Experimentally disambiguating models of sensory cue integration

Peter Scarfe

Vision and Haptics Laboratory, School of Psychology and Clinical Language Sciences, University of Reading

Abstract

Sensory cue integration is one of the primary areas in which a normative mathematical framework has been used to (1) define the “optimal” way in which to make decisions based upon ambiguous sensory information and (2) compare these predictions to an organism’s behaviour. The conclusion from such studies is that sensory cues are integrated in a statistically optimal fashion. Problematically, numerous alternative computational frameworks exist by which sensory cues could be integrated, many of which could be described as “optimal” base on different optimising criteria. Existing studies rarely assess the evidence relative to different candidate models, resulting in an inability to conclude that sensory cues are integrated according to the experimenters preferred framework. The aims of the present paper are to summarise and highlight the implicit assumptions rarely acknowledged in testing models of sensory cue integration, as well as to introduce an unbiased and principled method by which to distinguish the probability with which experimental data is consistent with a set of candidate models.

Introduction

Integrating sensory information into a robust percept

Humans have access to a rich array of sensory data from both within and between modalities upon which to based perceptual estimates and motor actions. This sensory data is treated as consisting of quasi-independent sensory “cues”. For example, in the visual modality incoming sensory information is broken into cues such as horizontal and vertical disparity (Howard & Rogers, 2002; Rogers & Bradshaw, 1993), vergence (Tresilian & Mon-Williams, 2000; Tresilian, Mon-Williams, & Kelly, 1999), texture (Knill, 1998b), shading (Bulthoff & Mallot, 1988), perspective (Hershenson, 1999) and blur (Held, Cooper, & Banks, 2012; Vishwanath, 2012). There is no fully agreed definition of what constitutes a “cue” (Glennerster, Tcheang, Gilson, Fitzgibbon, & Parker, 2006; Ho, Landy, & Maloney, 2006; Knill, 1998a, 1998b; Zabulis &

Backus, 2004), however for the present purposes we will adopt the definition provided by Ernst and Bülthoff (2004, p. 163) and treat a cue as "… any sensory information that gives rise to a perceptual estimate".

For an observer making an estimate about a property of the world, S, the perceptual estimate of that property from the ith cue is given by

$$\hat{S}_i = f_i(S)$$

(1)

Here f_i represents the generative function relating the property in the world to the perceptual estimate of that property (Ernst & Banks, 2002). Given a set of cues from within or between modalities the question then becomes how information from cues is *integrated* to generate a robust percept of the world (Ernst & Bulthoff, 2004). Mathematically, there are multiple ways in which this could occur (Jones, 2016; Tassinari & Domini, 2008; Trommershauser, Körding, & Landy, 2011), however currently the most popular theory is that of "modified weak fusion" (MWF) (Landy, Maloney, Johnston, & Young, 1995; Maloney & Landy, 1989). Whilst there are clearly multiple benefits of combining and integrating sensory information (Ernst & Bulthoff, 2004), MWF posits that a key goal (or optimising criteria) of sensory integration is to maximise the precision of the integrated cues sensory estimate. We first describe the key predictions of MWF and how they can be tested experimentally. This is essential to examining the probability with which alternative models can be distinguished from MWF in an experimental setting.

MWF is applicable to the integration of any number of cues, however, here is will be described in terms of two cues, $\hat{S}_A$ and $\hat{S}_B$, which provide redundant information about a property of the world, S. If each cue is corrupted by statistically independent Gaussian noise with variances σ_A^2 and σ_B^2 such that each cue can be represented a Gaussian probability density function, it can be shown, given some additional assumptions, that the integrated cues estimate, $\hat{S}_C$, is given by a simple weighted average (Cochran, 1937; Oruc, Maloney, & Landy, 2003).

$$\hat{S}_C = w_A\hat{S}_A + w_B\hat{S}_B$$

(2)

The cue "weights", $w_A$and w_B, are determined by the relative reliability of each cue, $w_A = r_A/(r_A + r_B)$ and $w_B = r_B/(r_A + r_B)$, where the reliability of the i^{th} cue is defined as the inverse of its variance $r_i = 1/\sigma_i^2$.

$$w_A = \frac{1/\sigma_A^2}{1/\sigma_A^2 + 1/\sigma_B^2} \tag{3}$$

$$w_B = \frac{1/\sigma_B^2}{1/\sigma_A^2 + 1/\sigma_B^2} \tag{4}$$

The weights of the cues sum to unity ($w_A + w_B = 1$) and the variance of the integrated cues estimator is given by

$$\sigma_C^2 = \frac{\sigma_A^2 * \sigma_B^2}{\sigma_A^2 + \sigma_B^2} \tag{5}$$

As such, the standard deviation (sigma) of the Gaussian probability density function representing the integrated cues estimator is given by

$$\sigma_C = \sqrt{\frac{\sigma_A^2 * \sigma_B^2}{\sigma_A^2 + \sigma_B^2}} \tag{6}$$

The key benefit of integrating cues in this way is that the sigma of the integrated cues estimator is *always* less than or equal to the sigma of the most reliable of the individual sensory cues. As a result, integrating cues in this way is often termed optimal cue integration (Trommershauser et al., 2011). The maximal reduction in sigma is achieved when the two cues are equally reliable (Figure 1). That is, setting $\sigma_A = \sigma_B = \sigma$ and assuming $\sigma > 0$, solving Equation 6 gives

$$\sigma_C = \frac{\sigma}{\sqrt{2}} \tag{7}$$

As the reliability of the cues get progressive more unbalanced, the benefits of integrating cues in accordance with Equations 2 and 6 diminishes (Figure 1b). There are additional benefits of combining sensory information, for example, increased robustness through redundancy in perceptual estimates (Ernst & Bulthoff, 2004), however, if an organism is to benefit from integrating cues to the increase the precision of the integrated cue estimate, as proposed by MWF (Landy et al., 1995; Maloney & Landy, 1989), it would do so only when the cues are approximately matched in reliability. For unmatched cue reliabilities, the benefits of Equation 6, compared to simply choosing the most reliable of the two cues can become minimal. This is shown by the narrowing of the shaded region in Figure 1b as the ratio of single cue sigmas diverges from one.

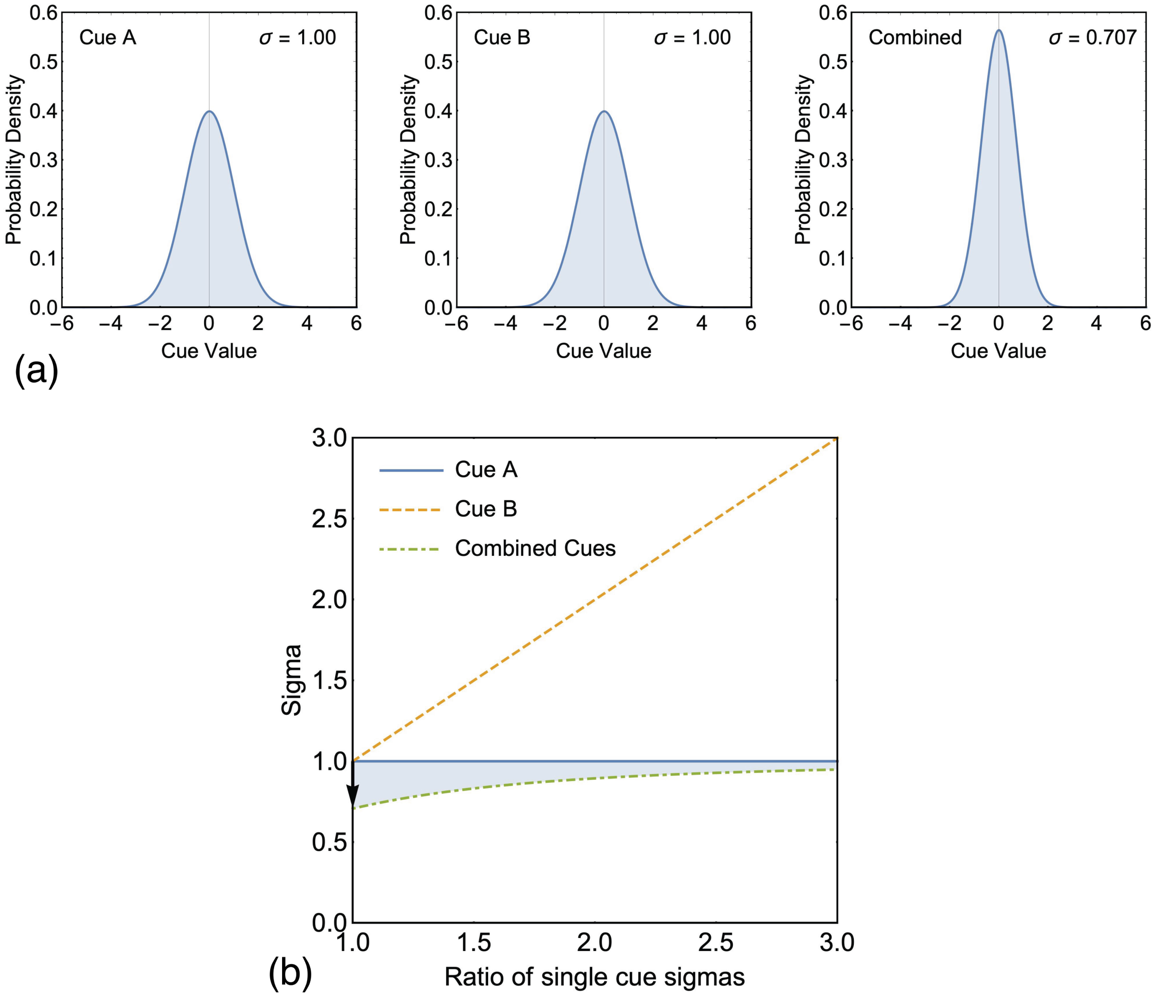

Figure 1: *(a) Shows a hypothetical example of combining two cues (A and B) with identical reliabilities (for both cues $\sigma = 1$). In this instance an observer would maximally benefit from combining cues in accordance with Equation 6 and obtain a $\sqrt{2}$ reduction in sigma (i.e. Equation 7). (b) Plots single cue sigmas and the integrated cues sigma associated with two hypothetical cues for a range of sigma ratios. A sigma ratio of one indicates that the two cues are equally reliable (as in (a)). A value greater than one indicates that Cue B is more variable than Cue A. The shaded region shows the increased precision afforded by integrating cues in accordance with Equation 6. The black arrow shows the maximally achievable increase in precision shown in (a).*

Clearly for Equation 2 and 6 to hold, the units of $\hat{S}_A$ and $\hat{S}_B$ have to be the same. As such MWF has a "cue promotion" stage prior to averaging, where cues are promoted so as to be in common units (Landy et al., 1995). The cue promotion stage, whilst critical, is rarely directly studied (although see Burge, Fowlkes, & Banks, 2010; Burge, Peterson, & Palmer, 2005), so experimenters tend to assume *a priori* that cues estimates are in the same units. As such, MWF is often referred to as "weighted averaging" in reference to Equation 2. A related issue is that the units in which $\hat{S}_A$ and $\hat{S}_B$ are modelled has to be equivalent to the units that the observer is using to make perceptual estimates. Again, this is normally assumed *a priori*, however, it is possible to examine this assumption directly (though is rarely done). For example, Hillis et al. (2004) investigated the perception of slant from texture and disparity cues. Here the estimates need to be in common units of "slant" (e.g. degrees or radians), as such, they conducted a control experiment to ensure that observers were judging disparity defined surfaces in units of "slant", rather than on the basis of the gradient of horizontal disparity. Clearly, the slant units across cues (degrees or radians) would also need to be the same.

Equation 2 has a couple of additional interesting properties. Firstly, it holds regardless of the difference in the estimates provided by each cue ($\hat{S}_A$ and $\hat{S}_B$). This is problematic, as it makes little sense to integrate cues if they provide wildly discrepant estimates of an environmental property. For example, if two cues suggest that the size of an object is 5.5cm and 5.4cm, and a third that the object is 500cm, it would be sensible to ignore the massively discrepant perceptual estimate. To get around this problem the perceptual system is assumed to be *robust*, such that cues are "vetoed" if they provide discrepant estimates (Landy et al., 1995). In the case of only two cues, as examined here, the situation is more complex, as it is unclear which cue to veto. Experimentally, it has been shown that for two cues with a large discrepancy, veto does occur, but puzzlingly, the vetoed cue is not necessarily the less reliable cue (Girshick & Banks, 2009).

Second, cues are integrated regardless of their perceptual bias. By "bias" we mean a difference between (a) the perceptual estimate of a property of the world and (b) the actual physical value of that property. This has been termed "external accuracy" (Burge, Girshick, & Banks, 2010). Bias is a problem, in part, because there is no reason to assume that cues which are most reliable are also least biased. As a result, there are a mathematically definable range of circumstances where integrating cues in accordance with Equations 2 and 6 results in perceptual estimates which are more precise, but *less accurate with respect to the world* (Scarfe & Hibbard, 2011). In these instances, observers integrated cue estimates are in effect "more precisely wrong" ("wrong" meaning inaccurate). Bias is difficult to account for in models of cue integration, as by definition, observers have no direct access to properties of the world (Descartes, 1641) and experimenters have no direct access to an observers internal perceptual estimates. Because of this, cues are generally assumed *a priori* to be unbiased, with any bias typically being attributed to unmodelled cue conflicts or response bias (Watt, Akeley, Ernst, & Banks, 2005).

The assumption of unbiased estimators would be reasonable if our sensory systems got error signals sufficient to calibrate out bias through the process of sensory adaptation (Adams, Banks, & van Ee, 2001; Henriques & Cressman, 2012; McLaughlin & Webster, 1967; Scarfe & Glennerster, 2014; Welch, Bridgeman, Anand, & Browman, 1993). However, this seems unlikely in all instances given the observed large reproducible perceptual biases in real world environments (Bradshaw, Parton, & Glennerster, 2000; Koenderink, van Doorn, Kappers, & Lappin, 2002; Koenderink, van Doorn, Kappers, & Todd, 2002; Koenderink, van Doorn, & Lappin, 2000; Wagner, 1985), as well as in expertly controlled experiments with computer generated stimuli where cue conflicts cannot be evoked to explain away perceptual bias (Watt et al., 2005). Importantly, it has been shown that the *integration* of sensory cues does not lead to the *calibration* of those same cues (Smeets, van den Dobbelsteen, de Grave, van Beers, & Brenner, 2006). As a result, it is now becoming more widely accepted that cues can provide biased sensory estimates and that this needs to be accounted for in models of cue integration (see Ernst & Di Luca, 2011). We discuss the effect of perceptual bias on testing the predictions of MWF below.

Experimentally testing MWF

Numerous studies have investigated sensory cue integration and purport to show that humans combine cues 'optimally' in accordance with MWF (Burge, Girshick, et al., 2010; Ernst, 2006; Ernst & Banks, 2002; Gepshtein, Burge, Ernst, & Banks, 2005; Girshick & Banks, 2009; Glennerster et al., 2006; Helbig & Ernst, 2007; Hillis, Ernst, Banks, & Landy, 2002; Hillis et al., 2004; Johnston, Cumming, & Landy, 1994; Johnston, Cumming, & Parker, 1993; Knill & Saunders, 2003; Lovell, Bloj, & Harris, 2012; Saunders & Chen, 2015; Scarfe & Hibbard, 2011; Svarverud, Gilson, & Glennerster, 2010; Watt et al., 2005). Here, to provide context for what follows, we describe the standard methodology used for testing the predictions of MWF (for a more detailed exposition see the excellent practical tutorial provided by Rohde, van Dam, & Ernst, 2016).

In its simplest form, testing MWF comes down to seeing if the numerical predictions made by Equations 2 and 6 correspond to observers' behaviour. Equation 2 predicts the perceptual estimate observers will make when cues are in conflict, whereas Equation 6 predicts the increased precision of the integrated cues estimate. Of the two predictions, Rohde, van Dam and Ernst (2016, p. 7) describe Equation 6 as the "essential prediction" of optimal cue integration, stating that "… noise reduction is the most important hallmark of optimal integration". They point out that seeing performance line with Equation 2 is "… by itself not sufficient to show that optimal integration occurs" (Rohde et al., 2016, p. 7). One of the reasons for this is that, as we will examine below, *identical* predictions to Equation 2 can be made by alternative models of perceptual processing, including those in which cues are not integrated in any way. As a result, "(i)f one can show only a bias in the results (equation (2)) but not a reduction in noise (equation (6)), one *cannot conclude that optimal integration occurred* ..." (Rohde et al., 2016, p. 10. Note: equation numbers have been changed to correspond to the equivalent equations in the current paper and italics added.).

In order to test the predictions made by Equations 2 and 6 an experimenter needs estimates of the internal parameters σ_A, σ_B, $\hat{S}_A$ and $\hat{S}_B$ (here the hat refers to an estimate on the part of the observer). To estimate the sigma parameters of individual and integrated cues estimates a two-alternative forced choice (2AFC) task is typically adopted. For example, in Ernst and Banks (2002) observers were presented with a ridge defined by vision, haptics, or both, across two intervals. One interval contained the "standard" stimulus, which was a fixed height, and the other the "comparison" stimulus, which was of a variable height. By varying the height of the comparison stimulus one can map out a psychometric function, which is typically fit with a Cumulative Gaussian function (based on the assumption that both cues can be represented by Gaussian probability density functions, so the difference between these distributions will also be Gaussian). The experimental estimate of the sigma of the cue (or integrated cues) is taken to be the sigma of the fitted Cumulative Gaussian divided by $\sqrt{2}$. The division being necessary as there are two stimuli being judged, one per interval (Green & Swets, 1974). By entering the single cue sigma's into Equation 6 one can get a numerical prediction of the integrated cues sigma, which can be compared with that observed experimentally.

To test the predictions of Equation 2 a "perturbation analysis" is typically used whereby cue conflicts are experimentally introduced into the combined cues stimuli and the weight assigned to each cue is estimated by the extent to which each cue determines the integrated cues percept (Young, Landy, & Maloney, 1993). For example, Ernst and Banks (2002) presented observers with visual-haptic stimuli in which the bar height specified by vision and haptics was in conflict. The predicted weights given to vision and haptics were used to examine the extent to which vision or haptics would determine the integrated cues percept and this compared to observer's behaviour.

To demonstrate this idea, following on from the example above, we can add a perturbation of value Δ to the cue $\hat{S}_B$, such that

$$\hat{S}_C = w_A\hat{S}_A + w_B(\hat{S}_B + \Delta) \tag{8}$$

Recognising that $\hat{S}_A = \hat{S}_B$ and $w_A = 1 - w_B$ we get

$$\hat{S}_C = (1 - w_B) * \hat{S}_A + w_B * (\Delta + \hat{S}_A)$$

(9)

Solving for w_B we get

$$w_B = \frac{\hat{S}_C - \hat{S}_A}{\Delta} \tag{10}$$

Finally, recognising that the numerator is the change in percept compared to the single cue estimates ($\hat{S}_C - \hat{S}_A = \Delta\hat{S}$) gives

$$w_B = \frac{\Delta\hat{S}}{\Delta} \tag{11}$$

Thus, cue weighting can be inferred from the ratio of the change in the integrated cues percept, $\Delta\hat{S}$, and the perturbation, Δ, added to the cue, $\hat{S}_A$. Again, the key insight is that the relative reliabilities of the individual cues determines the integrated cues percept. In experiments such as Ernst and Banks (2002), equal and opposite perturbations are added to each cue, but the result is the same; single cue sensitivities (σ_A and σ_B) are used to predict cue weights, which are in turn used to predict the integrated cues percept when cues are in conflict.

As can be seen from Equations 8 through 11, a core assumption of a perturbation analysis is that single cue estimates are not biased relative to the physical value of the stimulus ($\hat{S}_A = \hat{S}_B = S$). If single cues estimates are biased, the *physical* perturbation an experimenter includes in an experiment will not be equivalent to the *perceptual* estimate of this perturbation. Under these circumstances, when using a perturbation analysis, the inferred weights given to cues will be misestimated relative to the true values. Due to the difficulties inherent in measuring perceptual bias experimenters typically assume that perceptual estimates are *by definition* unbiased. Indeed, the 2AFC procedure used to measure single cue functions offers no way in which to measure perceptual bias as the mean of the fitted function will be zero by definition. In the following we detail the consequences of perceptual bias for inferring weights via a perturbation analysis as it serves to reiterate the reason why Equation 6 is the "essential prediction" (Rohde et al., 2016, p. 7) of optimal cue integration.

Let's assume we have two equally weighted cues $\hat{S}_A$ and $\hat{S}_B$ and that each cue is unbiased, signalling the correct value of an environment property, S, such that $\hat{S}_A = \hat{S}_B = S$. Let's also assume that we have done a perfect job of matching cue reliabilities in our experiment such that $w_A = w_B = \frac{1}{2}$. We conduct a perturbation analysis to measure cue weights by adding a perturbation value Δ to the cue $\hat{S}_B$. Now Equation 8 can be written as

$$\hat{S}_C = \frac{\hat{S}_A}{2} + \frac{\hat{S}_B + \Delta}{2} \tag{12}$$

We can ask what level of bias, β, would need to be present in $\hat{S}_A$ to completely eliminate any effect of the cue perturbation Δ, such that $\hat{S}_C = S$.

$$\hat{S}_C = S = \frac{\hat{S}_A + \beta}{2} + \frac{\hat{S}_B + \Delta}{2} \tag{13}$$

Solving Equation 13 for β gives

$$\beta = -\Delta \tag{14}$$

Thus, if $\hat{S}_A$ is biased by an equal and opposite amount to the perturbation added to $\hat{S}_B$, all evidence of optimal cue integration will be eliminated. More generally, this is the case for *any* level of cue weighting. Given

$$\hat{S}_C = w_A\left(\hat{S}_A + \beta\right) + w_B(\hat{S}_B + \Delta) \tag{15}$$

Recognising that $w_A + w_B = 1$, that cue weights are determined cues variances (Equation 3 and 4) and setting $\hat{S}_A = \hat{S}_B = \hat{S}_C$ in Equation 15 and solving for β gives

$$\beta = -\Delta\frac{\sigma_A^2}{\sigma_B^2} \tag{16}$$

Therefore, for *any* relative weighting of cues, all evidence of optimal cue integration as measured by a perturbation analysis can be eliminated if one (or more generally both) of the cues are biased. The bias needed to do this is a function of the perturbation and the relative reliability of the cues. This is equivalent to the idea of there being perceptual metamers i.e. perceptually indistinguishable stimuli which each consist of different values/magnitudes of the constituent cues (Backus, 2002; Hillis et al., 2002). Given that the typical perturbation added in an experiment is small so as not to elicit cue veto, cues only need to be biased by a small amount to significantly interfere with accurately determining cue weights through a perturbation analysis. For example, Rohde et al. (2016) recommend that Δ be 1 to 1.5 (and no larger than 2) *Just Noticeable Differences* (JND), where a JND is given by $\sigma\sqrt{2}$. This illustrates a further reason why "… noise reduction is the most important hallmark of optimal integration" (Rohde et al., 2016, p. 7).

MWF and Bayesian Inference

MWF is itself a special case of a wider theoretical framework which models perception as a process of Bayesian inference; formalised in terms of Bayes Theorem (Equation 17). Studies on cue integration have been widely used to support the idea that the human nervous system performs some kind of Bayesian inference during perceptual processing (Kersten, Mamassian, & Yuille, 2004; Knill & Richards, 1996). Indeed, cue integration has been described as the "… poster child for Bayesian inference in the nervous system" (Beierholm, Shams, Kording, & Ma, 2009, p. 1).

$$p(S|I) = \frac{p(I|S)p(S)}{p(I)}$$

(17)

In the Bayesian framework, the task of the observer defined as estimating the most probable state of the world, S, given the current sensory information, I. The information available to the observer to do this is contained in the posterior probability distribution, $p(S|I)$, which is determined by current and past sensory information. Current sensory information is represented by the likelihood function, $p(I|S)$, which embodies the generative function which transforms the property of the world, S, into sensory information I (see Equation 1). Past sensory information is represented by the prior, $p(S)$, which embodies the prior probability over states of the world, independent of current sensory information. $p(I)$ is the prior probability over the sensory data and is typically considered as a normalising constant allowing $p(S|I)$ to integrate to one (Mamassian, Landy, & Maloney, 2002).

