Comparing sensitivity estimates from MLDS and forced-choice methods in a slant-from-texture experiment

Guillermo Aguilar
Modelling of Cognitive Processes Group, Department of Software Engineering and Theoretical Computer Science, Technische Universität Berlin, Berlin, Germany
Bernstein Center for Computational Neuroscience, Berlin, Germany

Felix A. Wichmann
AG Neuronale Informationsverarbeitung, Mathematisch-Naturwissenschaftliche Fakultät, Eberhard Karls Universität, Tübingen, Germany
Bernstein Center for Computational Neuroscience, Tübingen, Germany
Max-Planck-Institut für Intelligente Systeme, Tübingen, Germany

Marianne Maertens
Modelling of Cognitive Processes Group, Department of Software Engineering and Theoretical Computer Science, Technische Universität Berlin, Berlin, Germany

Maximum likelihood difference scaling (MLDS) is a method for the estimation of perceptual scales based on the judgment of differences in stimulus appearance (Maloney & Yang, 2003). MLDS has recently also been used to estimate near-threshold discrimination performance (Devinck & Knoblauch, 2012). Using MLDS as a psychophysical method for sensitivity estimation is potentially appealing, because MLDS has been reported to need less data than forced-choice procedures, and particularly naive observers report to prefer suprathreshold comparisons to JND-style threshold tasks. Here we compare two methods, MLDS and two-interval forced-choice (2-IFC), regarding their capability to estimate sensitivity assuming an underlying signal-detection model. We first examined the theoretical equivalence between both methods using simulations. We found that they disagreed in their estimation only when sensitivity was low, or when one of the assumptions on which MLDS is based was violated. Furthermore, we found that the confidence intervals derived from MLDS had a low coverage; i.e., they were too narrow, underestimating the true variability. Subsequently we compared MLDS and 2-IFC empirically using a slant-from-texture task. The amount of agreement between sensitivity estimates from the two methods varied substantially across observers. We discuss possible reasons for the observed disagreements, most notably violations of the MLDS model assumptions. We conclude that in the present example MLDS and 2-IFC could equally be used to estimate sensitivity to differences in slant, with MLDS having the benefit of being more efficient and more pleasant, but having the disadvantage of unsatisfying coverage.

Introduction

Maximum likelihood difference scaling (MLDS) is a psychophysical method that allows the efficient characterization of perceptual scales (Knoblauch & Maloney, 2012; Maloney & Yang, 2003). Observers are asked to judge appearance differences for suprathreshold stimuli that vary along some dimension of interest, and a scale is constructed based on the reported differences in appearance. The method has been used to study appearance in a variety of visual domains such as color differences (Maloney & Yang,

Citation: Aguilar, G., Wichmann, F. A., & Maertens, M. (2017). Comparing sensitivity estimates from MLDS and forced-choice methods in a slant-from-texture experiment. *Journal of Vision*, *17*(1):37, 1–18, doi:10.1167/17.1.37.

doi: 10.1167/17.1.37 Received August 5, 2016; published January XX, 2017 ISSN 1534-7362

2003), texture properties (Emrith, Chantler, Green, Maloney, & Clarke, 2010), surface glossiness (Obein, Knoblauch, & Viènot, 2004), transparency (Fleming, Jäkel, & Maloney, 2011) and material properties (Paulun, Kawabe, Nishida, & Fleming, 2015), as well as for the assessment of perceived image quality in compression-degraded images (Charrier, Maloney, Cherifi, & Knoblauch, 2007).

Recently, MLDS has been used to link stimulus appearance with stimulus discriminability. Assuming an underlying signal detection model, Devinck and Knoblauch (2012) have demonstrated a quantitative agreement between sensitivity estimates derived from perceptual scales (MLDS) and sensitivity estimates assessed with a traditional forced-choice procedure for the watercolor effect. Their finding is remarkable given the long effort in psychophysical research of relating discrimination and appearance in a unified framework.

