haptic and stereo estimates.

4. Brief pointers to the philosophy of perception.

Though the Bayesian approach to modeling cue combination is the dominant approach in perception science, the computational modeling details of this work are just starting to find their way into discussions in philosophy of perception (though see Rescorla forthcoming; see also Columbo and Series 2012). In this section we briefly point to places where connections are there to be drawn and explored, with the science potentially stimulating, enriching, and constraining the work in philosophy.

As we have seen, the chief aim in this Bayesian cue combination modeling tradition is to explain how different sources of information are combined in reaching upshot estimates of single worldly properties, like size or like slant. This general kind of perceptual combination is what de Vignemont (this volume) discusses and explores as “integrative binding”.

There are, first of all, apparent connections to Molyneux’s question. Molyneux’s question is explored in the Sinha and in the Van Cleve contributions to this volume. As Van Cleve discusses, formulations of Molyneux’s question can differ subtly and importantly. But Molyneux’s question is essentially the question of ‘whether a person born blind would, upon regaining sight, recognize by vision the shape of an object previously known by touch’. For detailed discussion of the implications of how this question (and related questions) is/are answered, see (again) the Van Cleve and the Sinha contributions to this volume.

On the working outlook of cue combination modeling, the same spatial properties are accessed and estimated through different sources. This is in line with the idea, dating to Aristotle, that there are perceptual ‘common sensibles’—here spatial common sensibles. Such common sensible views have often, traditionally, been associated with a proposed or hypothesized “yes” on Molyneux’s question. This observed, the connection between the cue combination computational modeling and commitments on Molynuex is not straightforward. A full charting would include exploring just how subjects gain the ability to access the same spatial properties across information gained from touch and sight—perhaps through some form of learning. There is much still to be learned about how the relevant kind(s) of perceptual learning is/are

achieved (on the modeling of learning, see van Dam, Parise, and Ernst, this volume, and Ernst 2012).

Here is another kind of connection between the modeling we have described and topics in the philosophy of perception.

A number of contributors to this volume discuss how and when *different* properties—say, slant and color--are bound to the same object or object surface. See especially the contributions to this volume by O'Callaghan, by Deroy, and by Bayne. De Vignemont (this volume) refers to this sort of perceptual-binding achievement as "additive binding".

Explaining this kind of binding of multiple properties to an object/event is not the primary aim of the kind of Bayesian cue combination modeling described in this note. By way of (over-simplified) illustration: in the case of Bayesian cue combination, the modeling task is to understand how to combine information that is at least partly redundant—haptic and visual information about slant, say. By contrast, in the 'multiple property' cases focused on by O'Callaghan, Deroy, and Bayne, the sensory information specifying one property (say slant) and the sensory information specifying the other property (say, color) is not redundant—two different properties are indicated (to be bound to the same object).

Nonetheless, there is an important connection between the Bayesian cue combination case (redundant information) and the 'multiple property' case (non-redundant information). A core challenge faced by perceptual systems, that must be resolved if different properties are to be accurately bound to the same object or surface, is determining whether the information guiding the detection of each property derives from the same object. Once again, see the

contributions to this volume by Deroy, O'Callaghan, and Bayne; see also the de Vignemont contribution. Meeting this challenge requires solving a "correspondence problem" (or a problem of "causal inference") of the sort described in section **IIIB** above, and in the van Dam, Parise, and Ernst (this volume) section on "The correspondence problem".[8]

References

Bayne, T. 2013. The multisensory mature of perceptual consciousness. (This volume.)

Colombo, M. & Series, P. 2012. Bayes on the brain—on Bayesian modeling in neuroscience. *The British Journal for the Philosophy of Science*. 63: 697-723.

Deroy, O. 2013. The unity assumption and the many unities of consciousness. (This volume.)

de Vignemont, F. 2013. Multimodal unity and multimodal binding. (This Volume.)

Domini, F. & Caudek, C. (2011). Combining image signals before three-dimensional reconstruction: The intrinsic constraint model of cue integration. In J. Trommershäuser, K.Kording, M.S.Landy (Eds), *Sensory Cue Integration*. Oxford: O.U.P., 120-143.

Ernst, M.O. 2012. Optimal multisensory integration: Assumptions and limits. In B.E. Stein (Ed), *The New Handbook of Multisensory Processes*. MIT Press, 1084–1124.

Ernst, M.O. & Di Luca, M. (2011). Multisensory perception: From integration to remapping. In J. Trommershäuser, K.Kording, M.S.Landy (Eds), *Sensory Cue Integration*. Oxford: O.U.P. 224-250.

Hillis, J. M., Ernst, M. O., Banks, M. S., & Landy, M. S. 2002. Combining sensory information: Mandatory fusion within, but not between, senses. *Science*, 298: 1627–1630.

[8] Thanks to Cesare Parise for helpful discussion, as well as for our Figure 1.

Knill, D. C. 2003. Mixture models and the probabilistic structure of depth cues. *Vision Research*, *43*: 831–854.

Knill, D. C. 2007. Robust cue integration: A Bayesian model and evidence from cue-conflict studies with stereoscopic and figure cues to slant. *Journal of Vision*, 7:5 1–24.

van Dam, L.C.J., Parise, C.V., & Ernst, M. 2013. Modeling multisensory integration. (This volume.)

Landy, M.S., Banks, M.S., and Knill, D.C. 2011.Ideal observer models of cue integration. In J. Trommershäuser, K.Kording, M.S.Landy (Eds), *Sensory Cue Integration*. Oxford: O.U.P., 5-29.

Rescorla, M. Forthcoming. Bayesian perceptual psychology. In M. Matthen, *The Oxford Handbook of the Philosophy of Perception*. Oxford: O.U.P.

O'Callaghan, C. 2013. Intermodal Binding. (This volume.)

Seydell, A., Knill, D.C., & Trommershäuser, J. 2010. Adapting Bayesian priors for the Integration of visual depth cues. Journal of Vision, 10, 1-27

Seydell, A., Knill, D.C., & Trommershäuser, J. 2011. Priors and learning in cue integration. In J. Trommershäuser, K.Kording, M.S.Landy (Eds), *Sensory Cue Integration*. Oxford: O.U.P., 155-172.

Shams, L., & Beierholm, U. R. 2010. Causal inference in perception. *Trends in Cognitive Sciences*, *14*: 425–432.

Sinha, P. 2013. [report on the Project Prakesh 'Molyneux subjects'.] (This volume.)

van Cleve, J. 2013. Berkeley, Reid, and Sinha on Molyneux's Question. (This volume.)