If the prior is uniform or is much broader than the likelihood function (and dropping the normalising constant) Equation 17 simplifies to

$$p(S|I) = p(I|S)$$

(18)

For the case with conditionally independent cues A and B, this can be written as

$$p(S|I_A, I_B) = p(I_A|S)p(I_B|S)$$

(19)

Given the posterior probability distribution over states of the world, the observer has to make a single perceptual estimate, $\hat{S}$. The "best" way to make an estimate is determined by the costs associated with making errors (Kording, 2007). These costs are embodied in the "loss function" for a given task. Common choices include choosing the *maximum* of the posterior probability distribution, which minimizes the delta loss function, or choosing the *mean* of the posterior probability distribution, which minimizes mean squared error loss. When the likelihood functions of the cues are Gaussian, choosing the mean or maximum of the posterior are equivalent. Making estimates in this way is termed maximum likelihood estimation (MLE) (Ernst & Banks, 2002).

On a given trial in an experiment MLE provides a single estimate $\hat{S}$, of the underlying world property S, which is based on the information provided by the two cues, S_A and S_B (i.e. the information provided by the experimenter in the experiment from the two cues, not the observers estimates of these, $\hat{S}_A$ and $\hat{S}_B$). As pointed out by Beierholm, Körding, Shams and Ma (2009), the posterior, $p(S|I_A, I_B)$, is therefore not strictly speaking the experimentally observed distribution of estimates across trials, $p\left(\hat{S}\middle|\hat{S}_A, \hat{S}_B\right)$. However, this too will be Gaussian if the underlying distributions are normally distributed. In this case, performing estimation in this way will be mathematically equivalent to Equations 2 and 6 (Figure 2).

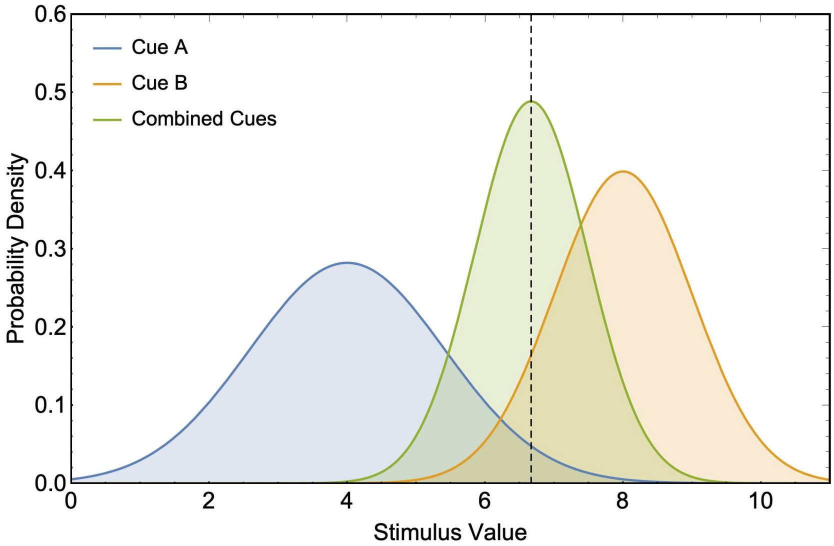

Figure 2: *Shows the equivalence of modelling cue integration as Bayesian Inference and as a weighted average for a two-cue example. The Gaussian PDF for Cue A has a mean of 4 and sigma of 2 and the PDF for Cue B a mean of 8 and sigma of 1. Given that Cue B is more reliable than Cue A, the Integrated Cues PDF is shifted toward that of Cue B. The plotted distributions are calculated analytically by multiplying probability distributions via Equation 19 (with normalisation). As predicted by Equations 2 and 6, the mean of integrated cues PDF (dotted line) is 6.67 and its sigma 0.82.*

Bayesian Inference provides a broader, more widely applicable framework than MWF. As a result it is acknowledge that MWF is likely to provide a local approximation to the reality of how cues are integrated (Landy, Banks, & Knill, 2011). Any model, in any domain, is likely to be only an approximation of the underlying phenomena. One of the primary benefits of MWF is therefore its simplicity, as formulating "full" Bayesian models can be much more complex (Schrater & Kersten, 2000). Indeed, the wider Bayesian framework has been criticised as being so permissive that it is in effect unfalsifiable (Bowers & Davis, 2012). The simplicity of MWF (and its easier falsifiability) arises from the additional assumptions MWF makes about the nature of perceptual processing. Most of the assumptions of MWF result from assuming that perceptual estimates can be represented as statistically independent Gaussian distributions over a linear perceptual scale. Without these assumptions Equations 2 and 6 do not hold and either a fuller Bayesian analysis is needed, or corrections need to be made to the MWF equations, if possible.

A clear case where the Gaussian assumption *does not hold* is in the case circularly distributed variables such as planar direction. With circularly distributed variables the von Mise distribution should be used. A full Bayesian approach could be taken here but simplified equations similar to MWF can be derived with some additional assumptions (Murray & Morgenstern, 2010). However, many studies simply assume that over the stimulus domain tested, Gaussian distributions provide a good enough approximation to the underlying von Mise distributions (Hillis et al., 2004). When statistical independence does not hold, corrections to the weighted averaging equations can again be derived to account for correlated noise (Oruc et al., 2003), without resort to a full Bayesian model. Correlated noise will likely be more problematic for cues with the same modality than across modalities, however, experimentally, regardless of modalities, the correlation between cues is assumed to be zero or so small that MWF provides a valid approximation.

Non-linear perceptual scales and perceptual bias cause a number of problems in modelling cue integration which cannot be easily accounted for. For example, it has been suggested that in the case of the perception of slant from texture the standard 2AFC methodology used misestimates the variance of the underlying estimators (Todd, Christensen, & Guckes, 2010). This misestimate is proposed to arise from systematic biases in observers judgements of slant produced by confounding 2D cues which are unrelated to the perception of slant, but which may be used to discriminate between the images of two textured surfaces (Todd et al., 2010; Todd & Thaler, 2010). Whilst this example is an area of active debate (Saunders & Chen, 2015; Todd, 2015), it is known and acknowledged that perceptual scales are *not* linear (Rohde et al., 2016), as a result, the domain over which cue integration is investigated it typically restricted and over this domain the scale is assumed to be a close approximation to being linear (e.g. Hillis et al., 2004).

Alternative models to MWF

Although MWF is the most widely accepted model of cue integration, there are numerous alternative ways in which sensory cues could be integrated, many of which take into account the reliability of sensory cues (Arnold, Petrie, Murray, & Johnston, 2019; Domini & Caudek, 2009; Jones, 2016; Tassinari & Domini, 2008). Much of the difference between models comes down to the computational architecture of the underlying system and many of these models could be described as “Bayesian” in nature (Beierholm et al., 2009; K. P. Kording et al., 2007; Trommershauser et al., 2011). Therefore, as within any area of science, the question comes down to designing experiments which are able to distinguish between competing models of the underlying phenomena. Problematically, until recently, very few papers compared the predictions of MWF to alternative models of sensory integration in any rigorous fashion (for exceptions see de Winkel, Katliar, Diers, & Bulthoff, 2018; Lovell et al., 2012). This has been recognised as a clear weakness in claims that cues are integrated “optimally” in accordance with MWF (Arnold et al., 2019)

An additional problem is that readers are often required to judge the fit of the data to MWF “by eye”, without any accompanying statistics detailing the fit of the model to the data (e.g. Ernst & Banks, 2002; Hillis et al., 2004). Indeed, a recent review has suggested that the adherence to optimal cue integration (without comparison to alternative models) can be assessed visually and has provided a visual taxonomy of “optimal”, “sub-optimal”, “ambiguous”, “near optimal” and “supra-optimal” performance (Rohde et al., 2016, p. 23). This visual taxonomy, of judging the fit to the predictions of optimal cue integration based upon visual inspection of (1) the data, (2) the error bars around the data, and (3) the predictions of optimal cue integration, has started to be used by researchers to assess “optimality” of experimental data through adherence to MWF (Negen, Wen, Thaler, & Nardini, 2018).

A visual taxonomy is problematic for a number of reasons. Firstly, across a number of disciplines, including psychology, behavioural neuroscience and medicine, leading researchers have been shown to have fundamental and severe misconceptions about how error bars relate to statistical significance and how they can be used to support statistical inferences from data (Belia, Fidler, Williams, & Cumming, 2005; Cumming, Fidler, & Vaux, 2007). Second, as will be seen, alternative models of cue integration provide highly correlated predictions with one another. Therefore, “eyeballing” the fit to a single model based on visual inspection is likely to lead to fundamental mistakes in inferring the extent to which a given model fits the data. Finally, as we will demonstrate, there are computational techniques which can be easily used to assess the fit of a set of candidate models to data in a far more objective way.

Outline of the current study

Here we present a technique which can be used to determine the probability with which a given experiment on sensory integration can experimentally distinguish between alternative models of the underlying phenomena. This technique consists of simulating end-to-end experiments (behaviour of observers in an experiment, fitting of psychometric functions, estimation of parameters from data, and final statistical analysis) in which observers act in accordance with a candidate model (or models), and examining the probability with which the data can be distinguished from the predictions made by a range of alternative models. Given the ubiquity of MWF, we focus primarily on the extent to which the predictions of MWF can be distinguished from two popular alternative models, however, our methods are fully generalisable and can be used to compare any set of models. As we will see, due to alternative models providing highly correlated predictions with MWF, it can become very difficult to experimentally distinguish between candidate models of sensory cue integration.

Methods and Results

Alternative models compared to MWF

For the simulations presented, the predictions of MWF were compared to two alternative models (a) choose the cue with the minimum sigma (MS), and (b) probabilistic cue switching (PCS). There are clearly numerous other models which could have been chosen for comparison (Jones, 2016), however, these two models have the benefits of (1) being conceptually similar to MWF, (2) require experimental estimation of the same parameters, and (3) are reducible to comparably simple equations, with no requirement to model the combination of full probability density functions. They also have the benefit of having been compared to the predictions of MWF in previous papers.

MS cue is conceptually very similar to MWF, in that, when the reliabilities of the two cues are imbalanced the mean and sigma of the integrated cues estimator in MWF is determined nearly entirely by the more reliable of the two cues. When one cue dominates the integrated cues percept it has been termed ‘cue capture’ (Ernst & Banks, 2002). The similarities of the predictions to MWF can be seen in Figure 3 where we re-plot discrimination thresholds for the visual, haptic and integrated cue estimators from Ernst and Banks (2002) (see Appendix A). As can be seen, for the 0, 67 and 200% noise conditions the threshold for the integrated cues estimator is visually indistinguishable from the discrimination thresholds of the most reliable of the individual cues (visual or haptic). Therefore, the *only* condition in this paper which can test the MWF model, relative to choosing the cue with the minimum sigma, is the 133% noise condition where the reliabilities of the two cues are nearly identical (grey rectangle). Be aware that this statement is statement based upon a *visual judgement from a graph of the data*. But, it is one which will be backed up computationally below.

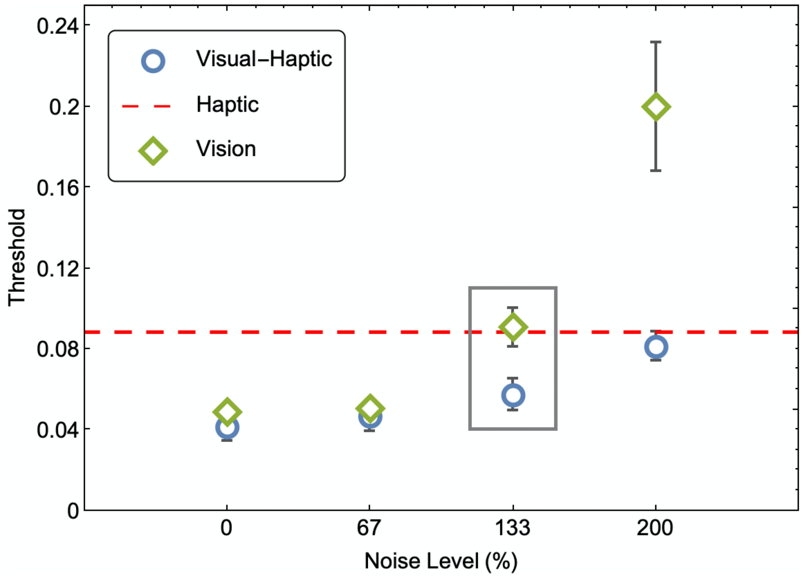

Figure 3: *Replot of the threshold data from Ernst and Banks (2002) Figure 3d. The threshold is defined as the difference between the 84% and 50% point of the underlying psychometric function. Thus, smaller thresholds represent more precise perceptual estimates. Thresholds are plotted against the % of noise in the visual modality stimulus (see Ernst & Banks, 2002 for full details). The only datapoint which can distinguish MWF from simply choosing the most reliable of the two cues is the 133% noise level stimulus (grey rectangle).*

When choosing the cue with the minimum sigma the sigma of the integrated cues estimator is given by

$$\sigma_C = \sqrt{min\{\sigma_A^2, \sigma_B^2\}} \tag{20}$$

The mean of the integrated cues estimator is simply that of whichever cue is most reliable (either $\hat{S}_A$ or $\hat{S}_B$).

Probabilistic cue switching (PCS) (Byrne & Henriques, 2013; de Winkel et al., 2018; Nardini, Jones, Bedford, & Braddick, 2008; Serwe, Drewing, & Trommershauser, 2009) proposes that observers do not integrate cues to form a single perceptual estimate, rather, they use a single cue at a given time, and switch between cues with the probabilities p_A and p_B (where $p_A + p_B = 1$). The mean and sigma of the integrated cues estimator is given by

$$\hat{S}_C = p_A\hat{S}_A * p_B\hat{S}_B \tag{21}$$

and

$$\sigma_C = \sqrt{p_A\left(\hat{S}_A^2 + \sigma_A^2\right) + p_B\left(\hat{S}_B^2 + \sigma_B^2\right) - \left(p_A\hat{S}_A * p_B\hat{S}_B\right)^2} \tag{22}$$

The probabilities p_A and p_B are determined by the relative reliabilities of each cue, such that $p_A = w_A$ and $p_B = w_B$. Substituting Equations (3) and (4) into (22) and simplifying gives

$$\sigma_C = \sqrt{\frac{\sigma_A^2\sigma_B^2\left(\hat{S}_A^2 - 2 * \hat{S}_A\hat{S}_B + \hat{S}_B^2 + 2 * (\sigma_A^2 + \sigma_B^2)\right)}{(\sigma_A^2 + \sigma_B^2)^2}} \tag{23}$$

When $\hat{S}_A = \hat{S}_B$, Equation 23 simplifies further to

$$\sigma_C = \sqrt{2}\sqrt{\frac{\sigma_A^2 * \sigma_B^2}{\sigma_A^2 + \sigma_B^2}} \tag{24}$$

The similarities between Equations (2) and (21), and Equations (6) and (24) are clear. Note in particular that Equations (2) and (21) provide *identical* predictions for the mean of the integrated cues estimator. In other words, for the mean of the integrated cues estimator, a model in which cues are *not integrated* and instead used *completely independently* can produce identical predictions to MWF. This is one of the core reasons that noise reduction is essential hallmark of optimal cue integration (Rohde et al., 2016).

Correlated predictions of alternative models

We have already seen how MWF and PCS can provide identical predictions regarding the mean of the integrated cues estimator. Here we examine similarities in the predictions of MWF, MS and PCS in terms of the sensitivity of the integrated cues estimator. Figure 4 plots the predictions for the sigma of the integrated cues estimator under our three candidate models; MWF, MS and PCS. Here, for PCS, the two cues have been set to have the same mean (i.e. Equation 24). As can be seen, the three models provide highly correlated predictions regarding the predicted sigma of the integrated cues estimator.

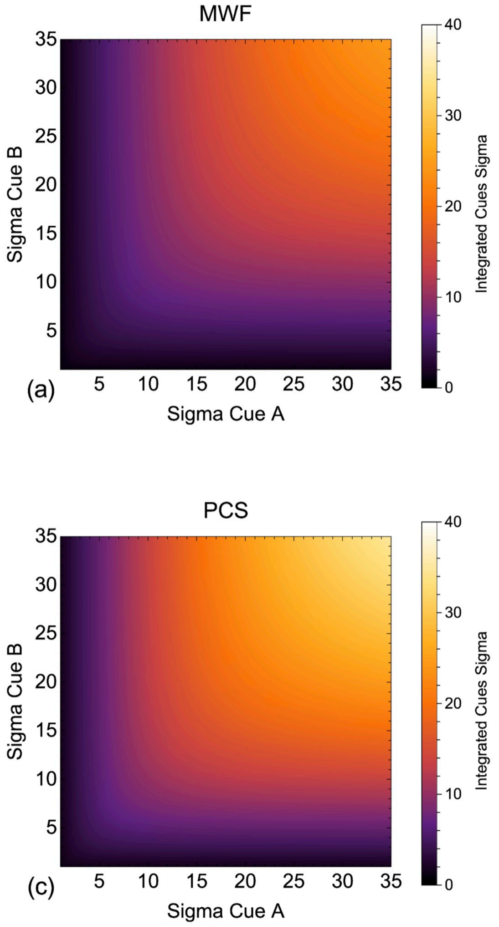

Figure 4: *Shows the integrated cues sigma, calculated for a range of two-cue sigma values under our three models of cue integration, (a) optimal cue integration (MWF; Equation 6), (b) choose the cue with the minimum sigma (MS; Equation 20) and (c) probabilistic cue switching (PCS; Equation 24).*

We can take the difference between the predictions of each model to examine the areas of the parameter space in which MWF can be most easily disambiguated from these two alternative models (Figure 5). Doing this shows that MS and PCS both provide maximally different predictions from MWF regarding the integrated cues reliability, when the sigmas of the individual cues are identical (positive diagonal). The absolute magnitude of the difference between predictions also increases with the sigma of the two cues (compare the bottom left to top right in each plot). Also plotted in Figure 5 are data points from two of the most widely cited papers on optimal cue integration, Ernst and Banks (2002) and Hillis, Watt, Landy and Banks (2004). Whilst some of these datapoints lay near the positive diagonal, many datapoints fall into areas of the parameter space which poorly distinguished MWF from MS and PCS based upon the core prediction of increased sensory precision.

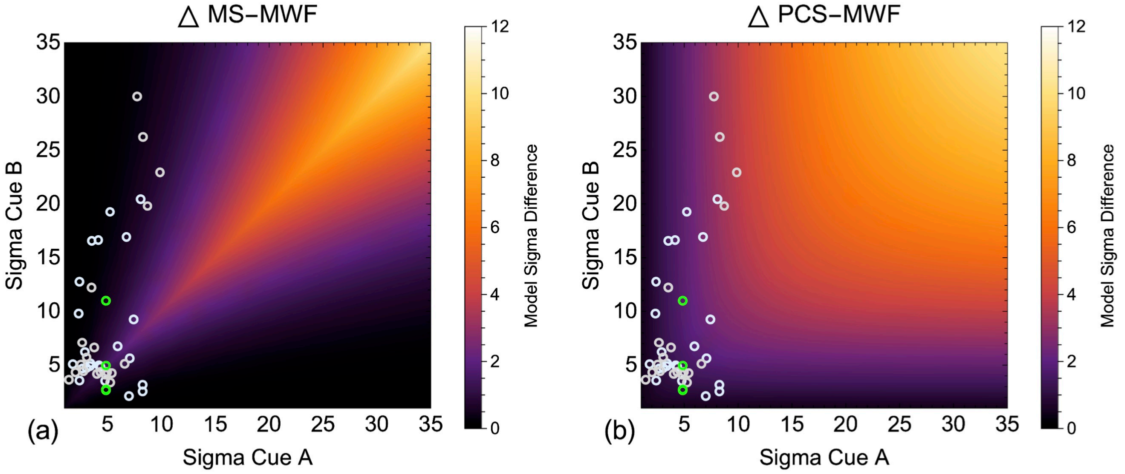

Figure 5: *Plots the difference in integrated cues sigma predictions calculated under (a) MS versus MWF and (b) PCS versus MWF. Green symbols show the cue sigma values from Figure 3d in Ernst and Banks (2002) for the perception of object height. Cyan and grey symbols show cue sigma values from Figure 11 in Hillis, Watt, Landy and Banks (2004) for the perception of surface slant (cyan symbols observer JMH and grey symbols observer ACD). If data are to best disambiguate models they should fall along the positive diagonal in each plot.*

General methods

All simulations were carried out in MATLAB (2020a) (MathWorks, Natick, MA, USA) on an 8-Core Intel Core i9 processor in a MacBook Pro running macOS 10.15. The extensive range of simulations reported were computationally expensive, so where possible they were distributed over the computers CPU processing cores using MATLAB's Parallel Processing Toolbox. The Palamedes toolbox was used to parametrically simulate observers and fit psychometric functions (Kingdom & Prins, 2010, 2016; Prins & Kingdom, 2009, 2018).

Simulation Set 1: Effects of relative reliability and number of observers in an experiment on distinguishing between candidate models

Methods

Observers were assumed to have access to two cues ($\hat{S}_A$ and $\hat{S}_B$) from which to make an integrated cues perceptual estimate ($\hat{S}_C$) about a property of the world. The mean of the two cues prior to any perturbation was the same (55mm as in Ernst and Banks (2002)). Cue A always had the same sigma $\sigma_A = 4.86$, which is approximately that of the haptic cue in Ernst and Banks (2002). Cue B had a sigma given by $\sigma_B = \sigma_A r$ where r varied between 1 and 4 in 27 linearly spaced steps. It has been suggested that to test for optimal cue integration the sigma ratio should be lay within the range 0.5 to 2 (Rohde et al., 2016, p. 15), however, it is clear that experimenters go beyond this reliability ratio (Figure 5). Therefore, we included simulated experiments beyond this recommended range to be more consistent with the existing experimental literature. For each reliability ratio we simulated experiments where there were 4 through 30 (in steps of 1) participants. Cue integration experiments are typically very time consuming, so there are normally few observers per experiment, but a substantial amount of data collected per observer (Rohde et al., 2016). For example, Ernst and Banks (2002) and Hillis, Watt, Landy and Banks (2004), each used four observers. Our highest observer number per experiment therefore represents an upper limit to the observers one might reasonably expect to see in a cue integration study.

The procedure described was repeated for three levels of cue conflict and four data collection regimes. The simulated conflicts, Δ, were 0, 3 and 6mm (as in Ernst and Banks (2002)). Conflicts were added by perturbing each cue by opposite amounts equal to half of the total cue conflict (i.e. 0, ±1.5 and ±3mm), that is $S_A = 55 + \Delta/2$ and $S_B = 55 - \Delta/2$. Estimated from the data of Ernst and Banks (2002), the (above zero) conflicts represented approximately 0.8 and 0.4 JNDs, which is around the recommended magnitude of cue conflict to use in a perturbation analysis (Rohde et al., 2016). In Ernst and Banks (2002) there were conditions with equal and opposite cue conflicts applied (i.e. ±3mm and ±6mm *total* cue conflict) in order avoid perceptual adaptation. With real observers this is needed as if one cue always received a negative perturbation and the other cue always received a positive perturbation, over time the brain may recalibrate the cues (Burge, Girshick, et al., 2010). We did not replicate this here as our simulated observers have no mechanisms of adaptation and all of their responses are statistically independent of one another.

We simulated performance and estimated three psychometric functions for each observer in each experiment. Two single cue functions, corresponding to the stage at which an experimenter estimates singles cue sensitivities, and an integrated cues condition where observers behaved in accordance with MWF. Observers were simulated in accordance with a Cumulative Gaussian function consistent with the underlying mean and sigma of the Gaussian probability density function representing the internal estimator. Functions were sampled with the method of constant stimuli, under four data collection regimes. The method of constant stimuli was selected as this is the most widely used procedure for estimating a psychometric function. Rohde, van Dam and Ernst describe it as "… the simplest and least biased method to measure a complete psychometric function". (p.15).

In sampling a psychometric function using the method of constant stimuli an experimenter has to make four decisions, (1) the range of stimuli to be presented, (2) which stimulus value this range should be centred upon, (3) how many points to sample across the stimulus space, and (4) how many times to sample each stimulus. There are no deterministic rules for deciding any of these, though clearly, the more thoroughly one samples the whole psychometric function, the better the experimental estimates of the underlying parameters. In terms of coverage, it is widely excepted that to gain reliable estimates of the parameters of a psychometric function one should include sampled points where the observer can clearly discriminate the stimuli (Wichmann & Hill, 2001a, 2001b), however, "(t)here is no need to use many, finely spaced stimulus levels. Concentrating responses at just a few appropriately distributed stimulus levels should suffice to obtain reliably estimates of the parameters of a PF" (Kingdom & Prins, 2016, p. 57 (PF being short for Psychometric Function)).