Relating stimulus appearance—the stimulus subjective magnitude—to discrimination—the ability to discriminate stimuli—dates back to the roots of psychophysical research. Fechner (1860) proposed that by summing equal subjective, just-noticeable differences (JND) and assuming Weber's law, a function could be constructed which relates stimulus subjective magnitude and physical magnitude (Baird, 1978). Soon Fechner's suggestion was criticized, theoretically as well as for lack of experimental evidence to support it (reviewed in detail in Krueger, 1989).

Stevens (1957, 1975) later proposed that subjective magnitude could be directly measured from observer responses to suprathreshold stimuli. He devised direct methods to measure subjective magnitude and derived (power) functions that would relate subjective and physical magnitude (Gescheider, 1997; Stevens, 1975; but c.f., Treisman, 1964a, 1964b). However, Steven's proposal was met with equal criticism, partly because of the scale's lack of predictive power for discriminability and partly because of the methodological concerns of asking observers to numerically estimate or provide ratings of perceived sensation (e.g., Baird, 1989). Although a considerable amount of work has been done trying to unify discrimination and appearance, so far the debate still continues and mixed experimental evidence has been found (e.g., Hillis & Brainard, 2007; Krueger, 1989; Ross, 1997). Thus, the finding of Devinck and Knoblauch (2012) that appearance and sensitivity can be linked via MLDS is promising, because it suggests that a suprathreshold method like MLDS could be used to predict sensitivity to near-threshold stimulus differences. Apart from potential theoretical implications, Devinck and Knoblauch's finding may be beneficial from a purely methodological point of view, because MLDS requires a considerably smaller amount of data than traditional discrimination methods. Because of its efficiency

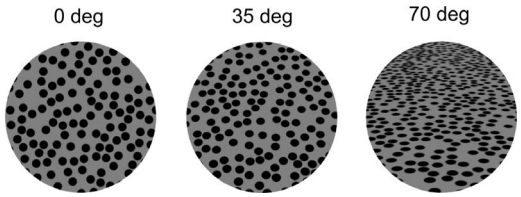

Figure 1. Example stimuli showing surfaces of different slants covered with the "polka dots" texture. Here we used the method of triads for MLDS where observers judge which of the pairs exhibit a larger difference in perceived slant, the left-middle pair or the right-middle pair. Most observers would report that the right-middle pair (35, 70) contains the larger slant difference, although the physical slant difference is identical between the pairs: (0, 35) versus (35, 70).

MLDS could be used to identify experimental settings in which appearance and discrimination judgments are consistent, by comparing sensitivity measured in discrimination tasks (e.g., two-interval forced-choice) with sensitivity derived from MLDS. The goal of this work was to further explore—theoretically and empirically—the possibility to use MLDS to predict near-threshold discrimination performance using a slant-from-texture task.

Slant-from-texture tasks

We measure perceptual scales in a slant-from-texture experiment. The perceptual scale that relates apparent and physical slant in slant-from-texture tasks has a nonlinear shape and it therefore provides an interesting test case for predicting sensitivity at different positions of the MLDS based scale. Slant-from-texture stimuli have been used extensively in the study of depth and surface perception (e.g., Knill, 1998; Rosas, Wichmann, & Wagemans, 2004; Saunders & Backus, 2006; Todd, Thaler, & Dijkstra, 2005; Velisavljević & Elder, 2006), because texture gradients can evoke a strong impression of 3-D slant in the absence of other cues (Saunders, 2003). Stimuli are surfaces that are covered with a texture pattern such as randomly placed circular elements (or "polka dots"; Figure 1). The surface is slanted at varying degrees relative to the fronto-parallel position resulting in characteristic changes in the polka dot patterns. The slanted texture is viewed through an aperture to isolate texture cues from other pictorial cues such as the shape and borders of the surface (Knill, 1998; Todd, Christensen, & Guckes, 2010). Using this type of stimuli, it has been found that sensitivity to slant is lower when the surface is close to the fronto-parallel position than when the surface is slanted away from it (Knill, 1998; Rosas et al., 2004). The difference in sensitivity between 0° and 70° can be up to ten-fold (Knill, 1998).