Rohde et al. (2016) in discussing the various factors which feed into this decision process conclude that "(i)n most cases a fixed set of seven or nine comparison stimuli can be identified that suits most observers" (p. 14). Here we adopt the upper of these suggestions. The range of the sampling space, 20mm, was based upon that of Ernst and Banks (2002). The sampling space was always centred upon the true mean of the psychometric function. For single cue functions the mean was $\hat{S}_i = 55\text{mm} \pm \Delta/2$, and for the integrated (MWF) cue function the mean was given by Equation 2, where $\hat{S}_A = \hat{S}_A + \Delta/2$ and $\hat{S}_B = \hat{S}_B - \Delta/2$. Centring the function on the true mean represents a best-case scenario for estimating the (normally unknown) function parameters. The 9 stimulus values used were linearly spaced across the range.

In terms of numbers of trials per stimulus level, for cue integration experiments, Rohde, van Dam and Ernst (2016) suggest that where the mean and slope of the PSE need to be estimated around 150 trials should be used. In contrast, Kingdom and Prins (2010) suggest that "although there is no hard-and-fast rule as to the minimum number of trials necessary, 400 trials is a reasonable number to aim for when one wants to estimate both the threshold and slope of the PF" (p. 57. PF, being Psychometric Function). In a simulation study, Wichmann and Hill (2001a) found that for some of their simulated sampling schemes 120 samples in total per function was often "… too small a number of trials to be able to obtain reliable estimates of thresholds and slopes …" (p. 1302). Therefore, here, in separate simulations, we examined sampling with 10, 25, 40 and 55 trials per stimulus level. This gave us 90, 225, 360, and 495 trials in total per function, which encompassed the above recommendations.

Piloting showed that for the cue reliability range, cue perturbations, and sampling regimes used throughout the present study, these parameters resulted in well fit psychometric functions (see Appendix B and also the criteria adopted for rejected functions detailed below). Whilst not as widely used, we could have used an adaptive method by which to sample the psychometric function (Leek, 2001). We opted not to do so for a number of reasons. First, to be consistent with the most widely used psychophysical methods used in the literature (Rohde et al., 2016). Second, to avoid issues in justifying *which* of the many adaptive methods to use, for example: Psi (Kontsevich & Tyler, 1999), Psi-Marginal (Prins, 2013), QUEST (Watson & Pelli, 1983), QUEST+ (Watson, 2017), the "best PEST" (Pentland, 1980) or a staircase procedure (Kingdom & Prins, 2016). Third, with all adaptive methods, a "stopping criteria" needs to be defined (e.g. number of reversals in a staircase procedure), as well as the allowable steps between stimulus values, both of which need to be justified.

Fourth, if using a staircase procedure, different staircases rules converge at different points on the psychometric function, so the question becomes which rule (or rules) to use (Kingdom & Prins, 2016). Finally, for a number of adaptive methods there are issues related to getting "stuck" in specific areas of the parameter space, causing the algorithm to repeatedly sample high intensity stimulus values (see Prins, 2013 for an extended discussion). Therefore, overall there seemed far more open questions to address if an adaptive procedure were adopted. An additional reason why adaptive procedures are used is because the experimenter does not know the parameters of the underlying function(s), which was not the case here with our simulated observers. With the above permutations, for the first set of simulations, we simulated 27 (reliability ratios) x 27 (number of observers / experiment) x 4 (data collection regimes) x 3 (cue conflicts) x 100 (repetitions of experiment) = 874800 simulated experiments. In total these experiments contained 14871600 simulated observers.

Simulating observers and fitting functions

Simulated data were fit with Cumulative Gaussian functions by maximum likelihood using the Palamedes toolbox. Whilst other fitting methods could be used, for example, fitting based on a Bayesian criterion (Kingdom & Prins, 2010; Kuss, Jakel, & Wichmann, 2005; Schutt, Harmeling, Macke, & Wichmann, 2016), fitting by maximum likelihood was chosen as it is currently the most widely used technique in the literature (Kingdom & Prins, 2010; Wichmann & Hill, 2001a, 2001b). In order to appropriately fit a function to an observers data three assumptions need to be met, (1) the observer does not improve or degrade at the task they are performing over time, (2) each perceptual judgement an observer makes is statistically independent of all others, and (3) performance of the observer can be well characterised by the psychometric function that the experimenter is fitting to the data (Kingdom & Prins, 2010, 2016). As we are parametrically simulating observers, we know that all these assumptions are met.

This is clearly not the case in an experimental setting. Here there is clear evidence that the decisions made by an observer on given trial can be influenced by previous decisions the observer has made (Fischer & Whitney, 2014; Frund, Haenel, & Wichmann, 2011; Kiyonaga, Scimeca, Bliss, & Whitney, 2017; Lages & Jaworska, 2012; Liberman, Fischer, & Whitney, 2014; Liberman, Manassi, & Whitney, 2018; Liberman, Zhang, & Whitney, 2016; Xia, Leib, & Whitney, 2016). Techniques exist to account for this potential "non-stationarity" in observers behaviour during fitting of a psychometric function (Frund et al., 2011; Schutt et al., 2016), but currently these methods have not been widely adopted. In terms of fitting the *correct* psychometric function this largely comes down to the experimenters current understanding of the computational mechanisms underlying behaviour. Techniques exist to fit smooth functions to data without assuming a parametric model (Zychaluk & Foster, 2009), but these are not widely used and often much of the power of applying a parametric model comes from an understanding of the model in relation to the inferred computational architecture of sensory processing.

For all simulations we modelled observers as making zero lapses, so during fitting functions we fixed the lapse rate to be zero. This sidesteps problems related to the extent to which fitting with a (1) zero, (2) fixed non-zero, or (3) variable (but bounded) non-zero lapse rate, effects inferences about the mean and sigma of the "true" psychometric function (Prins, 2012; Wichmann & Hill, 2001a, 2001b). This again comes down to the fact that we as experimenters have no direct access to the underlying computational mechanisms that produce behaviour. Therefore, the decisions made regarding simulating observers represent a best-case scenario under which we can estimate the underlying psychometric function parameters and therefore distinguish between candidate models of the data. The simulation data presented therefore likely represents an *overestimate* of an experimenter's ability to distinguish between alternative models of cue integration.

The mean and standard deviation of the fitted functions were taken as the experimental estimates of the observers' true internal parameters. In cases where a function could not be fit due to the simulated data being (1) at / around chance performance across all stimulus levels, or (2) a step function, the data for that simulated observer were removed from the analysis (see also Appendix B). Overall this represented 0.047% of the data. Poorly fit functions were most prevalent when sampling each stimulus level 10 times, giving 90 trials per psychometric function (removed observers for each number of "trials per psychometric function": 90 trials / function = 0.183%, 225 trials / function = 0.0025%, 360 trials / function = 0.00006%, and 495 trials / function = 0%). Thus, all other things being equal, 150 trials per function (Rohde et al., 2016) would give somewhere between 0.0025% and 0.183% cases where an observers data could not be fit, and 400 trials per function (Kingdom & Prins, 2010, 2016) somewhere between 0% and 0.00006% cases. Clearly, the more data one has, the better one is able to estimate experimental parameters. An alternative analysis where poorly fit functions are *replaced* by a newly simulated observer (rather than removed) results in identical conclusions being made throughout the paper.

Comparing the data to alternative models

For each simulated observer, the estimated mean and sigma of the single cue function with the lowest sigma was taken as the experimental prediction for the MS alternative model. For PCS, the estimated mean and sigma of the single cue functions were entered into Equations 21 and 23 to provide predictions of the mean and sigma of the integrated cues function. In a real experiment, functions are measured in a two-interval forced choice experiment so the sigma of the fitting functions need to be divided by $\sqrt{2}$ as there are two stimuli / intervals in the experiment (Green & Swets, 1974). As the functions were parametrically simulated here, this step was not needed. This procedure allowed us to compare the data from a population of *n* "optimal" observers behaving in accordance with MWF with the experimentally derived predictions of our two alternative models. We could have compared to the alternative models using the true underlying parameter values and predictions of each model, however *by definition* an experimenter only has access to experimentally derived estimates of these internal parameters, not their true values

To statistically test the behaviour of our simulated observers to the predictions of each model the comparisons can be posed in different ways. Here, for each experiment, we entered the data from our simulated observers into one-sample within-subjects t-test to compare the data to the point predictions of the two alternative models. For a given experiment, the mean value of the alternative model predictions across observers was taken as the point prediction for each model. Obviously, in an experimental setting, unless one specifically matches the relative reliability of cues across observers, an experiment is highly unlikely to have a set of optimal observers with the same single cue reliability ratios. However, the aim of these first simulations was to clearly show the effects of the (1) single cue reliability ratio and (2) number of observers in an experiment, on an experimenters' ability to distinguish the data of a set to optimal observers behaving in accordance with MWF from the two alternative candidate models, in a best case scenario where cue reliabilities are perfectly matched. In our second set of simulations (see below) we examine the case where we have a heterogeneous population of observers with different cue reliability ratios. This is a weaker way in which to test the MWF model, but one which is more representative of a cue typical integration experiment.

With statistical tests such as the one-sample t-test used here, the experimenter is making assumptions about the structure of the underlying data. For example, for a one-sample t-test, the data should be normally distributed, measured on an interval or ratio scale and all observations should be independent of one another. The latter two assumptions we know to be met. In terms of normality, for those cue integration experiments which present statistical comparisons, we have never seen the results of normality tests presented. This might be due to the fact that given the very small number of observers in a typical cue integration experiment (e.g. four observers (Ernst & Banks, 2002; Hillis et al., 2004)) it would be difficult or impossible to reliably estimate the normality of the data. Thus, adopting parametric tests rather than non-parametric tests was considered a reasonable choice. Using a non-parametric Wilcoxon signed rank test results in the same conclusions being made throughout the paper but with a *decreased* ability to distinguish between alternative models due to the reduced power of nonparametric, compared to parametric, tests.

Group analysis: integrated cues sensitivity

First, we examine the extent to which MWF, MS and PCS can be distinguished on the basis of the sensitivity of the integrated cues estimator, $\widehat{\sigma_C}$. Figures 6 shows the results of running a one-sample t-test on the sigmas of the Cumulative Gaussian functions fit to our simulated MWF observers in each experiment to see if the data statistically differed from the sigma value predicted by MS. The shading of each pixel in the 27 by 27 grid represents the percentage of our 100 simulated experiments in which the results of a population of observers behaving in accordance with MWF could be statistically distinguished from the numerical predictions of MS (Figure 6).

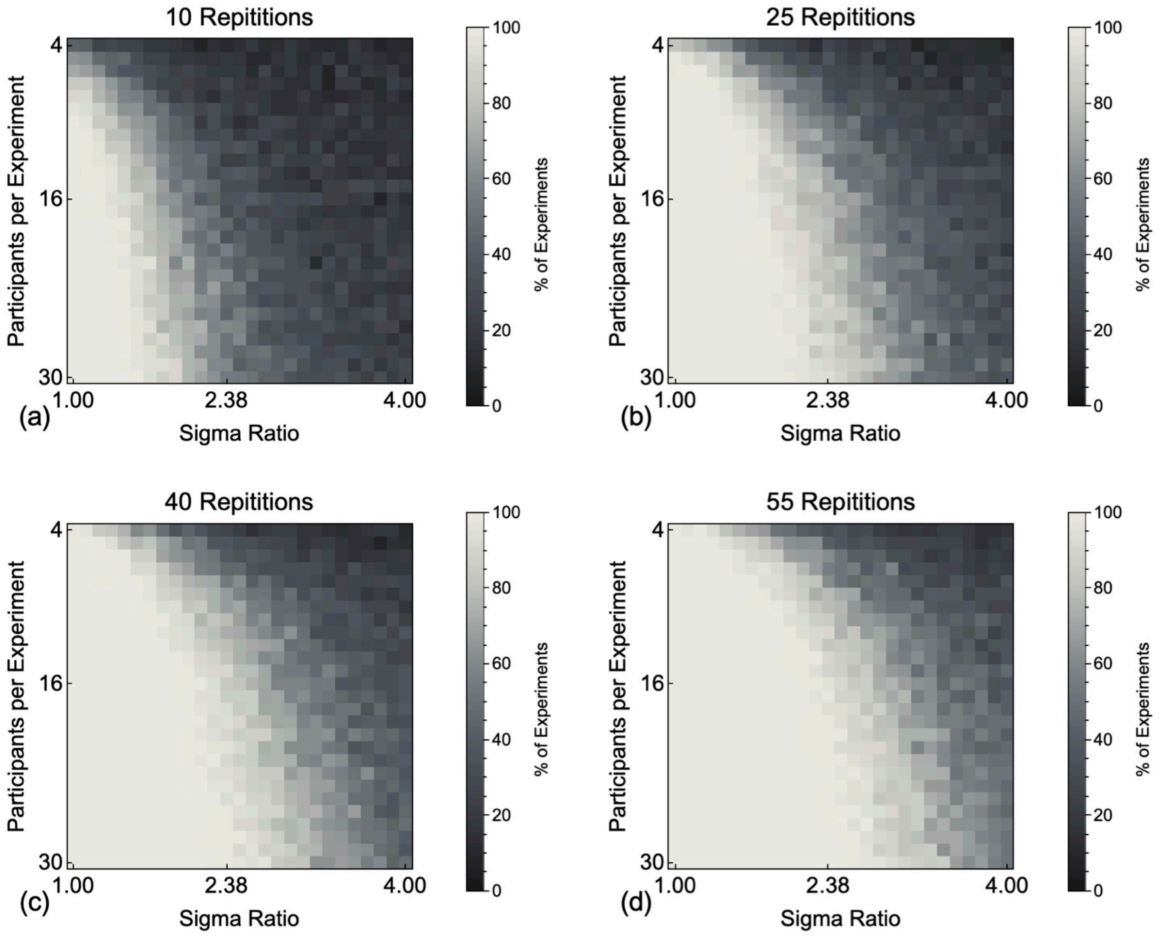

Figure 6: *Shows the percentage of experiments in which the sigmas of the Cumulative Gaussian functions fit to our simulated population of MWF observers could be statistically distinguished from the experimentally derived prediction of choosing the cue with the minimum sigma (MS). Each pixel in the image shows this percentage as calculated across 100 simulated experiments, for a given sigma ratio and number of participants. The four panes show this for (a) 10, (b) 25, (c) 40 and (d) 55, simulated trials per stimulus level on the psychometric function.*

Consistent with the correlated predictions of candidate models (Figure 5), as the sigma of the individual cues becomes unbalanced, it becomes more and more difficult to experimentally distinguish between MWF and MS. This is especially apparent with the low number of observers that characterise typical cue integration experiments. As would be expected, when more data is collected per function models can be more easily distinguished. In Figure 7 we plot the these results for just the simulated experiments with four observers (i.e. the same number of observers in Ernst and Banks (2002) and Hillis et al. (2004)). The vertical grey line shows the maximum recommended sigma ratio to use in cue integration experiments (Rohde et al., 2016), whereas the dashed grey line shows the point at which there is a 50% chance of distinguishing models in a given experiment.

In an ideal world an experimenter would run an experiment and have a 100% chance of distinguishing between candidate models of the underlying phenomena. It is not clear how much lower this probability could go before being considered unacceptably low by an experimenter. However, Figure 7 highlights that with a representative number of observers in a typical cue integration experiment, to have any reasonable chance of distinguishing MWF and MS, one needs to (1) collect a large amount of data per participant and (2) very closely match the reliabilities of the individual cues. Collecting 150 trials per function across four observers with a sigma ratio of 2 would result in an approximately 25% chance of distinguishing these models based in the “essential prediction” of MWF. Thus, the experimental conditions suggested by Rohde et al. (2016) may need to be improved upon to stand any reasonable chance of distinguishing models.

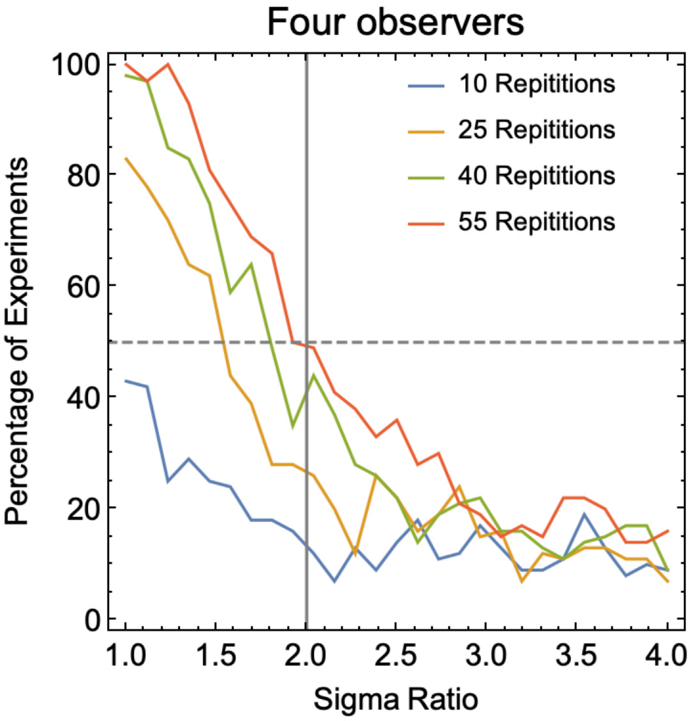

Figure 7: *Plots the percentage of experiments in which the sigmas of the Cumulative Gaussian functions fit to a simulated population of four MWF observers could be statistically distinguished from the experimentally derived prediction of choosing the cue with the*

minimum sigma (MS). The dashed grey line represents the point at which there is a 50% chance of distinguishing the data from the predictions of MS. The vertical grey line shows the maximum recommended sigma ratio to use in cue integration experiments (Rohde et al., 2016).

Figures 8 shows plots the results in the same format as Figure 6, but this time for distinguishing our simulated MWF observers from the predictions of PCS. As would be expected from comparing Equations 6 and 24, the sigmas of the Cumulative Gaussian functions fit to our simulated MWF observers can be easily distinguished from the sigma value predicted by PCS. This is true cross all sigma ratios and data collection regimes. Even in simulated experiments with only four observers (Figure 9) the models can be well disambiguated with all, but the most minimal data collection regime.

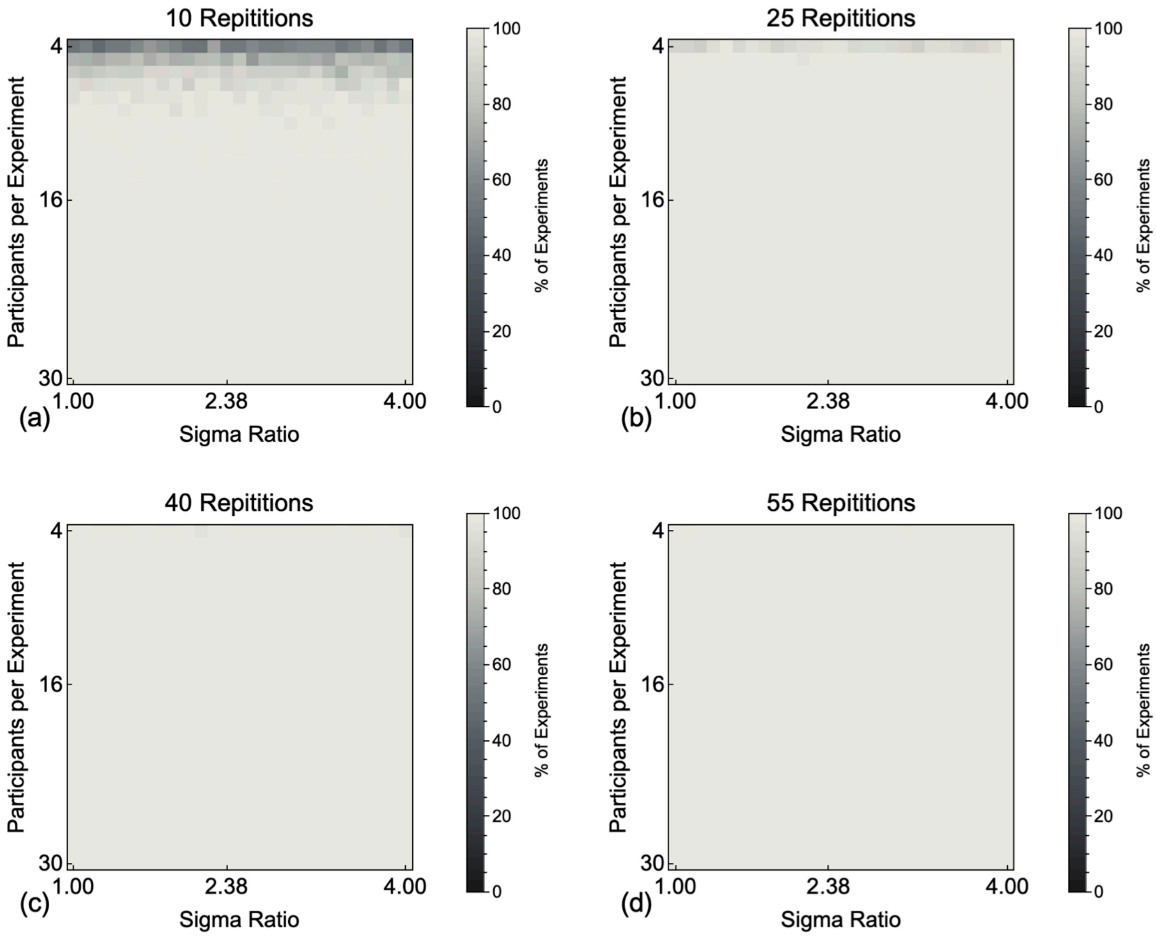

Figure 8: *Shows the percentage of experiments in which the sigmas of the Cumulative Gaussian functions fit to our simulated population of MWF observers could be statistically distinguished from the experimentally derived prediction of PCS. Each pixel in the image shows this percentage as calculated across 100 simulated experiments, of a given sigma ratio and*

number of participants. The four panes show this for (a) 10, (b) 25, (c) 40 and (d) 55, simulated trials per stimulus level on the psychometric function.

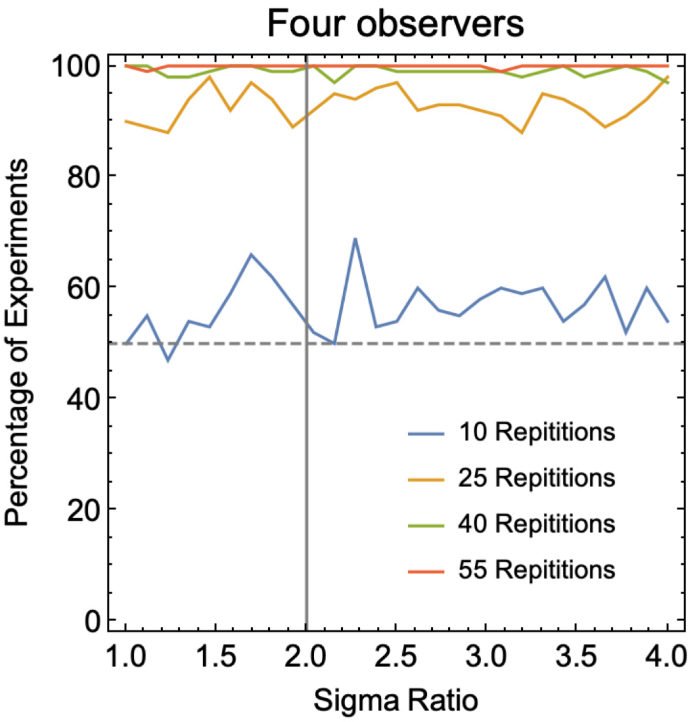

***Figure 9:** Plots the percentage of experiments in which the sigmas of the Cumulative Gaussian functions fit to a simulated population of four MWF observers could be statistically distinguished from the experimentally derived prediction of PCS. The dashed grey line represents the point at which there is a 50% chance of distinguishing the data from the predictions of PCS. The vertical grey line shows the maximum recommended sigma ratio to use in cue integration experiments (Rohde et al., 2016).*

Group analysis: integrated cues percept

Next we examine the extent to which MWF can be distinguished from MS and PCS based upon the predicted integrated cues percept when a discrepancy is experimentally introduced between cues (a ‘perturbation analysis’ (Young et al., 1993)). With zero cue conflict the only differences in $\hat{S}_A$, $\hat{S}_B$ and $\hat{S}_C$ will be due to the simulated data collection and the effect this has on the fit of the psychometric function. Therefore, as expected, when this is the case the predictions of MWF are experimentally near indistinguishable from the predictions of both MS (Figure 10) and PCS (Figure 11). Of note is that there are “false positives” where statistically a population of MWF observers can be distinguished from the predictions of the alternative models, even though the underlying parameters are identical. This is the case for around 16% of simulated experiments for MS (for 10, 25, 40 and 55 repetitions per function, the percentages are 15.99%, 16.09%, 16.21%, and 15.83%) and around 14% of simulated experiments for PCS (for 10, 25, 40 and 55 repetitions, the percentages are (13.62%, 13.69%, 13.82%, and 13.57%). The small difference between the false positives for comparison with MS and PCS is due to the effect that the sigma of the simulated function has on the variability of the *inferred* mean of the function across participants (this differs between models, see Figure 4). More succinctly, whilst the mean and sigma of a Cumulative Gaussian functions are mathematically independent, our ability to *infer* these parameters by fitting psychometric functions to data is not.

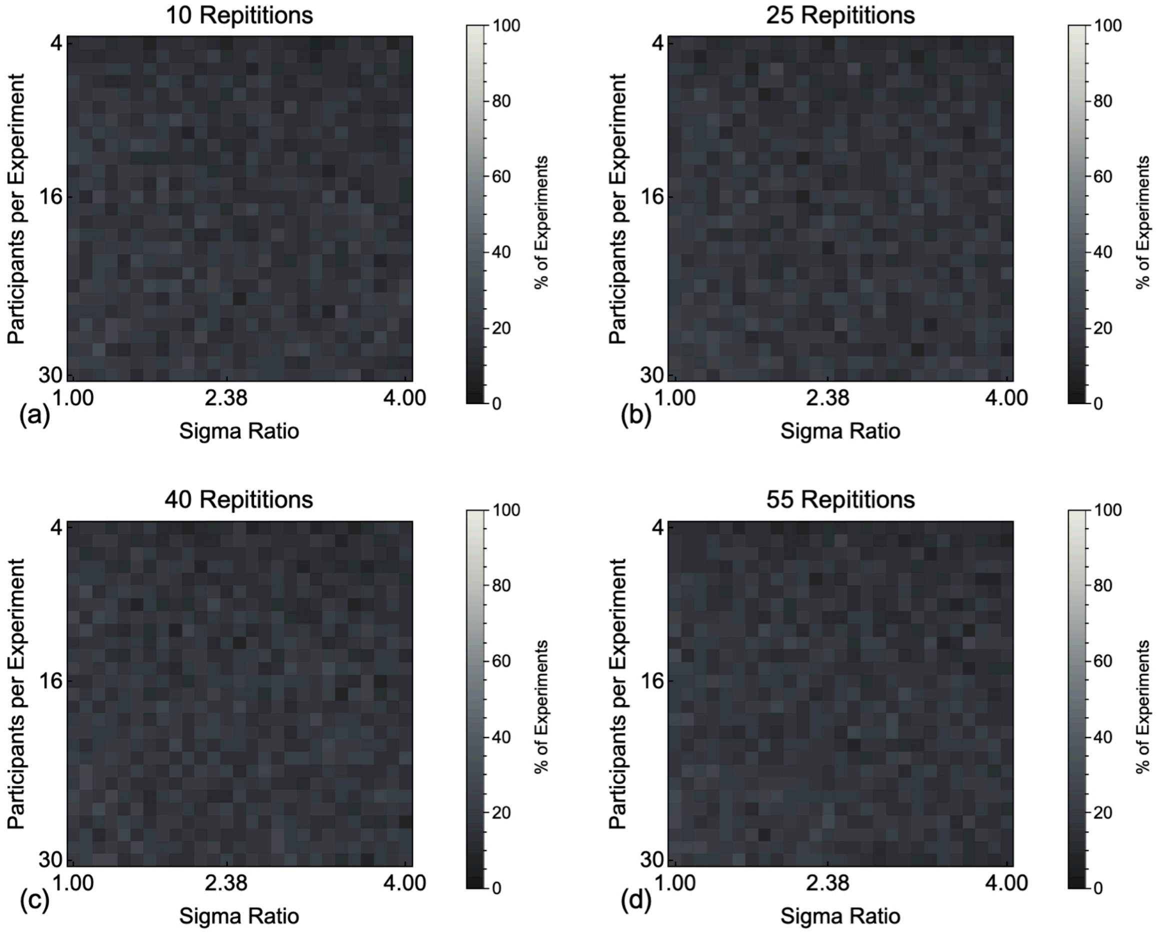

Figure 10: *Shows the percentage of experiments in which the mean of the Cumulative Gaussian functions fit to our simulated population of MWF observers could be statistically distinguished from the experimentally derived prediction of MS, when there is zero cue conflict. Each pixel in the image shows this percentage as calculated across 100 simulated experiments, of a given sigma ratio and number of participants. The four panes show this for (a) 10, (b) 25, (c) 40 and (d) 55, simulated trials per stimulus level on the psychometric function.*

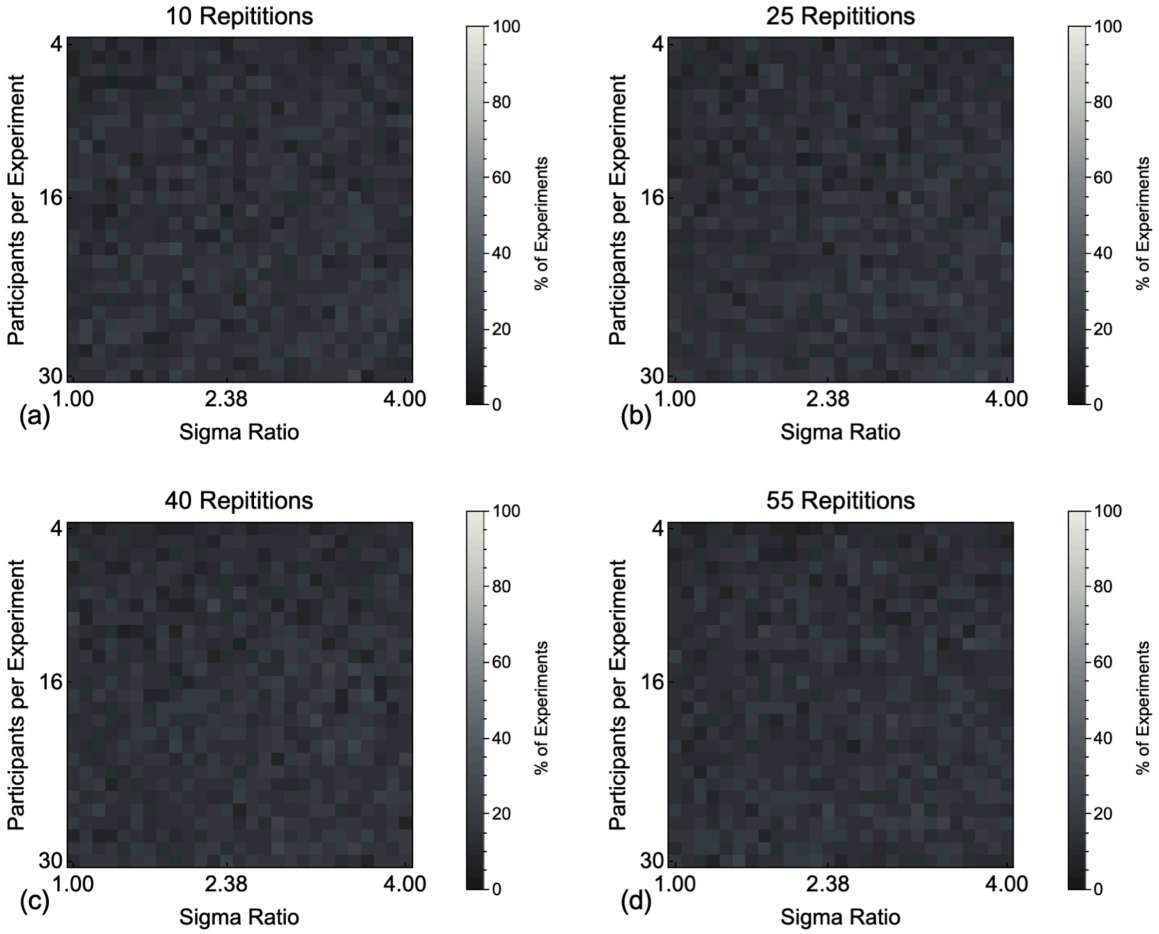

***Figure 11:** Shows the percentage of experiments in which the mean of the Cumulative Gaussian functions fit to our simulated population of MWF observers could be statistically distinguished from the experimentally derived prediction of PCS, when there is zero cue conflict. Each pixel in the image shows this percentage as calculated across 100 simulated experiments, of a given sigma ratio and number of participants. The four panes show this for (a) 10, (b) 25, (c) 40 and (d) 55, simulated trials per stimulus level on the psychometric function.*

Figures 12 and 13 show the data for the 3mm and 6mm cue conflict simulations. As can be seen, the predictions of MWF and MS can now be distinguished, but as with distinguishing models on the basis of the sigma's, the ability to distinguish between models is strongly affected by the relative reliability of the cues and the data collection regime. Consistent with expectations, the probability of distinguishing between models is greater with a larger cue conflict (compare Figure 13 to Figure 12). Due to PCS and MWF providing identical predictions regardless of the experimental cue conflict the only times a population of MWF observers are distinguishable from the predictions of PCS again represent false positives (Figures 14 and 15).

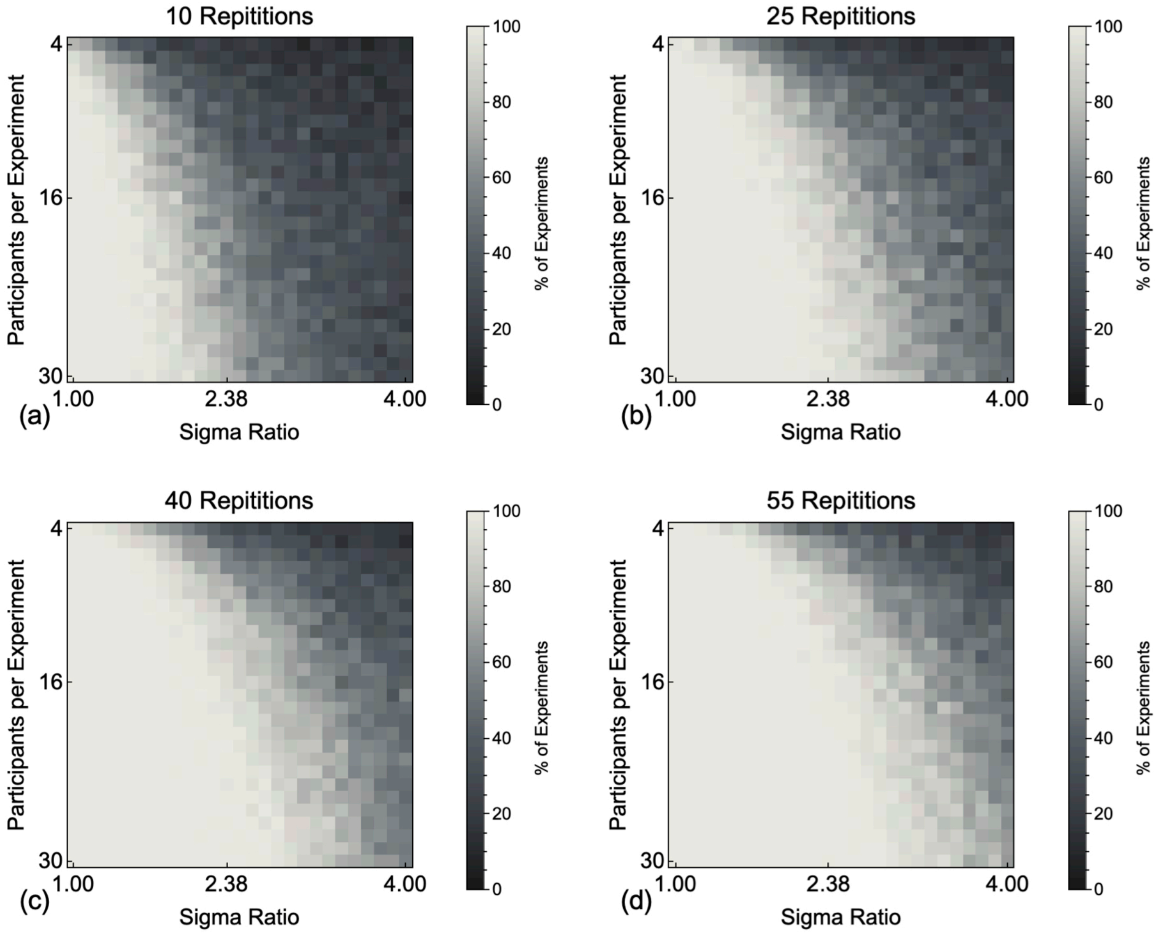

Figure 12: *Shows the percentage of experiments in which the mean of the Cumulative Gaussian functions fit to our simulated population of MWF observers could be statistically distinguished from the experimentally derived prediction of MS with an experimental cue conflict of 3mm. Each pixel in the image shows this percentage as calculated across 100 simulated experiments, of a given sigma ratio and number of participants. The four panes show this for (a) 10, (b) 25, (c) 40 and (d) 55, simulated trials per stimulus level on the psychometric function.*

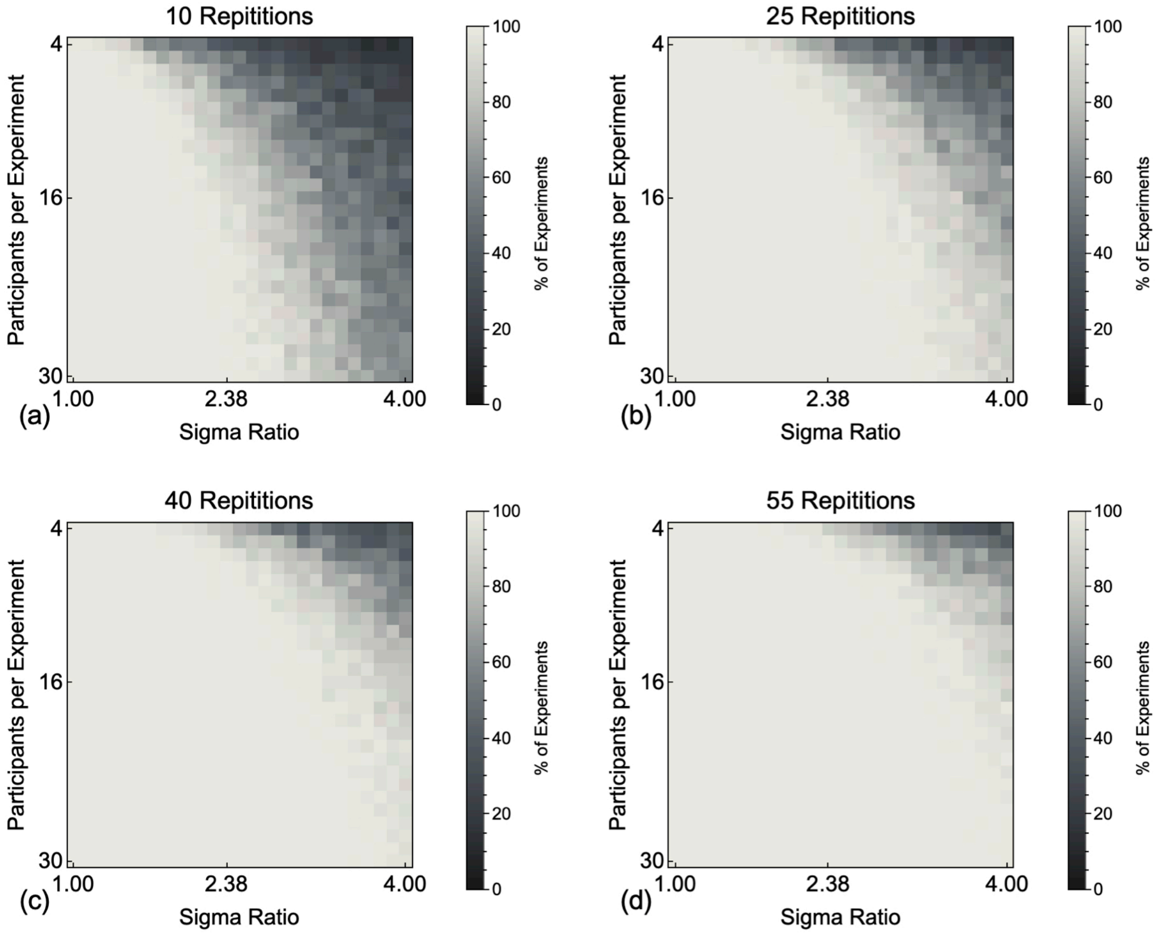

Figure 13: *Shows the percentage of experiments in which the mean of the Cumulative Gaussian functions fit to our simulated population of MWF observers could be statistically distinguished from the experimentally derived prediction of MS with an experimental cue conflict of 6mm. Each pixel in the image shows this percentage as calculated across 100 simulated experiments, of a given sigma ratio and number of participants. The four panes show this for (a) 10, (b) 25, (c) 40 and (d) 55, simulated trials per stimulus level on the psychometric function.*

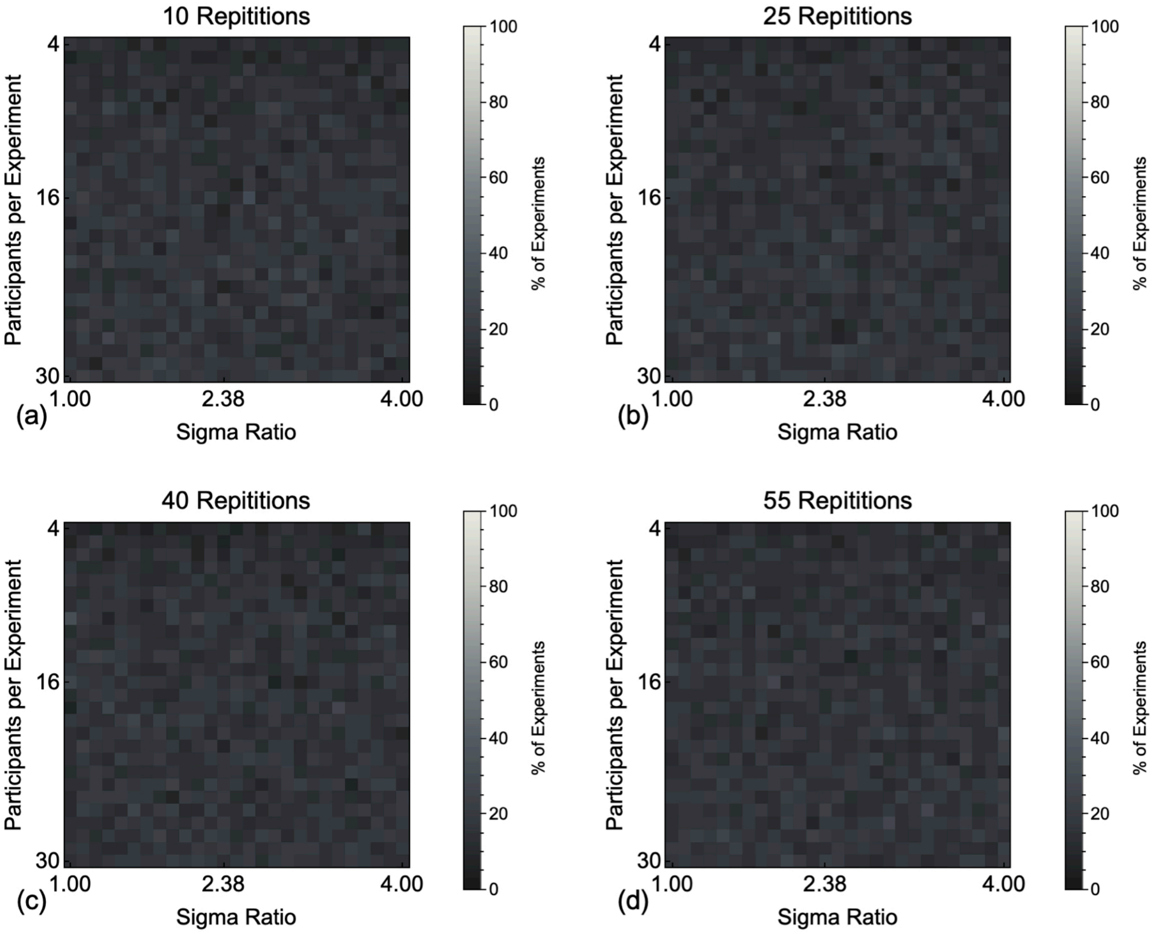

Figure 14: *Shows the percentage of experiments in which the mean of the Cumulative Gaussian functions fit to our simulated population of MWF observers could be statistically distinguished from the experimentally derived prediction of PCS with an experimental cue conflict of 3mm. Each pixel in the image shows this percentage as calculated across 100 simulated experiments, of a given sigma ratio and number of participants. The four panes show this for (a) 10, (b) 25, (c) 40 and (d) 55, simulated trials per stimulus level on the psychometric function.*

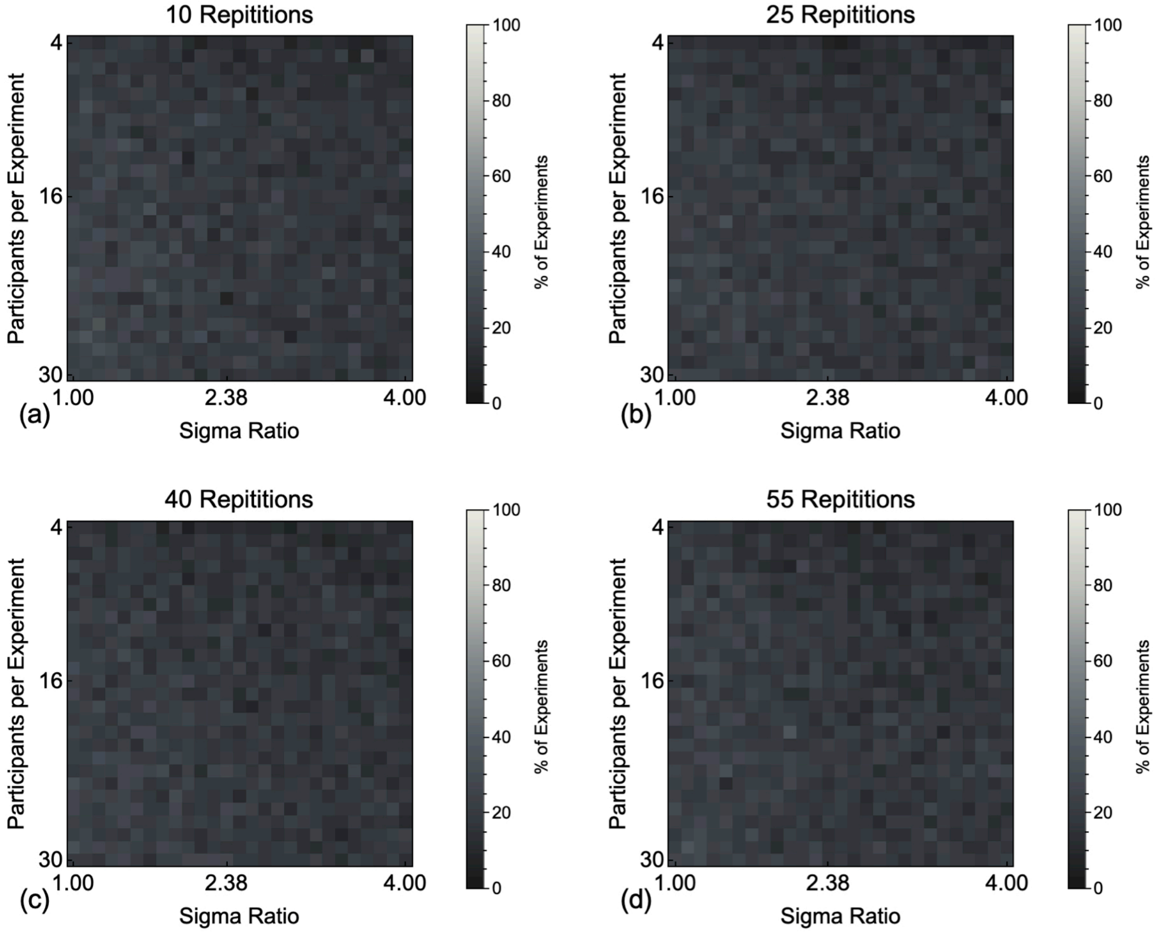

Figure 15: *Shows the percentage of experiments in which the mean of the Cumulative Gaussian functions fit to our simulated population of MWF observers could be statistically distinguished from the experimentally derived prediction of PCS with an experimental cue conflict of 6mm. Each pixel in the image shows this percentage as calculated across 100 simulated experiments, of a given sigma ratio and number of participants. The four panes show this for (a) 10, (b) 25, (c) 40 and (d) 55, simulated trials per stimulus level on the psychometric function.*

In Figure 16 we show the ability to experimentally distinguish between MWF and MS based upon the predicted integrated cues percept for just the simulated experiments with four observers (Ernst & Banks, 2002; Hillis et al., 2004). When there is no cue conflict (Delta 0) there is a false positive rate of around 12% across all data collection regimes and sigma ratios. For both cue conflict values (Delta 3 and 6), the closer the reliability of cues is matched, and the more data collected, the better one is able to discriminate our population of MWF observers from MS. For a Delta of 3, the ability to distinguish models rapidly drops off within the range of sigma ratios acceptable for a cue integration experiment (Rohde et al., 2016). With a sigma ratio of 3 and above, performance is comparable to that of the false positive rate. By comparison, with a Delta of 6, within the range of sigma ratios acceptable for a cue integration experiment the ability to discriminate between models is good, with performance dropping substantially for only the most minimal data collection regime.