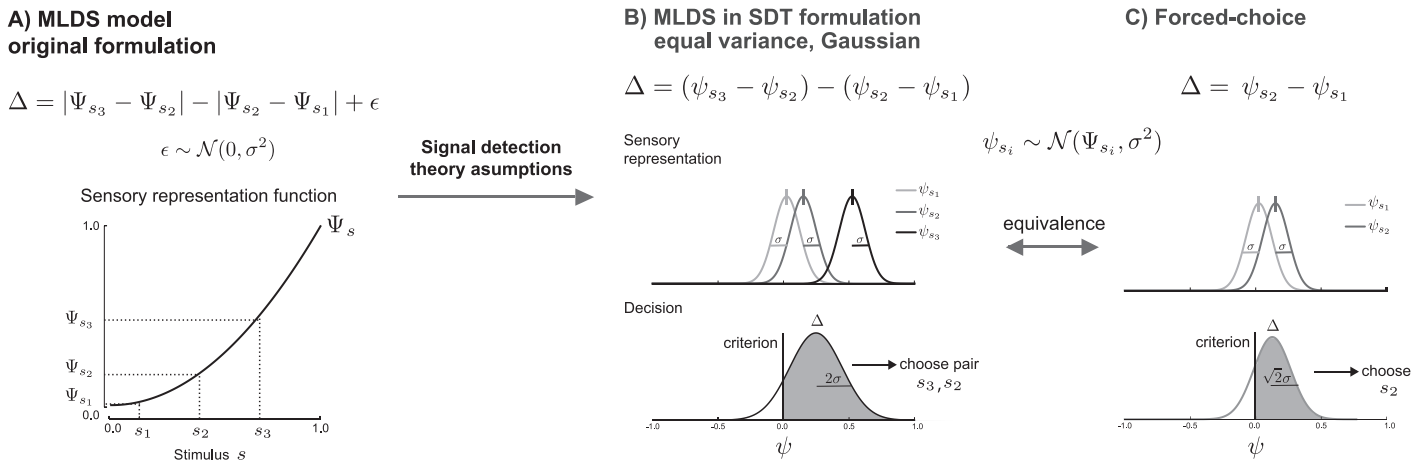

Figure 2. MLDS in the signal detection framework. (A) In its original formulation the decision variable (Δ) in MLDS is defined as the difference between intervals ($|\Psi_{s_3}, \Psi_{s_2}|$and $|\Psi_{s_2}, \Psi_{s_1}|$), and this difference is corrupted by Gaussian noise (ε). (B) In the signal detection formulation of MLDS, the noise originates only from the sensory representations (ψ_s) which are assumed to be independent Gaussian random variables with equal variance. In the signal detection version of MLDS, the model is equivalent to forced-choice methods (C) at the level of the sensory representation. See text for details.

MLDS and the signal detection model

The decision model underlying the MLDS framework is depicted in Figure 2A. It is assumed that different stimulus levels s_i are associated with discrete perceptual responses Ψ_{s_i}, and that observers compare different stimuli by judging the differences between the perceptual responses. The decision variable is assumed to be corrupted by decision noise, ε, which is assumed to be Gaussian distributed with zero mean and variance σ^2. MLDS estimates the perceptual scale together with the noise associated with the judgments (Knoblauch & Maloney, 2008; Maloney & Yang, 2003).

The same perceptual process can be rephrased in a signal detection framework by shifting the noise from the decision process to the sensory representation (see Figure 2B). In this way, the original MLDS scale can be transformed to a normed scale in which the units on the perceptual axis represent differences in units of d'. This transformation has been suggested by Devinck and Knoblauch (2012) to compare supra- and near-threshold judgments in the watercolor effect. A detailed description of the transformation and the MLDS model is provided in the Appendix.