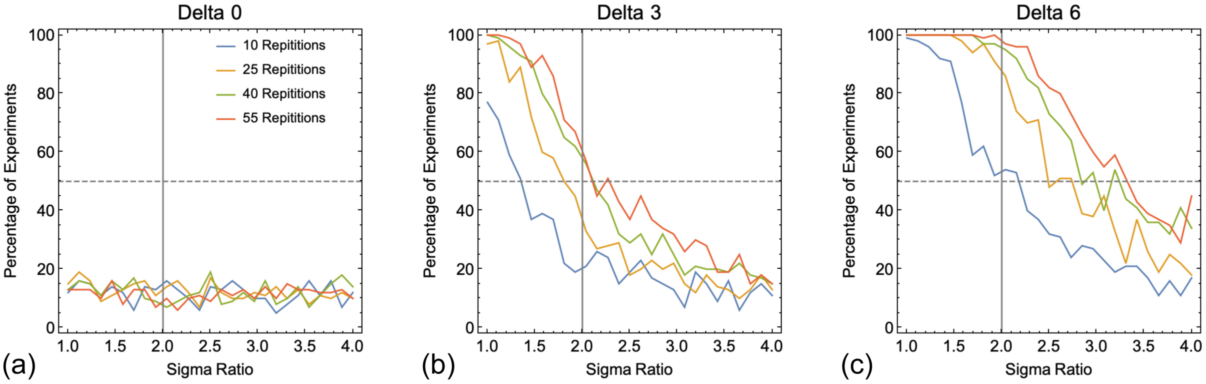

Figure 16: *Plots the percentage of experiments in which the PSE's of the Cumulative Gaussian functions fit to a simulated population of four MWF observers could be statistically distinguished from the experimentally derived prediction of MS. The dashed grey line represents the point at which there is a 50% chance of distinguishing the data from choosing the cue with the minimum variance. The vertical grey line shows the maximum recommended sigma ratio to use in cue integration experiments (Rohde et al., 2016).*

One of the most striking things about the analysis presented is just how rapid the drop-off in an experimenter's ability to distinguish a population of "optimal" MWF observers from the predictions of the two alternative candidate models is, as the reliability of cues becomes unmatched. This is especially true when examining the "essential prediction" of increased integrated cues reliability, in comparison to simply choosing the more reliable cue. MWF observers are easily distinguished from the predictions of PCS in terms of the cue reliability, but impossible to distinguish based upon the integrated cues percept when cues are in conflict. MWF observers can be more easily distinguished from MS based upon the integrated cues percept, but only dramatically so for larger cue conflicts. Problematically, distinguishing models based upon the integrated cues percept alone is *not* sufficient to demonstrate that observers are behaving in accordance with MWF (Rohde et al., 2016).

Simulation Set 2: Using variation across experimental observers to distinguish between models

The simulations presented above were designed to show the effects of the number of observers in and experiment and the relative cue reliability on an experimenter's ability to distinguish a population of optimal observers from two alternative models: choose the cue with the minimum sigma and probabilistic cue switching. As such, all simulated observers in an experiment had matched cue reliabilities. This focuses data collection in areas of the parameter space where MWF, MS and PCS provide most divergent predictions regarding the key signature of MWF (increased sensory precision) (Takahashi, Diedrichsen, & Watt, 2009), but it is not representative of a typical cue integration experiment where there may be variation in cue reliabilities across observers (Hillis et al., 2004; Scarfe & Hibbard, 2011) and properties of the stimuli may naturally (Hillis et al., 2004) or artificially (Ernst & Banks, 2002; Helbig & Ernst, 2007) be used to modulate the relative reliability of cues across experimental conditions. Therefore, in a second set of simulations we examine the case where a where individual observers in an experiment have different relative cue reliabilities.

Methods

For these simulations we focused on comparing MWF and MS. The comparison with PCS is less interesting as its predictions as regards the integrated cues percept are by definition indistinguishable from MWF (Figures 14 and 15, and Equation 21), whereas PCS and MWF are easily distinguished upon the basis of integrated cues sensitivity (Figure 8, Equation 23). In contrast MWF and MS can be distinguished from one another upon both the integrated cues percept and its precision, with this ability clearly being modulated by experimental parameters such as the relative reliability of cues, the number of observers in an experiment and the data collection regime. As before, observers were simulated as having access from two cues ($\hat{S}_A$ and $\hat{S}_B$) from which to make an integrated cues perceptual estimate ($\hat{S}_C$). These cues were in conflict with one another such that $S_A = 55 + \Delta/2$ and $S_B = 55 - \Delta/2$, where, in separate experiments, Δ was either 3 or 6mm.

For each observer in a given experiment, Cue A always had the same variability $\sigma_A = 4.86$, which is approximately that of the haptic cue in Ernst and Banks (2002), whereas Cue B had a variability $\sigma_B = \sigma_A r$ where for each observer r was between 0.5 and 2 (i.e. Cue B twice as reliable as Cue A, through to Cue B half as reliable as Cue A). These limits are consistent with recommendations for the maximum reliability ratio to use in experiments on cue integration (Rohde et al., 2016). To select values with equal probability between these limits, for each observer we generated a random number $\mathrm{x}_i \in [-1, 1]$, and set $r = 2^{x_i}$. Thus, each observer had a different predicted integrated cues PSE and sigma. Separate simulations were run with 4, 12 and 36 observers per simulated experiment, and for 10, 25, 40 and 55 trials per stimulus level. For each combination of (a) data collection regime, (b) number of observers per experiment, and (c) cue conflict (4 × 3 × 2), we simulated 1000 experiments i.e. 32000 experiments with 416000 observers in total.

As before, for each observer in each experiment we simulated performance in two single cues conditions. The parameters derived from the fitted single-cue functions were then fed into the equations for MWF and MS to give predictions for the two candidate models (both in terms of the sigma and PSE of the integrated cues estimator). Next, we simulated the performance of each observer behaving in accordance with both MWF and MS. This allowed us to compare the behaviour of a heterogenous population of observers behaving in accordance with either MWF or MS to the experimentally derived predictions for each model (MWF or MS).

With a heterogenous population of observers the relationship between predicted and observed data are often compared using a linear regression analysis. For example, Burge, Girshick and Banks (2010) examined the perception of slant from disparity and haptic cues and reported an R^2 of 0.60 for predicted versus observed integrated cues sensitivity. Knill and Saunders (2003) also examined the perception of slant, but from disparity and texture cues, and reported R^2 values between around 0.15 and 0.46 for the predicted and observed cue weighting for different base slants. Svarverud et al. (2010) examined "texture-based" and "physical-based" cues to distance and reported R^2 values of about 0.95 for predicted and observed cue weights. The median value of the an R^2 value in these studies is 0.53; in all instances the authors concluded that observers were combining cues optimally in accordance with MWF.

Following these studies, a regression analysis was adopted here. For each experiment the data from the population of observers behaving in accordance with either MWF or MS were plotted against the predictions of each of the two candidate models. The data were fit with first order polynomial by least squares and an R^2 value for the fit of each model to the data calculated. Thus, there were four possible regression comparisons: (1) "MWF v MWF" – experimentally derived predictions of the MWF model, plotted against the data of a population of observers behaving in accordance with MWF; (2) "MS v MS" – experimentally derived predictions of the MS model, plotted against the behaviour of a population of observers behaving in accordance with MS; (3) "MWF v MS" – experimentally derived predictions of the MWF model, plotted against the data of a population of observers behaving in accordance with MS; and (4) "MS v MWF" – experimentally derived predictions of the MS model, plotted against the data of a population of observers behaving in accordance with MWF.

In what follows we will refer to (1) and (2) as "consistent" predicted and observed data, as the simulated data and predictions are from the same model, conversely, we refer to (3) and (4) as "inconsistent" predicted and observed data, as the simulated data and predictions arise from different models. A set of example data from 36 observers behaving in accordance with MWF, with 55 samples per stimulus value and a delta of 3mm, can be seen in Figure 17 for the "MWF v MWF" and "MS v MWF" comparisons. This example represents the upper limit of observers in a typical cue combination experiment (Rohde et al., 2016) and the upper limit of trials per stimulus level for a psychometric function (Kingdom & Prins, 2016). The upper two plots in Figure 17 plot the PSE data from the MWF observers against the experimentally derived predictions of the two candidate models. The green and red dashed lines show the true underlying PSE for each cue.

When plotting the observed data from a population of MWF observers against the experimentally derived predictions of the MWF model (Figure 17a) the data fall between the dashed lines representing the PSE of each cue clustered along the grey unity line. When the data from the MWF observers are plotted against the predictions of the MS model (Figure 17b) the datapoints deviate from the unity line and are clustered around the vertical dashed lines representing the PSE of each cue. The data are clustered around these lines, rather than falling directly upon them as the experimenter has no direct access to the internal parameters of the observer, only estimates of these. Asymptotically, with an infinite amount of data per function, all data in Figure 17a would fall between the dashed vertical lines on the unity line, and all data in Figure 17b directly on the dashed vertical lines.

The lower two plots in Figure 17 plot the observed sigma data from the MWF observers against the experimentally derived predictions of the two candidate models. Here, the dashed red line shows the fixed sigma of Cue A and the green dashed line the minimum possible sigma for Cue B. Thus, if there were no sampling noise, (1) all data points in Figure 17c and Figure 17d would fall to the left of the dashed red line, and (2) in Figure 17d all data would fall to the right of the green dashed line. What is most striking from this example is that the observed R^2 values for both PSE's and sigmas are directly comparable to those found in the literature, regardless of whether the data from a population of MWF observers fit fitted with a regression against the predictions of *either* MWF or MS.

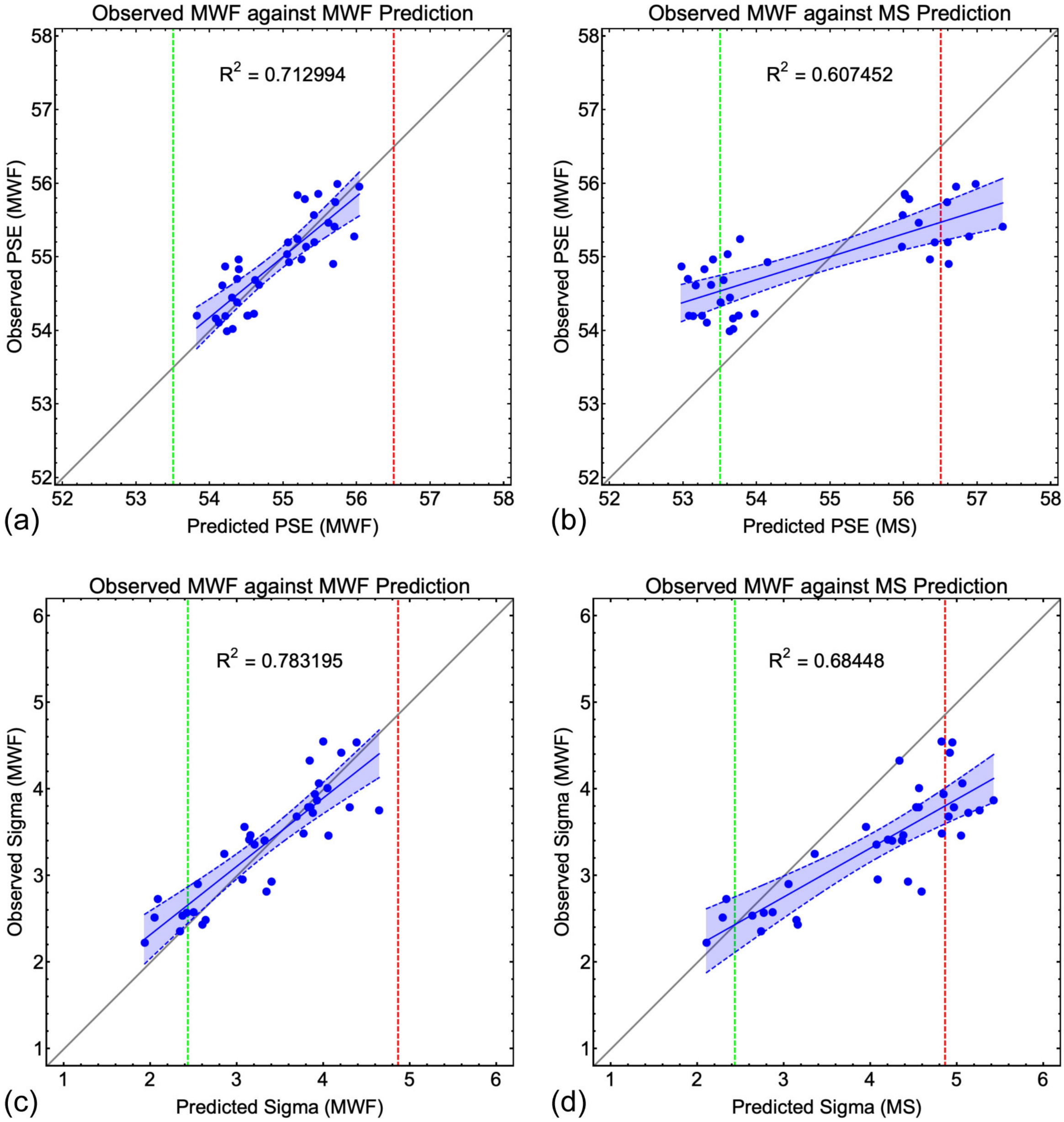

Figure 17: *Linear regression example where the data (blue points) from 36 observers behaving in accordance with MWF (with 55 samples per stimulus value and a delta of 3mm) are plotted against the predictions of the two candidate models (MWF (a and c) and MS (b and d)) for both PSE (a and b) and Sigma (c and d). The least squares first order polynomial is shown as the solid blue line, the dashed blue lines and shaded region in each graph show the 95% confidence bounds around the fit. In (a) and (b) the dashed red line shows the true underlying PSE for Cue A, and the green dashed line shows the true underlying PSE for Cue B. In (c) and (d) the red dashed line shows the (fixed) sigma for Cue A, and the dashed green line the minimum possible sigma for Cue B (which varied across simulated observers).*

Figure 18 shows histograms of the observed R^2 values for the same example, but across all 1000 simulated experiments. The raw histograms are shown overlaid with smooth kernel distributions, given by

$$\hat{F}_X(x) = \frac{1}{nh}\sum_{i=1}^{n} \mathcal{K}\left(\frac{x - x_i}{h}\right) \tag{25}$$

Here $\mathcal{K}$ is a Gaussian kernel function, $x_i \in [0,1]$ (i.e. the domain of the R^2 value is 0 to 1), and $\hat{F}_X$ is the estimate of the unknown probability density function F_x.

The key parameter of interest is the extent to which these distributions overlap, as this determines the extent to which an the R^2 value from fitting predicted to observer data can be used to distinguish between candidate models of cue integration. The overlap of two smooth kernel distributions $\hat{F}_X$ and $\hat{F}_Y$ can be estimated via numerical integration (Pastore & Calcagni, 2019)

$$\hat{\eta}(X,Y) = \int_{1}^{0} min\left(\hat{F}_X(z), \hat{F}_Y(z)\right) dz \tag{26}$$

Numerically the overlap value lays between 0 (no overlap) and 1 (full overlap). This is shown inset into each graph in Figure 18. As can be seen there is substantial overlap in the distribution of R^2 values, especially so for the predicted and observed PSEs. Any overlap in the distributions is clearly problematic for using R^2 values to conclude that observers are behaving in accordance with MWF.

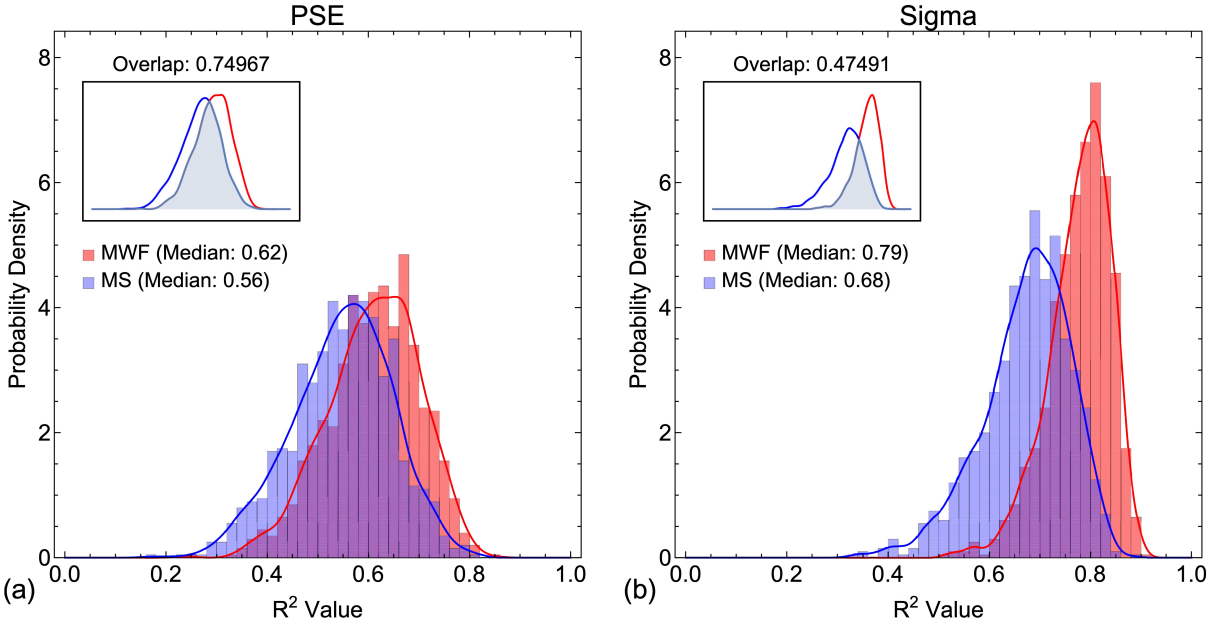

Figure 18: *Shows the full distribution for the R^2 value across all 1000 simulated experiments for the example shown in Figure 17 (36 observers per simulated experiment, each behaving in accordance with MWF, 55 samples per stimulus value and a delta of 3mm). (a) shows data for the PSE and (b) data for sigma. Data are show as bar histograms and as smoothed histograms (smoothed Gaussian kernel distribution; Equation 25). Blue data show the case where the data from the simulated MWF observers is plotted against the predictions of MWF, red data show the case where the data from the simulated MWF observers is plotted against the predictions of MS. The median for each data set is shown in the graphs. The inset graph shows the overlap of the smoothed histograms (Equation 26). Note that the axes of the inset graphs is smaller to ensure clarity of the overlapping region.*

Data across all comparisons for both PSE and sigma are shown in Figures 19, 20 and 21, for the 4, 12, and 36 participants per experiment conditions respectively. As one would expect, with more data collected per function and more observers per experiment the R^2 values improve, with a maximal median of ~0.7-0.8. Problematically, this pattern is present regardless of whether one is plotting consistent predicted and observed data (MWF v MWF and MS v MS), or inconsistent predicted and observed data (MWF v MS and MS v MWF). Across all plots there is the large overlap in the distributions of R^2 values when plotting “consistent” and “inconsistent” predicted and observed data. With fewer observers per experiment (4 and 12 versus 36) the overlap increases greatly, to the extent that with four observers per experiment the data have near complete overlap.