In order to apply this transformation, the following assumptions are made: (a) The sensory representations associated with each stimulus level are Gaussian random variables with equal variance (σ^2). (b) They are independent. (c) The decision process is deterministic. (d) The sensory representation function is monotonically increasing. This produces only positive values of sensory response intervals so that the absolute value operation can be removed from the decision rule (Δ variable in Figure 2A and B). An MLDS decision model with the above assumptions is equivalent to a signal detection model with equal-variance and Gaussian distributed sensory representations, as depicted in Figure 2C.[1]

Objectives

We want to test whether and to what extent we can assume the equivalence of MLDS and forced-choice procedures for estimating sensitivity as it was reported by Devinck and Knoblauch (2012) for the Watercolor effect. We first examine the theoretical equivalence between both methods by means of simulations. We use a known observer model to generate sensitivity estimates for both methods. In the present analysis we evaluate the adequacy of MLDS to predict sensitivity using the 2-AFC method as the standard of reference as the latter has proven its usefulness in the estimation of sensitivity over time. We quantify the amount of agreement between the two methods in the presence of different violations in the assumptions underlying MLDS. We then test the empirical consistency between sensitivity estimates derived with MLDS and forced-choice procedures in two experiments with a slant-from-texture task. In Experiment 1 observers judge suprathreshold slant differences and perceptual scales are derived from the judgments using MLDS. From these scales we derive sensitivity estimates (thresholds) at different slant levels. In Experiment 2 observers judge near-threshold slant differences in a two-interval forced-choice (2-IFC) task. Sensitivity estimates (thresholds) are derived from psychometric functions for the same slant levels as in Experiment 1.

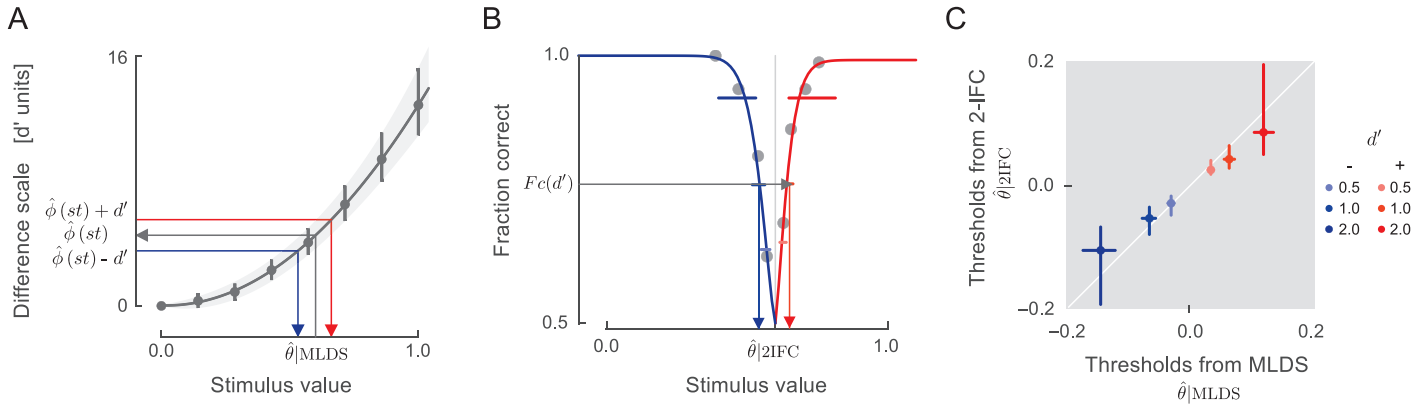

Figure 3. Comparison of MLDS and forced-choice thresholds. (A) Difference scale for a simulated MLDS experiment using the sensory function depicted in Figure 2A. A cubic spline (dark solid line) is fitted to the scale values (circles). The procedure to read out thresholds is illustrated by arrows. Here we read out the threshold ($\hat{\theta}|_{\text{MLDS}}$) for a standard $st = 0.6$ (vertical gray line) at a performance level of $d' = 1$ for comparisons above (red arrow) and below (blue arrow) the standard. B. Psychometric functions from a simulated 2-IFC procedure with the same sensory function for comparisons above (red) and below (blue) the standard (vertical gray line). Thresholds ($\hat{\theta}|_{\text{2IFC}}$) were derived from the fraction correct corresponding to a performance level $d' = 1$ ($Fc = 0.76$). (C) Thresholds derived with each method are plotted against each other. They are expressed relative to the standard ($st = 0.6$) for comparisons above (red colors) and below (blue colors) the standard. The main diagonal white line indicates identity. Error bars indicate 95% C.I.