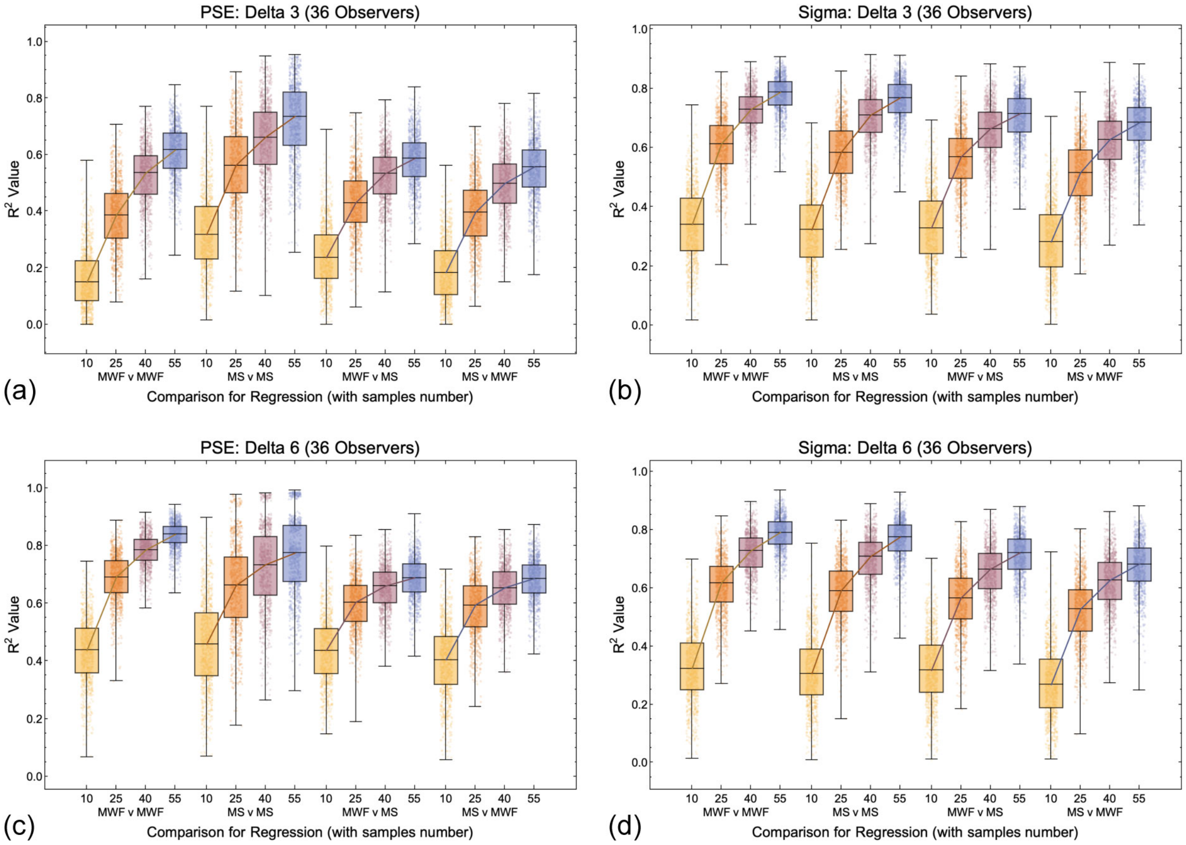

Figure 19: *Box and whisker plots showing the distribution of R^2 values for all conditions and comparisons in which there were 36 simulated observers per experiment. The central box line shows the median (also shown as a line connecting the boxes), the limits of the boxes show the 25% and 75% quantiles and the limits of the bars (whiskers) show the maximum and minimum values. Also shown are all 1000 datapoints per condition (dots). For increased clarity the dots have been randomly jittered laterally.*

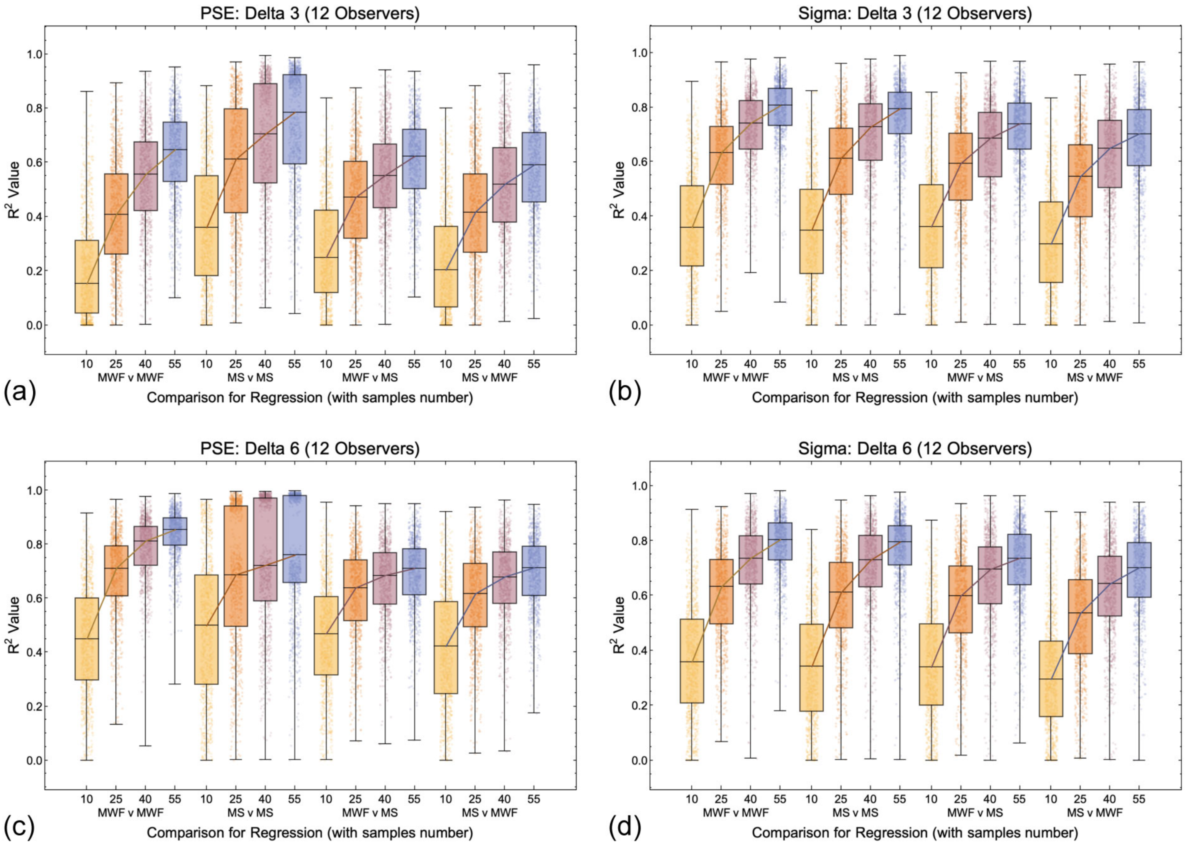

Figure 20: *Box and whisker plots showing the distribution of R^2 values for all conditions and comparisons in which there were 12 simulated observers per experiment. The central box line shows the median (also shown as a line connecting the boxes), the limits of the boxes show the 25% and 75% quantiles and the limits of the bars (whiskers) show the maximum and minimum values. Also shown are all 1000 datapoints per condition (dots). For increased clarity the dots have been randomly jittered laterally.*

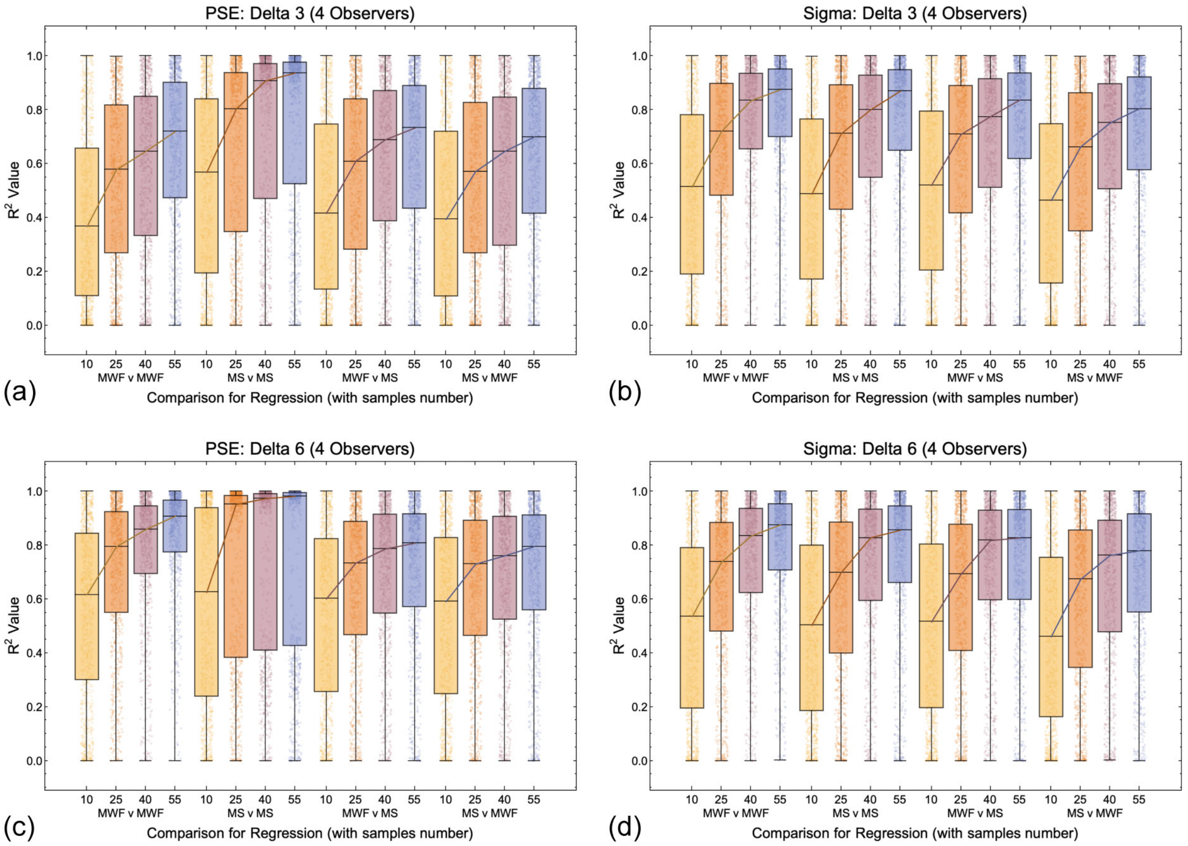

Figure 21: *Box and whisker plots showing the distribution of R^2 values for all conditions and comparisons in which there were 4 simulated observers per experiment. The central box line shows the median (also shown as a line connecting the boxes), the limits of the boxes show the 25% and 75% quantiles and the limits of the bars (whiskers) show the maximum and minimum values. Also shown are all 1000 datapoints per condition (dots). For increased clarity the dots have been randomly jittered laterally.*

Figure 22 shows the overlap (Equation 26) for the distributions where a population of observers behaving in accordance with MWF were compared to the experimentally derived predictions of MWF and MS. Whereas, in Figure 23 shows the overlap for the distributions were a population of observers behaving in accordance with MS were compared to the experimentally derived predictions of MWF and MS. Consistent with Figures 19, 20 and 21, the distribution overlap decreases with increasing amounts of data collected per function. As expected, for the PSE distributions, the distribution overlap is less with a Δ of 6mm versus 3mm, and the delta magnitude has no effect on the overlap of the sigma distributions. As is clear, distribution overlap is greater than 50% (overlap fraction of 0.5) for virtually all conditions. A large overlap means that it is not possible to determine whether a given R^2 value arises from plotting and analysing "consistent" and "inconsistent" predicted and observed data. This strongly questions one's ability to use R^2 values to assess the extent to which a set of data is consistent with the predictions of MWF.

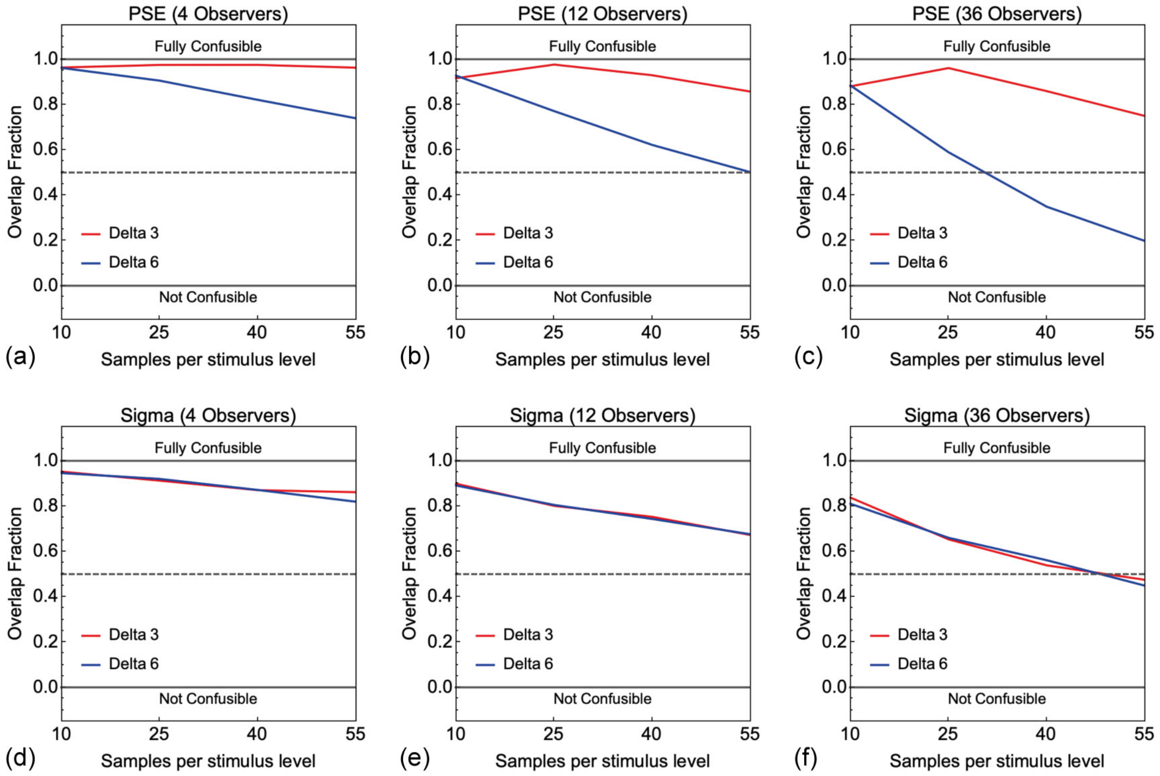

Figure 22: *Overlap of the smooth kernel distributions of R^2 value values produced from fitting a first order polynomial to observed data from a set of MWF observers against the experimentally derived predictions of MWF and MS. An overlap value of 1 (upper solid grey line) means that the distributions completely overlap and are fully confusable (100% overlap) and overlap of 0 (lower solid grey line) means that the distributions do not overlap at all and are thus not confusable (0% overlap). The dashed grey line shows the case where the distributions overlap by 50%. Panels (a) through (c) show data for PSE's, and (d) through (f) data for sigmas. Columns are number of observers per simulated experiment.*

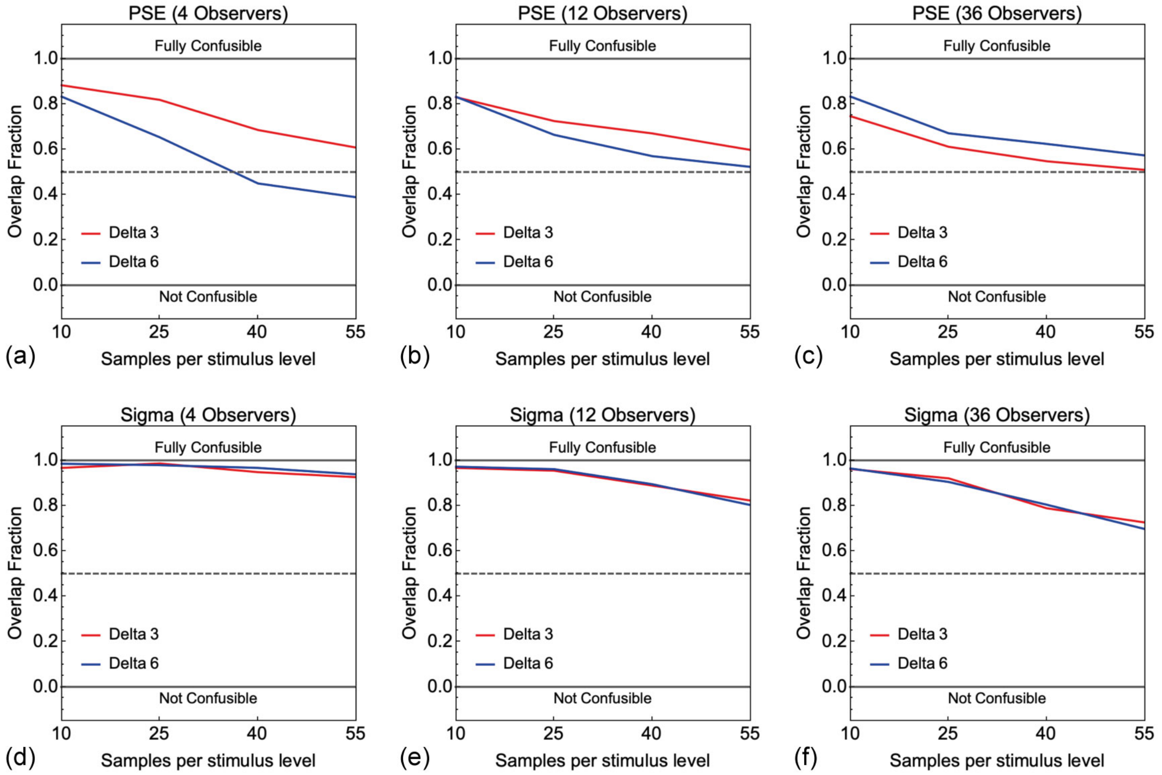

Figure 23: *Overlap of the smooth kernel distributions of R^2 value values produced from fitting a first order polynomial to observed data from a set of MS observers against the experimentally derived predictions of MWF and MS. An overlap value of 1 (upper solid grey line) means that the distributions completely overlap and are fully confusable (100% overlap) and overlap of 0 (lower solid grey line) means that the distributions do not overlap at all and are thus not confusable (0% overlap). The dashed grey line shows the case where the distributions overlap by 50%. Panels (a) through (c) show data for PSE's, and (d) through (f) data for sigmas. Columns are number of observers per simulated experiment.*

Interestingly we also observed that for a low number of observers per experiment (especially coupled with low samples per stimulus level), in addition to the R^2distributions having near complete overlap, there was a distinct peak around an $R^2 \approx 0$. This is because with such an impoverished data collect regime the simulated data could show both a positive or negative relationship between the predicted and observed PSE or sigma. Thus, functions with an $R^2 \approx 0$ all fell into the minimum histogram bin but could be due to either a shallow positive *or negative* relationship between the predicted and observed data (i.e. slope is a signed variable, whereas, R^2 is unsigned). Together with the previous simulations this highlights the fact that a low number of observers per experiment (Ernst & Banks, 2002; Hillis et al., 2004), coupled with a small number of samples per point on a psychometric function (Rohde et al., 2016), is insufficient to test models of sensory cue integration.

A further problem with this type of analysis is that the R^2 value does not measure the deviation of the data from predictions of a model (dark grey diagonal lines in Figure 17), rather, the R^2 gives a measure of the fit *relative to the fitted polynomial*. As such, one can obtain an $R^2 = 1$ when the predicted and observed data have highly discrepant values (Figure 24a) as well as the opposite predicted relationship (Figure 24b). Clearly an experimenter would not conclude that the data in Figure 24b support the MWF model, however, it is not clear what would be conclude in the case of Figure 24a. The magnitude of the offsets (black arrows) can take *any* value and still give an $R^2 = 1$. Thus, a regression analysis negates one of the key benefits of MWF (and other models), which is the ability to, in advance, predict *absolute* value of the integrated cues percept and its reliability and compare this to that observed experimentally (Lipton, 2005).

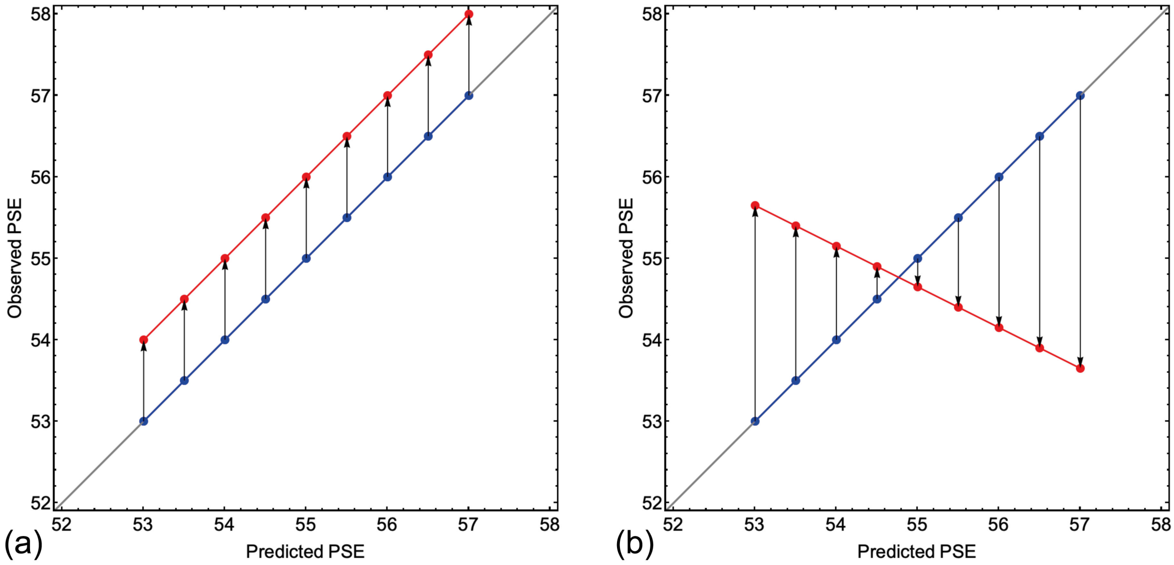

Figure 24: *Shows some hypothetical cue integration data for predicted and observed PSE's in the same format to Figure 17a and b. The blue data points show example data where there is an absolute one-to-one relationship between predicted and observed PSE's. The blue line shows the best fit (least squares) first order polynomial, which aligns with the absolute one-to-one relationship shown by the dark grey line. The red data points are offset relative to the blue data points in (a) this offset is $y = x + 1$, in (b) this offset is $y = 82.15 - 0.5 * x$. The red line shows the best fit (least squares) first order polynomial. In all cases $R^2 = 1$.*

Discussion

Seeing perception as a process of probabilistic Bayesian inference offers a powerful and elegant way in which to make principled predictions about how observers should make perceptual decisions (Adams & Mamassian, 2004; Kersten et al., 2004; Knill & Pouget, 2004;

Knill & Richards, 1996; Mamassian et al., 2002). This is part of what is called an *Ideal Observer Analysis*, whereby a normative mathematical model can be used to define the "optimal" way in which to make discissions based upon noisy and ambiguous sensory information and these predictions can then be compared to human performance (Geisler, 2011). The MWF model of cue integration is one such normative model (Landy et al., 2011; Landy et al., 1995; Maloney & Landy, 1989) and a large body of research purports to show that human observers integrate sensory cues in accordance with MWF, both within and between the senses (Ernst & Banks, 2002; Helbig & Ernst, 2007; Hillis et al., 2004; Scarfe & Hibbard, 2011; Svarverud et al., 2010). As such, sensory cue integration has been described as the "… poster child for Bayesian inference in the nervous system" (Beierholm et al., 2009, p. 1).

In any area of science it is the job of a scientist to design experiments which are able to best distinguish between alternative models of the underlying phenomena. Unfortunately, in the area of cue integration, this is rarely done. Despite there being a wide range of competing models for how human observers might integrate information from sensory cues (Beierholm et al., 2009; Jones, 2016; K. P. Kording et al., 2007; Mamassian et al., 2002; Trommershauser et al., 2011), in many instances the results of an experiment are simply visually inspected relative to the predictions of the experimenters preferred model. Indeed, a recent tutorial detailing how to test for optimal cue integration has proposed a visual taxonomy by which one can judge the fit of a set of data to MWF by visual inspection of the data, associated error bars and model predictions (Rohde et al., 2016). This taxonomy has started to be used researchers to assess the adherence to MWF (Negen et al., 2018).

This approach is problematic as leading researchers have fundamental misconceptions about how error bars relate to statistical significance and the way in which they can be used to support statistical inferences from data (Belia et al., 2005; Cumming et al., 2007). This includes the fields of psychology and behavioural neuroscience where models of sensory cue integration are tested. A visual taxonomy is even more problematic because candidate models often make highly correlated predictions to one another (Arnold et al., 2019 and Figures 4 and 5; Jones, 2016). Therefore, "eyeballing" data relative to an experimenters' preferred model is not a sufficient method to (a) determine the fit of the preferred model to the data, or (b) distinguish between competing models, and will likely result in inferential errors about computational phenomena that have resulted in the measured behaviour.

In the present paper we first sought to draw attention to the many, often unacknowledged, assumptions an experimenter makes when fitting and modelling data from a cue integration experiment and second, introduce a principled objective method by which to determine the probability with which alternative models of the data can be distinguished in a given experiment. The second aim was accomplished by simulating end-to-end experiments and examining the probability with which a population of observers behaving in accordance with MWF (or MS) could be experimentally distinguished from the predictions of a set of alternative models. We showed that the low number of observers in typical cue integration experiments, coupled with unmatched cue reliabilities, results in a widespread inability to test the adherence of a set of data to the predictions of MWF.

As one would expect, there was a clear link between (1) the correlated predictions of different models for a given metric, and (2) an experimenter's ability to distinguish between these models based on this metric. PCS is a clear example of this; it can be easily distinguished from MWF in terms of the sigma of the integrated cues estimator but is impossible to distinguish on the basis of the integrated cues percept. The problem inherent in distinguish between models for sensory integration was present even under conditions where all observers in an experiment have matched cue reliabilities (where MWF predicted the greatest gain in sensory precision) (Takahashi et al., 2009). At all decision points the simulations were designed to be (1) consistent with published guidelines stating how to test models of cue integration (Rohde et al., 2016), (2) consistent with the existing literature (Ernst & Banks, 2002), and (3) consistent with best practice as regards experimental methods (Frund et al., 2011; Kingdom & Prins, 2016; Prins, 2012, 2013; Rohde et al., 2016; Wichmann & Hill, 2001a, 2001b).