To anticipate, the amount of agreement between sensitivity estimates from the two methods varied substantially across observers. The simulations showed that disagreement between the methods might be due to violations of the model assumptions underlying MLDS.

Simulations

The sensory representation was modelled as a power function, $\Psi(s) = s^e$, with exponent $e = 2.0$ (Figure 2A). We used an exponent greater than one so that sensitivity would increase with stimulus intensity, which is the case for slant-from-texture (Knill, 1998). The sensory representation function was used to simulate responses of a model observer for the MLDS and the 2-IFC procedure. It was assumed to be a Gaussian random variable with the mean corresponding to $\Psi(s)$ and unique variance σ^2 (Figure 2B through C). An example simulation is depicted in Figure 3. Thresholds were derived for a standard value of $st = 0.6$ from MLDS scales (panel A) and from psychometric functions in a 2-IFC task (panel B).

MLDS thresholds

We performed the MLDS experiment with the method of triads (Knoblauch & Maloney, 2012; Maloney & Yang, 2003). A triad consists of three stimuli, s_1, s_2, and s_3. To simulate a triad, the generative model (Figure 2A) assigns perceptual responses, Ψ_i, to each of the three stimuli, s_i. The simulated observer decides which of the pairs, (s_1, s_2) or (s_2, s_3), contains the bigger difference in perceived slant according to the decision model depicted in Figure 2B.

MLDS data (simulated and observed) were analyzed with the R package *MLDS*, available in CRAN (Knoblauch & Maloney, 2008) and with python routines based on *numpy* and *scipy* libraries. A python wrapper of the *MLDS* routines together with all subsequent analysis routines is available online (http://github.com/TUBvision/mlds).

We first estimated a perceptual scale from the simulated responses by employing the standard MLDS routines available in R (Knoblauch & Maloney, 2008; see Appendix, MLDS with the method of triads section for a detailed description of the estimation procedure). We then derived sensitivity estimates from the perceptual scale following the procedure suggested by Devinck and Knoblauch (2012). To do this we reparametrized the original unconstrained scale so that the scale values are expressed in units of d'. The details underlying the reparametrization are explained in Appendix, Difference scales and signal detection theory. In the simulation we derived sensitivity estimates for eight standard values (experiments were done with four standard values). Due to the nonlinear shape of the perceptual scale, the local slopes differed between different standard values and hence translated into different sensitivity levels along the stimulus dimension. For each standard we determined sensitivity at three performance levels ($d' = 0.5$, 1, and 2) above and below the standard. To derive the stimulus values that corresponded to each d' difference for a given standard, we interpolated between the sampled data

points with a cubic spline fit ($\hat{\phi}(s)$,shown as solid dark gray line in Figure 3A). The scale value, $\hat{\phi}(st)$in d' units, that corresponds to a particular standard stimulus (st) and performance level (d') was read from the fitted function. The readout can be described by

$$\hat{\phi}^{-1}\left(\hat{\phi}(st)\pm d'\right) = \hat{\theta}^{st}_{\pm d'}\Big|\text{MLDS} \quad (1)$$

in which the + (–) sign next to d' stands for comparison values above (below) the standard, and $\hat{\theta}^{st}_{\pm d'}|$MLDS stands for a particular sensitivity value in stimulus units as estimated by MLDS.

2-IFC thresholds

The same generative model is used to simulate responses in the 2-IFC procedure. In each trial, one response is generated for the standard and one for the comparison value. Perceptual responses are compared according to the decision model depicted in Figure 2C. We simulated the same number of trials that we ran in the behavioral experiments (see Experiments, both Observers and Procedure experiment 1: MLDS sections).

To allow the comparison of thresholds across different standard slants, we report comparison values in terms of differences relative to each standard. We fitted separate psychometric functions for positive and negative comparison values (smaller and larger than the standard). Psychometric functions were Weibull functions (F) with the guess rate (γ) set to 50% chance level. The lapse rate (λ), slope, and position parameters of the psychometric function were estimated using Bayes inference (Kuss, Jäkel, & Wichmann, 2005). We used the *psignifit4* implementation (Schütt, Harmeling, Macke, & Wichmann, 2016) for function fitting, estimation of confidence intervals, and analysis of goodness of fit. Each psychometric function was estimated from a total of 320 trials (4 comparison values × 80 repeats) as in the experiments.

An example psychometric function for one standard slant is shown in Figure 3B. Performance thresholds were obtained from each psychometric function by finding the stimulus value that produces a percentage correct corresponding to a desired d'. Assuming the equal variance Gaussian case of a signal detection model (Green & Swets, 1966), d' can be converted to percentage correct and vice versa, and the threshold can be read out by

$$\hat{F}^{-1}(Fc') = \hat{\theta}^{st}_{\pm d'}|\text{2IFC}$$

where + (–) indicate comparisons above (below) the standard, $Fc' = \{0.28, 0.52, 0.84\}$ are the unscaled fractions correct (range between 0 and 1) that correspond to the raw fractions correct $Fc = \{0.64, 0.76, 0.92\}$ (range between 0.5 and 1.0). These fraction correct values Fc correspond to the performance levels of $d' = 0.5$, 1, and 2, respectively, in a two-alternative forced-choice task (Green & Swets, 1966).

Threshold comparison

In Figure 3C the thresholds derived with each method are plotted against each other. They are expressed as differences relative to the standard value. Perfect agreement between the two methods is indicated by the main diagonal. To evaluate the statistical significance of the differences between thresholds, we estimated the 95% CI for each of the thresholds using the bootstrap technique (for details see Simulations, Variability of threshold estimates section).

Thresholds were said to be in agreement when either one of the two confidence intervals of a data point (vertical or horizontal corresponding to 2-IFC and MLDS, respectively) crossed the unity line. This criterion ensures that the point estimate of one method is included in the 95% CI of the other method. In Figure 3C all data points coincided with the unity line, resulting in a 100% agreement.

We used this measure to quantify the degree to which the consistency between the thresholds. For eight different standard values we performed $n = 1{,}000$ simulations, and Figure 4A shows a summary of the results for the average of the empirically observed noise level, $\sigma = 0.07$ (green lines). Thresholds agreed in more than 90% of the cases, and the agreement was also high across a range of noise levels that we tested, from $\sigma = 0.035$ to $\sigma = 0.14$ (see Supplementary material), which includes all the values of sensory noise observed in the experiments.

Thresholds that could not be obtained

The estimation procedure in either of the methods sometimes failed when sensitivity was low. When the stimulus was in a range where the sensory function is too shallow, for example for values below 0.4 in the sensory function in Figure 2A, the interpolation of scale differences was not possible. Similarly, the psychometric function was sometimes so shallow that it did not allow the read out of a threshold at a given performance level. These "failure" cases provide an additional test of consistency between the two methods, because when sensitivity is genuinely low, both methods should fail to provide a threshold estimate. We counted the number of cases in which either one or both of the methods did not provide a threshold estimate for a given performance level. The results are shown in Figure 4A (gray lines). It can be read from the