In addition to this, many of the nuisance parameters which would impede an experimenter's ability to distinguish between models were not simulated. For example, for our simulated observers there was (1) statistical independence between trials, with no learning or boredom effects (Frund et al., 2011), (2) a known generative function underlying behaviour (Kingdom & Prins, 2016; Murray & Morgenstern, 2010), (3) no perceptual bias (Scarfe & Hibbard, 2011), (4) stimulus values for the psychometric function were centred on the true mean of the psychometric function, (5) simulated observers exhibited no lapses (Prins, 2012; Wichmann & Hill, 2001a, 2001b), (6) the simulated data were not contaminated by the effect of decisional (or other sources of) noise (Hillis et al., 2004), (7) cues were statistically independent from one another (Oruc et al., 2003) and (8) there were no conflicting sources of sensory information (Watt et al., 2005). As these assumptions are known to be violated in many, if not all, experimental settings, the simulations presented are highly likely to *overestimate* one's ability to distinguish between models.

Can single cues truly be isolated?

A grounding assumption of the cue integration literature (and the presented simulations) is that there exist separable sources of sensory information which provide independent perceptual estimates about properties of the world (Ernst & Bulthoff, 2004). In practice, it rapidly becomes apparent just how difficult it is to experimentally isolate sensory cues and to eliminate alternate cues which are not of interest. This is true even with artificial stimuli for which the experimenter has control over all aspects of the stimulus. For example, random dot stereograms (Julesz, 1971) which are used to measure thresholds for the disparity cue normally contain a potentially useful texture cue (changes in dot density) and a conflicting texture cue (fixed aspect ratio of the dots defining the surface). In an elegant exposition, Zabulis and Backus (2004) show the great lengths that one needs to go to in order to create a "texture" which can be used in random dot stereogram which contains no other cues to depth.

Watt et al. (2005) clearly showed how focus cues in computer generated stimuli, which have been typically assumed to be weak and of little utility, can in fact influence perceived depth and surface orientation. This feeds into the ongoing debate as to what "cues" observers use in an experimental setting to perform the tasks that experimenter asks of them, and whether these cues are the ones identified by the experimenter (Saunders & Chen, 2015; Todd, 2015; Todd et al., 2010). Ho, Landy and Maloney (2006), in a roughness constancy task, nicely showed how "pseudocues" (cues which were valid indicators to roughness under a single lighting condition, but not invariant under changes in lighting condition) could influence observers' perceptions of surface roughness. These examples all represent instances where "nuisance cues" contaminate measurements related to the cues of interest to the experimenter, but there are also instances where pre-existing identified cues will also contaminate measurements made by an experimenter.

In a fMRI imaging study Murphy, Ban and Welchman (2013) measured slant thresholds for texture and disparity cues. When measuring a "single-cue" texture threshold, they avoided monocular presentation, which is normally used to isolate the texture cue, as it "… is known to significantly affect both univariate and multivariate fMRI responses" (p. 192), so during measurements of a texture threshold disparity always signalled zero slant. Similarly, when measuring a "single-cue" disparity threshold, slant from disparity was varied but texture always signalled zero slant. A similar approach was taken by Svarverud et al. (2010) in a virtual reality study investigating "texture-based" and "physical-based" cues to distance. In each instance, "single-cue" functions were measured in the presence of the "other" class of cue which was held constant. In both of these cases measurements of single cue sensitivities will have been contaminated by the presence of a conflicting cue that was held constant. Under these circumstances it demonstratively false that the conflicting cue is weak and can thus be ignored.

Controlling for the effects of conflicting cues when measuring "single cue" sensitivities

Here, we briefly examine the consequences of inferring single cue sensitivities in the presence of a conflicting sensory cue and how this can be mathematically controlled for. Let's assume that an experimenter is using a two-interval forced choice experiment to measure the sensitivity of a cue S_A for judgements of size. On each trial, in one interval the experiment presents a "standard" stimulus and in the other interval a "comparison" stimulus, the difference between these being ΔS_A. The observer has to signal in which interval the "larger" stimulus was presented. Next, let's assume that this is done in present of a conflicting "nuisance" cue, ΔS_N, which is constant and signals that the stimulus is unchanged across intervals. This means that the "single cue" stimulus is in fact an integrated cues stimulus and following Equation (2) can be described as

$$\Delta S_c = w_A \Delta S_A + w_N \Delta S_N \tag{25}$$

For each stimulus value $\Delta S_c(i)$ the experimenter measures $p\left("larger"\middle|\Delta S_c(i)\right)$ and with the assumption that the "standard" and "comparison" stimuli can be represented by Gaussian probability density functions, maps out a psychometric function by (incorrectly) plotting $p\left("larger"\middle|\Delta S_c(i)\right)$ against $\Delta S_A(i)$ and fits a Cumulative Gaussian to the data. Clearly, the experimenter will incorrectly estimate σ_A from this fitted function. More specifically, they will overestimate σ_A because each stimulus that they present is in fact an attenuated version of that which they intended (i.e. $\Delta S_c(i) < \Delta S_A(i)$). The extent to which the experimenter misestimates σ_A will be a function of w_N (the weight given to the nuisance cue S_N, which is signally no change across intervals). As $\sigma_N \rightarrow \infty$, the weight given to the nuisance cue will approach zero, $w_N \rightarrow 0$, and σ_A will be estimated accurately. However, for any non-infinite value of σ_N, the experimenter will misestimate σ_A.

In effect, what one needs to do is "warp" the x-axis of the measured psychometric function such that one is plotting $p("larger")$ against $\Delta S_c(i)$ instead of $\Delta S_A(i)$ (Figure 24). To determine this "warping", we can ask, what scale factor, k, would we would need to apply to ΔS_A such that in all cases $\Delta S_c = \Delta S_A$.

Recognising that $w_N = 1 - w_A$, we can write this as

$$\Delta S_A = \Delta S_c = w_A(\Delta S_A * k) + (1 - w_A)\Delta S_N \tag{26}$$

Solving for k, we get

$$k = \frac{\Delta S_A - \Delta S_N}{\Delta S_A * w_A} \tag{27}$$

Given that $\Delta S_N = 0$ this simplifies to

$$k = \frac{1}{w_A} \tag{28}$$

Intuitively we can see that this makes sense, as when $w_A = 1$, no scaling is required to combat the attenuation caused by ΔS_N, because it receives zero weight, however, as soon as $w_A < 1$, scaling is needed (i.e. $k > 1$). Next, we can ask, given the true value of σ_A, what would be our estimate, $\hat{\sigma}_A$, of this be in the presence of the conflicting nuisance cue. To do this we recognise that for a probability density function of a random variable X distributed according to $F_X(x)$, the probability density function of a variable $Y = g(X)$ is also a random variable. If g is differentiable and $g\colon \mathbb{R} \to \mathbb{R}$ is a monotonic function, we can then use a *change of variables* to transform between probability density functions.

$$F_Y(y) = F_X(x)\left|\frac{dx}{dy}\right| \tag{29}$$

Here, $(x) = g^{-1}(y)$ and the support of Y is $g(x)$ with the support of X being x (Blitzstein & Hwang, 2015). For our example, the Gaussian probability density function representing our cue S_A, can be written as

$$F_X(x) = \frac{1}{\sigma_A\sqrt{2\pi}}e^{-\frac{(\mu-x)^2}{2\sigma_A^2}} \tag{30}$$

This function has a mean of μ and standard deviation of σ_A. From Equation 28, using the transform $x * k$, a *change of variables* gives

$$F_Y(y) = \frac{w_A}{\sigma_A\sqrt{2\pi}}e^{-\frac{(\mu-w_A*x)^2}{2\sigma_A^2}} \tag{31}$$

This represents our experimentally inferred probability density function for cue S_A. The standard deviation of $F_Y(y)$ is given by

$$\sigma_A{}' = \frac{\sigma_A}{w_A} \tag{32}$$

Therefore, if the weight given to a nuisance cue is greater than zero, w_A will be less than 1 and we will therefore overestimate the true value of σ_A. In Figure 24 we show a visual example of the process described above.

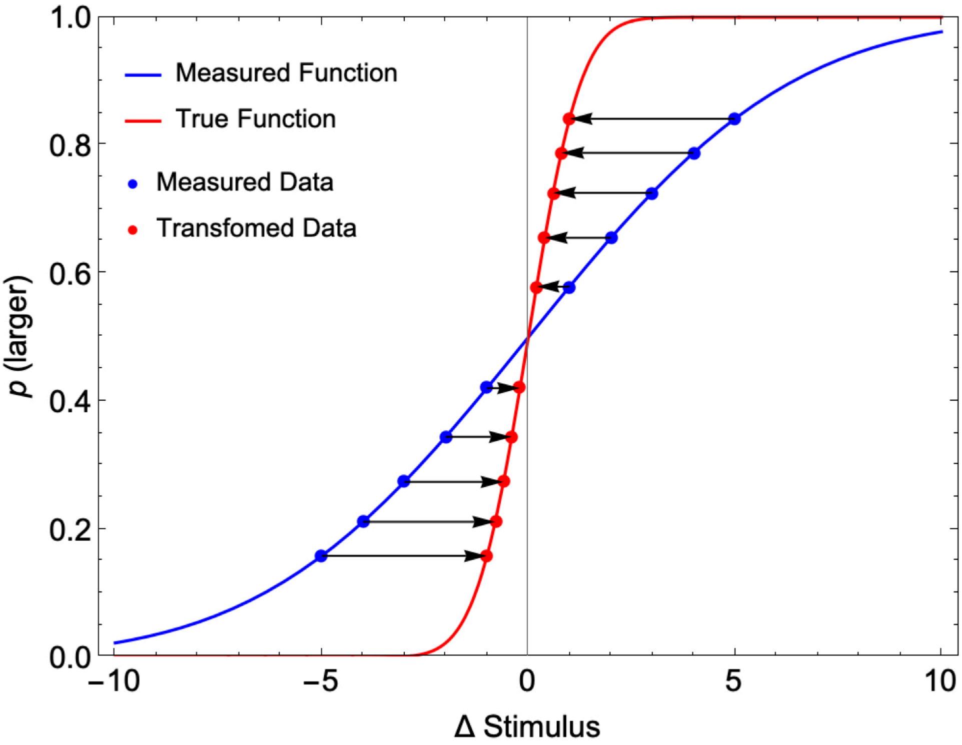

***Figure 24**: An experimenter presents a range of stimuli ΔS_A ($\Delta\ Stimulus$) and for each of these measures the probability of a "larger" response ("Measured Data", shown as blue points). This is done in the presence of a conflicting cue, S_N, which signals no change across intervals. For this visual example, $\sigma_A = 1$ and $\sigma_N = 0.5$, therefore $w_A = 0.2$ and $w_N = 0.8$ (Equations 3 and 4). The experimentally measured points are consistent with a measured psychometric function (blue Cumulative Gaussian function (given by Equation 31)). This function has a standard deviation $\hat{\sigma}_A = \frac{\sigma_A}{w_A} = 5$. In reality, each stimulus $\Delta S_A(i)$ is in fact cue conflict stimuli $\Delta S_C(i)$ (given by Equation 25), thus the data should be shifted along the x-axis toward $\Delta\ Stimulus = 0$ (black by arrows) to accurately plot the function. These shifted points ("Transformed Data", shown as red points, Equation 26) are consistent with the true underlying psychometric function for the cue S_A (red Cumulative Gaussian function (given by Equation 30)). This function is steeper than the (measured) blue function because for a measured p(larger), the Δ Stimulus was in fact smaller than the experimenter had planned (due to the cue conflict).*

Given the derivations above, we can work out the consequences this has for measuring the relative reliability of cues, which is the key variable needed for testing MWF. Let's say we have two cues S_A and S_B with standard deviations of σ_A and σ_B signalling a property of interest, S. We measure "single cue" sensitivity functions for each cue whilst holding the other cue constant (Murphy et al., 2013; Svarverud et al., 2010). Recognising that $1/\sigma_A^2 + 1/\sigma_B^2$ is a constant, c, we can rewrite Equations (3) and (4) as

$$w_A = \frac{1}{c * \sigma_A^2} \tag{33}$$

and

$$w_B = \frac{1}{c * \sigma_B^2} \tag{34}$$

Given Equation 32, we can write

$$\hat{\sigma}_A = c * \sigma_A^3 \tag{35}$$

and

$$\hat{\sigma}_B = c * \sigma_B^3 \tag{36}$$

These represent our experimental *estimates* of the true underlying standard deviations. They are each larger than the true underlying values as they have been measured in the presence of a cue signally no change (Figure 24). The ratio of these estimates is given by

$$\frac{\hat{\sigma}_A}{\hat{\sigma}_B} = \frac{\sigma_A^3}{\sigma_B^3} \tag{37}$$

The ratio of the true underlying sigma's, which is the property we wish to estimate, is given by

$$\frac{\sigma_A}{\sigma_B} = \sqrt[3]{\frac{\hat{\sigma}_A}{\hat{\sigma}_B}} \tag{38}$$

Therefore, if we infer from our experiment that $\sigma_A/\sigma_B = 1/27$ the true sigma ratio is in fact $1/3$. That is, we experimentally misestimate σ_A/σ_B by a factor of ~ 9. Studies which have measured the reliability of cues in the presence of a secondarily constant conflicting cue, (e.g. Murphy et al. (2013) and Svarverud et al. (2010)), are therefore likely to have significantly overestimated the true cue relative reliabilities. As such, the data in these studies cannot be used to accurately test MWF, without some form of correction (Equation 38). This analysis shows the critical importance of being able to (1) isolate singles cues satisfactorily, or if one is not able to, (2) correct for their influence on one another when inferring relative cue reliabilities.

Conclusion

The simplicity of the MWF equations for cue integration is deceptive, as a model's simplicity is generally directly correlated with the number of assumptions it makes about the underlying phenomena. With more assumptions and a simpler model, there is a greater chance that the assumptions of the model will not be met, and this will impact an experimenter's ability to accurately test the predictions of the model. For example, even the comparatively "simple" task of fitting a Cumulative Gaussian function to experimental data, that is one of the initial steps of a cue integration experiment, has many inbuilt assumptions which are rarely acknowledged or considered (Kingdom & Prins, 2016). Certainly, for the author (who has conducted cue integration experiments (Scarfe & Hibbard, 2011, 2013)) this was an illuminating experience.

Even if one can be satisfied that the assumptions of MWF hold in a given experimental situation, problematically, MWF provides correlated predictions with many other cue integration models (Arnold et al., 2019; Beierholm et al., 2009; Jones, 2016). Here we considered two such models; choose the cue with minimum sigma and probabilistic cue switching. It was shown that even when adopting published criteria describing how to best test the predictions of MWF (Rohde et al., 2016), it was very difficult to experimentally disambiguate between MWF and these alternative models. The analysis presented is only scratching the surface, as there are many other ways in which sensory cues could be integrated (Jones, 2016), some of which may be even more difficult to disambiguate from MWF.

Unfortunately, some of the most widely cited studies supporting MWF fail to consider alternative models satisfactorily, sample areas of the parameter space which poorly distinguish between competing models, and provide no statistical analysis related to the fit of the MWF to the data, or the relative fit of other models to the data. This questions the ability of these studies to conclude that sensory cues are integrated in an optimal fashion in accordance with MWF. Whilst one could interpret the results presented here in a pessimistic fashion, the opposite is true. The results presented show clear, simple and computationally attainable ways in which experimenters can correctly measure the variables needed to test models of cue integration and determine the probability with which a given experiment can distinguish between alternative models of the underlying phenomena.

Acknowledgements

The seeds of this project arose in the authors co-supervision, with Prof. Andrew Glennerster, of the PhD of Dr. Mark Adams. Preliminary results were presented at the Vision Sciences Society (VSS) 2016 meeting in St Pete's beach, Florida (Scarfe and Glennerster (2016)). Feedback at VSS from Prof. Marc Ernst, Prof. Mike Landy, Prof. Jenny Read, Prof. Paul Hibbard, Dr. Loes van Dam, Prof. Andrew Glennerster, Dr. Katie Gray and many other attendees helped convince the author that the project was of worthwhile interest to others. Prof. Andrew Glennerster and Dr. Katie Gray have provided advice and continued support at all stages of the project. Prof. Andrew Glennerster has been a key sounding board for the project and played an integral role in deriving equations 33 through 38. Thus, prompting the author to derive equations 26 through 32.

Appendix A: Recording data from Ernst and Banks (2002)

Single cues sensitivities used for the simulations reported were estimated from Figure 3d of Ernst and Banks (2002). In order to gain estimates of the single cue sensitivities we viewed Figure 3d (as a pdf file) on a 4K computer monitor, so that the graph filled the majority of the screen. We then took the pixel coordinates of (1) the data points, (2) the minimum and maximum error bar position for each data-point and (3) the minimum and maximum values on the *x*- and *y*-axes. We were then able to compute the relative position of each data point (and error bar) in pixel coordinates on the *x*- and *y*-axis and covert these to the units shown in the graph by using the measured correspondence between pixel coordinates and axis units. Visual comparison of Figure 3 of the present paper and Figure 3d of Ernst and Banks shows that close correspondence achieved.

There was some inconsistently in Ernst and Banks (2002) as to how a "threshold" or "discrimination threshold" was defined. On page 430 the authors state, "The discrimination threshold is defined as the difference between the point of subjective equality (PSE) and the height of the comparison stimulus when it is judged taller than the standard stimulus 84% of the time". However, on page 431 the authors state "… T_H and T_V are the haptic and visual thresholds (84% points in Fig. 3a)". It is the first definition which is consistent with the mathematics i.e. the difference between the PSE and 84% point of the function being equal to the sigma of the fitted Cumulative Gaussian function.

Therefore, we cross checked our thresholds estimated from Figure 3d, with the thresholds calculated from the integrated cues functions in Figure 3b. Thresholds from Figure 3b were taken to be the difference between the point of subjective equality (PSE) and the 84% point on the function. When compared to the thresholds estimated from Figure 3d the difference in estimates was very small (average across data points of 0.23). We were therefore happy that definition of threshold was that of page 430 and that we had accurately estimated the thresholds and understood their relationship to the properties of the psychometric functions reported in the paper. Note: that for the purposes of the present paper all that was needed is an approximation of the exact values.

Appendix B: Example functions and goodness of fit

Figure A1a shows the true underlying functions for the minimum, maximum and base (middle) sigma values used in the current study as well as the stimulus levels at which the functions were sampled. As can be seen, consistent with Ernst and Banks (2002) Figure 3a, all functions straddle high and low performance levels needed for well fit functions (Wichmann & Hill, 2001a, 2001b). Figures A1b-e show examples of how these functions were sampled with our four sampling regimes (10, 25, 40 and 55 trials per stimulus level), with the maximum likelihood best fit functions and goodness of fit (see below) values shown in the legend. We have only shown these for just the $\Delta = 0$ case, as for all delta values used the sampling range was shifted so as to be centred on the true mean of the underlying function. As is clear, for all sampling regimes the data are well fit by the Cumulative Gaussian functions.

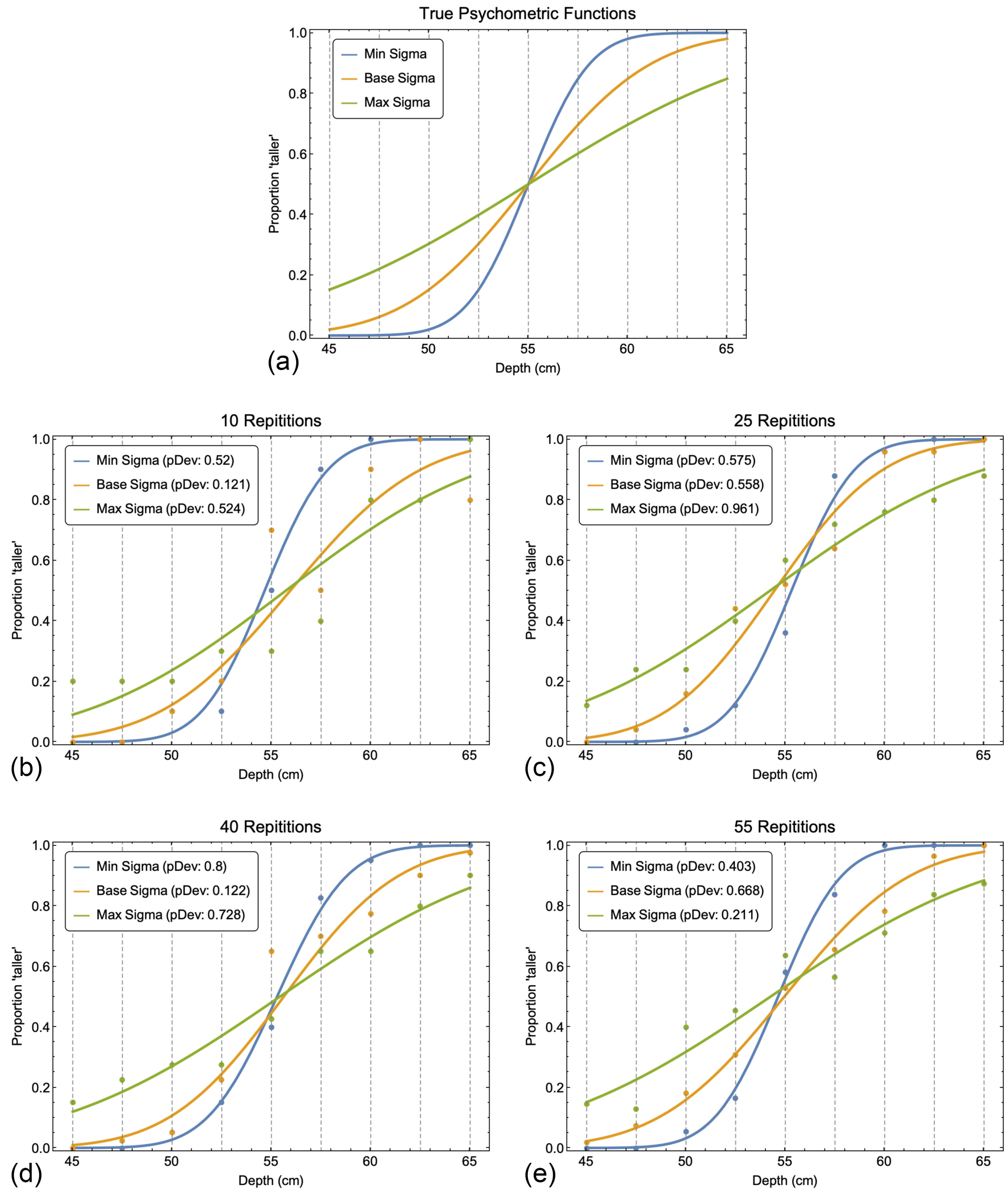

***Figure A1**: (a) shows the underlying "true" psychometric functions for the minimum, maximum and base (middle) sigma values used in the paper. The dashed vertical grey lines show the nine stimulus values at which these functions were sampled. (b) through (e) show examples of how the functions were sampled through simulation and fit with psychometric functions for the four data collection regimes used throughout the paper (10, 25, 40, and 55 repetitions per stimulus level). Inset in each graph is the goodness of fit value, pDev. This represents the probability with which the experimental data produced a higher likelihood ratio than that of the stimulated experiments based upon the target model. If this is greater than 0.05, the function is considered to fit the data well. See accompanying text for details.*

Within the cue integration literature, the goodness of fit of a function and the criteria upon which a fit is considered unacceptable is rarely if ever stated (for example Ernst & Banks, 2002; Helbig & Ernst, 2007; Hillis et al., 2002). Thus, it is impossible to tell if a goodness of fit test was performed, and if one was, which test which was used, and the criteria adopted for rejecting a fitted function. Given that the fit of data to the MWF model is normally assessed by eye, it is likely that this is also the case for the fit of individual psychometric functions (Kingdom & Prins, 2016). The Palamedes toolbox (Prins & Kingdom, 2009) used in the present study implements a bootstrapped likelihood ratio test to assess the goodness of fit of a psychometric function. The logic of the test is as follows (Kingdom & Prins, 2016).

As detailed in the main text, when fitting a psychometric function to some data the experimenter assumes: (1) the observer does not improve or degrade at the task they are performing over time, (2) each perceptual judgement an observer makes is statistically independent of all others, and (3) performance of the observer can be well characterised by the psychometric function that the experimenter is choosing to fit to the data. These assumptions combined can be referred to as the “target model”. The validity of the target model can be assessed by comparing it to a “saturated model” which only assumes (1) and (2). Thus, in the saturated model, the probability of response for one stimulus level is completely independent on the probability of response for any other stimulus level i.e. no psychometric function is assumed.

The target model is “nested” under the saturated model, as it is a single specific case of the saturated model. Thus, the likelihood associated with the fit of the target model can never produce a better fit than that of the less restrictive saturated model. For a given set of data one can calculate the likelihood ratio (likelihood of the target model / likelihood of the saturated model) which will, by definition, be less than or equal to 1. It will only be equal to one if the target and saturated models provide as good a fit as one another. The likelihood ratio test simulates a set of experiments through a bootstrap procedure and for each calculates the likelihood ratio. The probability with which the experimental data produces a higher likelihood ratio than that of the stimulated experiments is calculated (*pDev* in Figure A1). If this probability is less than 0.05% the goodness of fit is deemed poor. As with any *p*-value, the 0.05% cut-off is a completely arbitrary convention (Kingdom & Prins, 2016). Thus, some experimenters may adopt this and others not. This mirrors the open discussion about the use of *p*-values for general statistical analysis.

For the present study, it was computationally unfeasible to run a bootstrapped likelihood ratio test for each of the ~15.3 million simulated functions (even when using MATLAB’s Parallel processing toolbox to spread the computational load over the 8-Core Intel Core i9 available to the author this would have taken ~1-2 months of constant processing).

Nevertheless, we wanted to assess the extent to which the maximum likelihood fit functions would in general be considered well fit. Therefore, for the maximum and minimum cue sigma value used in the paper (i.e. shallowest and steepest psychometric functions), we simulated data for 1000 observers, fit Cumulative Gaussian psychometric functions to the data (as described in the main text) and assessed the goodness of fit using the bootstrapped likelihood ratio test (1000 bootstraps). We did this for our four sampling regimes: 10, 25, 40 and 55 trials per stimulus level.

Based upon the 0.05% criteria for a cut-off between well and poorly fit function (*pDev* in Figure A1), virtually all functions would have been classed as well fit, regardless of data collection regime of the slope of the underlying function (Table A1; overall average 94.95%). As would be expected, this was true for all Delta levels. This is because the sampling range was always centred on the true mean of the function, so the values for Delta 0, 3 and 6 in Table A1 are effectively replications of one another. This confirms across 24000 fitted functions what can be seen in the example functions of Figure A1 i.e. that the data are well fit by the psychometric functions. We can therefore be satisfied that the around 94.95% of all functions reported in the paper would have been classed as well fit based on this criteria. See also the criteria adopted for rejecting psychometric functions discussed in the main body of the text.

Sigma / Trials	**Delta 0**	**Delta 3**	**Delta 6**
Min 10	93.8%	95.5%	95.4%
Max 10	95.5%	95.2%	95.5%
Min 25	96.2%	94.5%	94.6%
Max 25	95.6%	96.7%	95.4%
Min 40	95.7%	95.5%	94.1%
Max 40	94.7%	95%	95.2%
Min 55	95.5%	94.1%	94.4%
Max 55	94.7%	93.7%	94.7%
Mean Value	**94.91%**	**95.03%**	**94.91%**

Table A1*: Shows the percentage of psychometric functions which would be classified as well fit based upon the bootstrapped likelihood ratio test described in the main text. The percentage of well fit functions is shown for the minimum and maximum sigma used in the simulations of the paper, and for each combination of trials per stimulus value on the psychometric function and cue conflict level (cue delta in mm).*

References

Adams, W. J., Banks, M. S., & van Ee, R. (2001). Adaptation to three-dimensional distortions in human vision. *Nat Neurosci, 4*(11), 1063-1064. doi:10.1038/nn729

Adams, W. J., & Mamassian, P. (2004). Bayesian combination of ambiguous shape cues. *J Vis, 4*(10), 921-929. doi:10.1167/4.10.7

Arnold, D. H., Petrie, K., Murray, C., & Johnston, A. (2019). Suboptimal human multisensory cue combination. *Sci Rep, 9*(1), 5155. doi:10.1038/s41598-018-37888-7

Backus, B. T. (2002). Perceptual metamers in stereoscopic vision. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in Neural Information Processing* (Vol. 14). Cambridge, MA: MIT Press.

Beierholm, U., Shams, L., Kording, K., & Ma, W. J. (2009). *Comparing Bayesian models for multisensory cue combination without mandatory fusion*. Paper presented at the Advances in Neural Information Processing Systems 20. Advances in Neural Information Processing Systems.

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychol Methods, 10*(4), 389-396. doi:10.1037/1082-989X.10.4.389

Blitzstein, J. K., & Hwang, J. (2015). *Introduction to Probability*. London: CRC Press.

Bowers, J. S., & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychol Bull, 138*(3), 389-414. doi:10.1037/a0026450

Bradshaw, M. F., Parton, A. D., & Glennerster, A. (2000). The task-dependent use of binocular disparity and motion parallax information. *Vision Research, 40*(27), 3725-3734.

Bulthoff, H. H., & Mallot, H. A. (1988). Integration of depth modules: stereo and shading. *J Opt Soc Am A, 5*(10), 1749-1758. doi:10.1364/josaa.5.001749

Burge, J., Fowlkes, C. C., & Banks, M. S. (2010). Natural-scene statistics predict how the figure-ground cue of convexity affects human depth perception. *Journal of Neuroscience, 30*(21), 7269-7280. doi:10.1523/JNEUROSCI.5551-09.2010

Burge, J., Girshick, A. R., & Banks, M. S. (2010). Visual-haptic adaptation is determined by relative reliability. *Journal of Neuroscience, 30*(22), 7714-7721. doi:10.1523/JNEUROSCI.6427-09.2010

Burge, J., Peterson, M. A., & Palmer, S. E. (2005). Ordinal configural cues combine with metric disparity in depth perception. *J Vis, 5*(6), 534-542. doi:10.1167/5.6.5

Byrne, P. A., & Henriques, D. Y. (2013). When more is less: increasing allocentric visual information can switch visual-proprioceptive combination from an optimal to sub-optimal process. *Neuropsychologia, 51*(1), 26-37. doi:10.1016/j.neuropsychologia.2012.10.008

Cochran, W. G. (1937). Problems arising in the analysis of a series of similar experiments. *Journal of the Royal Statistical Society, 4*, 102-118.

Cumming, G., Fidler, F., & Vaux, D. L. (2007). Error bars in experimental biology. *J Cell Biol, 177*(1), 7-11. doi:10.1083/jcb.200611141

de Winkel, K. N., Katliar, M., Diers, D., & Bulthoff, H. H. (2018). Causal Inference in the Perception of Verticality. *Sci Rep, 8*(1), 5483. doi:10.1038/s41598-018-23838-w

Descartes, R. (1641). *Meditations on First Philosophy: with Selections from the Objections and Replies*. Cambridge: Cambridge University Press.

Domini, F., & Caudek, C. (2009). The intrinsic constraint model and Fechnerian sensory scaling. *J Vis, 9*(2), 25 21-15. doi:10.1167/9.2.25

Ernst, M. O. (2006). A Bayesian view on multimodal cue integration. In G. Knoblich, I. M. Thornton, M. Grosjean, & M. Shiffrar (Eds.), *Human body perception from the inside out* (pp. 105-131). New York: Oxford University Press.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415*(6870), 429-433. doi:10.1038/415429a

Ernst, M. O., & Bulthoff, H. H. (2004). Merging the senses into a robust percept. *Trends Cogn Sci, 8*(4), 162-169. doi:10.1016/j.tics.2004.02.002

Ernst, M. O., & Di Luca, M. (2011). Multisensory perception: from integration to remapping. In J. Trommershauser, K. P. Körding, & M. S. Landy (Eds.), *Sensory Cue Integration* (pp. 224-250). Oxford: Oxford University Press.

Fischer, J., & Whitney, D. (2014). Serial dependence in visual perception. *Nat Neurosci, 17*(5), 738-743. doi:10.1038/nn.3689

Frund, I., Haenel, N. V., & Wichmann, F. A. (2011). Inference for psychometric functions in the presence of nonstationary behavior. *J Vis, 11*(6). doi:10.1167/11.6.16

Geisler, W. S. (2011). Contributions of ideal observer theory to vision research. *Vision Res, 51*(7), 771-781. doi:10.1016/j.visres.2010.09.027

Gepshtein, S., Burge, J., Ernst, M. O., & Banks, M. S. (2005). The combination of vision and touch depends on spatial proximity. *J Vis, 5*(11), 1013-1023. doi:10.1167/5.11.7

Girshick, A. R., & Banks, M. S. (2009). Probabilistic combination of slant information: weighted averaging and robustness as optimal percepts. *J Vis, 9*(9), 8 1-20. doi:10.1167/9.9.8

Glennerster, A., Tcheang, L., Gilson, S. J., Fitzgibbon, A. W., & Parker, A. J. (2006). Humans ignore motion and stereo cues in favor of a fictional stable world. *Curr Biol, 16*(4), 428-432. doi:10.1016/j.cub.2006.01.019

Green, D. M., & Swets, J. A. (1974). *Signal Detection Theory and Psychophysics*. Cambridge: Cambridge University Press.

Helbig, H. B., & Ernst, M. O. (2007). Optimal integration of shape information from vision and touch. *Exp Brain Res, 179*(4), 595-606. doi:10.1007/s00221-006-0814-y

Held, R. T., Cooper, E. A., & Banks, M. S. (2012). Blur and disparity are complementary cues to depth. *Curr Biol, 22*(5), 426-431. doi:10.1016/j.cub.2012.01.033

Henriques, D. Y., & Cressman, E. K. (2012). Visuomotor adaptation and proprioceptive recalibration. *J Mot Behav, 44*(6), 435-444. doi:10.1080/00222895.2012.659232

Hershenson, M. H. (1999). *Visual space perception: a primer*. London: MIT Press.

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*(5598), 1627-1630. Retrieved from <Go to ISI>://000179361600051

Hillis, J. M., Watt, S. J., Landy, M. S., & Banks, M. S. (2004). Slant from texture and disparity cues: optimal cue combination. *J Vis, 4*(12), 967-992. doi:10.1167/4.12.1

Ho, Y. X., Landy, M. S., & Maloney, L. T. (2006). How direction of illumination affects visually perceived surface roughness. *J Vis, 6*(5), 634-648. doi:10.1167/6.5.8

Howard, I. P., & Rogers, B. J. (2002). *Seeing in Depth: Depth Perception* (Vol. 2). Toronto: I Porteous.

Johnston, E. B., Cumming, B. G., & Landy, M. S. (1994). Integration of stereopsis and motion shape cues. *Vision Res, 34*(17), 2259-2275. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/7941420

Johnston, E. B., Cumming, B. G., & Parker, A. J. (1993). Integration of depth modules: stereopsis and texture. *Vision Res, 33*(5-6), 813-826. doi:Doi 10.1016/0042-6989(93)90200-G

Jones, P. R. (2016). A tutorial on cue combination and Signal Detection Theory: Using changes in sensitivity to evaluate how observers integrate sensory information. *Journal of Mathematical Psychology, 73*, 117-139.

Julesz, B. (1971). *Foundations of Cyclopean Perception*. Chicago: Chicago University Press.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annu Rev Psychol, 55*, 271-304. doi:10.1146/annurev.psych.55.090902.142005

Kingdom, F. A. A., & Prins, N. (2010). *Psychophysics: A Practical Introduction.* (1st ed.). London: Academic Press.

Kingdom, F. A. A., & Prins, N. (2016). *Psychophysics: A Practical Introduction.* (2nd ed.). London: Academic Press.

Kiyonaga, A., Scimeca, J. M., Bliss, D. P., & Whitney, D. (2017). Serial Dependence across Perception, Attention, and Memory. *Trends Cogn Sci, 21*(7), 493-497. doi:10.1016/j.tics.2017.04.011

Knill, D. C. (1998a). Ideal observer perturbation analysis reveals human strategies for inferring surface orientation from texture. *Vision Res, 38*(17), 2635-2656. doi:10.1016/s0042-6989(97)00415-x

Knill, D. C. (1998b). Surface orientation from texture: ideal observers, generic observers and the information content of texture cues. *Vision Res, 38*(11), 1655-1682. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/9747502

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci, 27*(12), 712-719. doi:10.1016/j.tins.2004.10.007

Knill, D. C., & Richards, W. (1996). *Perception as Bayesian Inference*. Cambridge: Cambridge University Press.

Knill, D. C., & Saunders, J. A. (2003). Do humans optimally integrate stereo and texture information for judgments of surface slant? *Vision Research, 43*(24), 2539-2558. Retrieved from <Go to ISI>://000185577700006

Koenderink, J. J., van Doorn, A. J., Kappers, A. M., & Lappin, J. S. (2002). Large-scale visual frontoparallels under full-cue conditions. *Perception, 31*(12), 1467-1475. doi:10.1068/p3295

Koenderink, J. J., van Doorn, A. J., Kappers, A. M., & Todd, J. T. (2002). Pappus in optical space. *Percept Psychophys, 64*(3), 380-391. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/12049279

Koenderink, J. J., van Doorn, A. J., & Lappin, J. S. (2000). Direct measurement of the curvature of visual space. *Perception, 29*(1), 69-79. doi:10.1068/p2921

Kontsevich, L. L., & Tyler, C. W. (1999). Bayesian adaptive estimation of psychometric slope and threshold. *Vision Res, 39*(16), 2729-2737. doi:Doi 10.1016/S0042-6989(98)00285-5

Kording. (2007). Decision theory: what "should" the nervous system do? *Science, 318*(5850), 606-610. doi:10.1126/science.1142998

Kording, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal inference in multisensory perception. *PloS one, 2*(9), e943. doi:10.1371/journal.pone.0000943

Kuss, M., Jakel, F., & Wichmann, F. A. (2005). Bayesian inference for psychometric functions. *J Vis, 5*(5), 478-492. doi:10.1167/5.5.8

Lages, M., & Jaworska, K. (2012). How Predictable are "Spontaneous Decisions" and "Hidden Intentions"? Comparing Classification Results Based on Previous Responses with

Multivariate Pattern Analysis of fMRI BOLD Signals. *Front Psychol, 3*, 56. doi:10.3389/fpsyg.2012.00056

Landy, M. S., Banks, M. S., & Knill, D. C. (2011). Ideal-observer models of cue integration. In J. Trommershauser, K. P. Körding, & M. S. Landy (Eds.), *Sensory Cue Integration* (pp. 5-29). Oxford: Oxford University Press.

Landy, M. S., Maloney, L. T., Johnston, E. B., & Young, M. (1995). Measurement and modeling of depth cue combination: in defense of weak fusion. *Vision Res, 35*(3), 389-412. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/7892735

Leek, M. R. (2001). Adaptive procedures in psychophysical research. *Percept Psychophys, 63*(8), 1279-1292. doi:10.3758/bf03194543

Liberman, A., Fischer, J., & Whitney, D. (2014). Serial dependence in the perception of faces. *Curr Biol, 24*(21), 2569-2574. doi:10.1016/j.cub.2014.09.025

Liberman, A., Manassi, M., & Whitney, D. (2018). Serial dependence promotes the stability of perceived emotional expression depending on face similarity. *Atten Percept Psychophys, 80*(6), 1461-1473. doi:10.3758/s13414-018-1533-8

Liberman, A., Zhang, K., & Whitney, D. (2016). Serial dependence promotes object stability during occlusion. *J Vis, 16*(15), 16. doi:10.1167/16.15.16

Lipton, P. (2005). Testing hypotheses: prediction and prejudice. *Science, 307*(5707), 219-221. doi:10.1126/science.1103024

Lovell, P. G., Bloj, M., & Harris, J. M. (2012). Optimal integration of shading and binocular disparity for depth perception. *J Vis, 12*(1). doi:10.1167/12.1.1

Maloney, L. T., & Landy, M. S. (1989). *A statistical framework for robust fusion of depth information.* Paper presented at the Proc. SPIE 1199, Visual Communications and Image Processing IV, Philadelphia, PA, United States.

Mamassian, P., Landy, M. S., & Maloney, L. T. (2002). Bayesian Modelling of Visual Perception. In R. P. N. Rao, B. A. Olshausen, & M. S. Lewicki (Eds.), *Probabilistic Models of the Brain: Perception and Neural Function* (pp. 13-36).

McLaughlin, S. C., & Webster, R. G. (1967). Changes in the straight-ahead eye position during adaptation to wedge prisms. *Percept Psychophysics, 2*(1), 37-44.

Murphy, A. P., Ban, H., & Welchman, A. E. (2013). Integration of texture and disparity cues to surface slant in dorsal visual cortex. *J Neurophysiol, 110*(1), 190-203. doi:10.1152/jn.01055.2012

Murray, R. F., & Morgenstern, Y. (2010). Cue combination on the circle and the sphere. *J Vis, 10*(11), 15. doi:10.1167/10.11.15

Nardini, M., Jones, P., Bedford, R., & Braddick, O. (2008). Development of cue integration in human navigation. *Curr Biol, 18*(9), 689-693. doi:10.1016/j.cub.2008.04.021

Negen, J., Wen, L., Thaler, L., & Nardini, M. (2018). Bayes-Like Integration of a New Sensory Skill with Vision. *Sci Rep, 8*(1), 16880. doi:10.1038/s41598-018-35046-7

Oruc, I., Maloney, L. T., & Landy, M. S. (2003). Weighted linear cue combination with possibly correlated error. *Vision Res, 43*(23), 2451-2468. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/12972395

Pastore, M., & Calcagni, A. (2019). Measuring Distribution Similarities Between Samples: A Distribution-Free Overlapping Index. *Front Psychol, 10*, 1089. doi:10.3389/fpsyg.2019.01089

Pentland, A. (1980). Maximum likelihood estimation: the best PEST. *Percept Psychophys, 28*(4), 377-379. doi:10.3758/bf03204398

Prins, N. (2012). The psychometric function: the lapse rate revisited. *J Vis, 12*(6). doi:10.1167/12.6.25

Prins, N. (2013). The psi-marginal adaptive method: How to give nuisance parameters the attention they deserve (no more, no less). *J Vis, 13*(7), 3. doi:10.1167/13.7.3

Prins, N., & Kingdom, F. A. A. (2009). *Palamedes: Matlab routines for analyzing psychophysical data.* http://www.palamedestoolbox.org.

Prins, N., & Kingdom, F. A. A. (2018). Applying the Model-Comparison Approach to Test Specific Research Hypotheses in Psychophysical Research Using the Palamedes Toolbox. *Front Psychol, 9*, 1250. doi:10.3389/fpsyg.2018.01250

Rogers, B. J., & Bradshaw, M. F. (1993). Vertical disparities, differential perspective and binocular stereopsis. *Nature, 361*(6409), 253-255. doi:10.1038/361253a0

Rohde, M., van Dam, L. C. J., & Ernst, M. (2016). Statistically Optimal Multisensory Cue Integration: A Practical Tutorial. *Multisens Res, 29*(4-5), 279-317. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/29384605

Saunders, J. A., & Chen, Z. (2015). Perceptual biases and cue weighting in perception of 3D slant from texture and stereo information. *J Vis, 15*(2). doi:10.1167/15.2.14

Scarfe, P., & Glennerster, A. (2014). Humans use predictive kinematic models to calibrate visual cues to three-dimensional surface slant. *Journal of Neuroscience, 34*(31), 10394-10401. doi:10.1523/JNEUROSCI.1000-14.2014

Scarfe, P., & Hibbard, P. B. (2011). Statistically optimal integration of biased sensory estimates. *J Vis, 11*(7). doi:10.1167/11.7.12

Scarfe, P., & Hibbard, P. B. (2013). Reverse correlation reveals how observers sample visual information when estimating three-dimensional shape. *Vision Res, 86*, 115-127. doi:10.1016/j.visres.2013.04.016

Schrater, P. R., & Kersten, D. (2000). How optimal depth cue integration depends on the task. *International Journal of Computer Vision, 40*(1), 73-91. Retrieved from <Go to ISI>://WOS:000165942300005

Schutt, H. H., Harmeling, S., Macke, J. H., & Wichmann, F. A. (2016). Painfree and accurate Bayesian estimation of psychometric functions for (potentially) overdispersed data. *Vision Res, 122*, 105-123. doi:10.1016/j.visres.2016.02.002

Serwe, S., Drewing, K., & Trommershauser, J. (2009). Combination of noisy directional visual and proprioceptive information. *J Vis, 9*(5), 28 21-14. doi:10.1167/9.5.28

Smeets, J. B., van den Dobbelsteen, J. J., de Grave, D. D., van Beers, R. J., & Brenner, E. (2006). Sensory integration does not lead to sensory calibration. *Proc Natl Acad Sci U S A, 103*(49), 18781-18786. doi:10.1073/pnas.0607687103

Svarverud, E., Gilson, S. J., & Glennerster, A. (2010). Cue combination for 3D location judgements. *J Vis, 10*(1), 5 1-13. doi:10.1167/10.1.5

Takahashi, C., Diedrichsen, J., & Watt, S. J. (2009). Integration of vision and haptics during tool use. *J Vis, 9*(6), 3 1-13. doi:10.1167/9.6.3

Tassinari, H., & Domini, F. (2008). The intrinsic constraint model for stereo-motion integration. *Perception, 37*(1), 79-95. doi:10.1068/p5501

Todd, J. T. (2015). Can a Bayesian analysis account for systematic errors in judgments of 3D shape from texture? A reply to Saunders and Chen. *J Vis, 15*(9), 22. doi:10.1167/15.9.22

Todd, J. T., Christensen, J. C., & Guckes, K. M. (2010). Are discrimination thresholds a valid measure of variance for judgments of slant from texture? *J Vis, 10*(2), 20 21-18. doi:10.1167/10.2.20

Todd, J. T., & Thaler, L. (2010). The perception of 3D shape from texture based on directional width gradients. *J Vis, 10*(5), 17. doi:10.1167/10.5.17

Tresilian, J. R., & Mon-Williams, M. (2000). Getting the measure of vergence weight in nearness perception. *Exp Brain Res, 132*(3), 362-368. doi:10.1007/s002210000333

Tresilian, J. R., Mon-Williams, M., & Kelly, B. M. (1999). Increasing confidence in vergence as a cue to distance. *Proceedings of the Royal Society of London Series B-Biological Sciences, 266*(1414), 39-44. Retrieved from <Go to ISI>://000078154000005

Trommershauser, J., Körding, K. P., & Landy, M. S. (2011). *Sensory Cue Integration*. Oxford: Oxford University Press.

Vishwanath, D. (2012). The utility of defocus blur in binocular depth perception. *Iperception, 3*(8), 541-546. doi:10.1068/i0544ic

Wagner, M. (1985). The metric of visual space. *Percept Psychophys, 38*(6), 483-495. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/3834394

Watson, A. B. (2017). QUEST+: A general multidimensional Bayesian adaptive psychometric method. *J Vis, 17*(3), 10. doi:10.1167/17.3.10

Watson, A. B., & Pelli, D. G. (1983). QUEST: a Bayesian adaptive psychometric method. *Percept Psychophys, 33*(2), 113-120. doi:10.3758/bf03202828

Watt, S. J., Akeley, K., Ernst, M. O., & Banks, M. S. (2005). Focus cues affect perceived depth. *J Vis, 5*(10), 834-862. doi:10.1167/5.10.7

Welch, R. B., Bridgeman, B., Anand, S., & Browman, K. E. (1993). Alternating prism exposure causes dual adaptation and generalization to a novel displacement. *Percept Psychophys, 54*(2), 195-204. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/8361835

Wichmann, F. A., & Hill, N. J. (2001a). The psychometric function: I. Fitting, sampling, and goodness of fit. *Percept Psychophys, 63*(8), 1293-1313. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/11800458

Wichmann, F. A., & Hill, N. J. (2001b). The psychometric function: II. Bootstrap-based confidence intervals and sampling. *Percept Psychophys, 63*(8), 1314-1329. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/11800459

Xia, Y., Leib, A. Y., & Whitney, D. (2016). Serial dependence in the perception of attractiveness. *J Vis, 16*(15), 28. doi:10.1167/16.15.28

Young, M. J., Landy, M. S., & Maloney, L. T. (1993). A perturbation analysis of depth perception from combinations of texture and motion cues. *Vision Res, 33*(18), 2685-2696. doi:Doi 10.1016/0042-6989(93)90228-O

Zabulis, X., & Backus, B. T. (2004). Starry night: a texture devoid of depth cues. *J Opt Soc Am A Opt Image Sci Vis, 21*(11), 2049-2060. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/15535362

Zychaluk, K., & Foster, D. H. (2009). Model-free estimation of the psychometric function. *Atten Percept Psychophys, 71*(6), 1414-1425. doi:10.3758/APP.71.6.1414