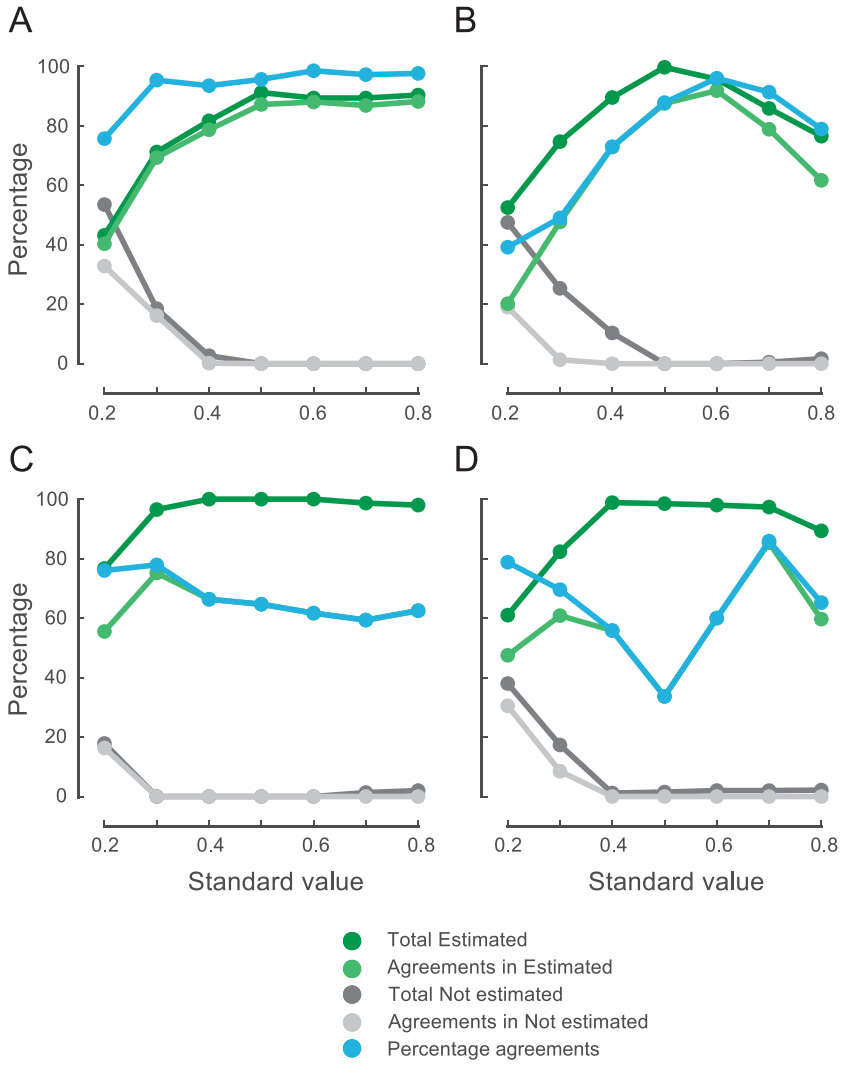

Figure 4. MLDS and forced-choice thresholds from simulations. At each standard level, thresholds for different performance levels could be successfully estimated (dark green) and have quantitative agreement between them (light green). There were cases in which thresholds could not be estimated (dark gray), from which an agreement occurred when both methods were unable to estimate it (light gray). The sum of agreement cases for estimated and not estimated thresholds is also shown (light blue). Percentage over 1,000 simulations: (A) Independent, equal-variance case with noise level $\sigma = 0.07$; (B) independent, unequal-variance case with increasing noise level from 0.035 to 0.14 (violation of equal-variance assumption); (C) same as (A) but with added uniform correlation in the sensory representation of $\rho = 0.8$ (violation of independence assumption); and (D) same as (B) but with added uniform correlation of $\rho = 0.8$ (both assumptions violated).

Figure that both methods did consistently fail to provide threshold estimates for standard values near zero.

Model assumptions

To test the effect of violations of some of the model assumptions on the agreements between thresholds, we repeated the simulations with a modified generative model. We introduced sensory noise that was not independent of the stimulus level but instead increased with the stimulus value. We also tested a model that included uniform correlations between the sensory representations (specific details in Supplementary material). These two modifications violate the assumptions of equal variance (assumption 1) and independence (assumption 2). As illustrated by Kingdom (2016) and tested in simulations by Maloney and Yang (2003), the scales themselves are insensitive to a violation of the equal variance assumption. However, violating the equal variance assumption did reduce the agreement between thresholds (Figure 4B) in particular for extreme standard values where the simulated noise

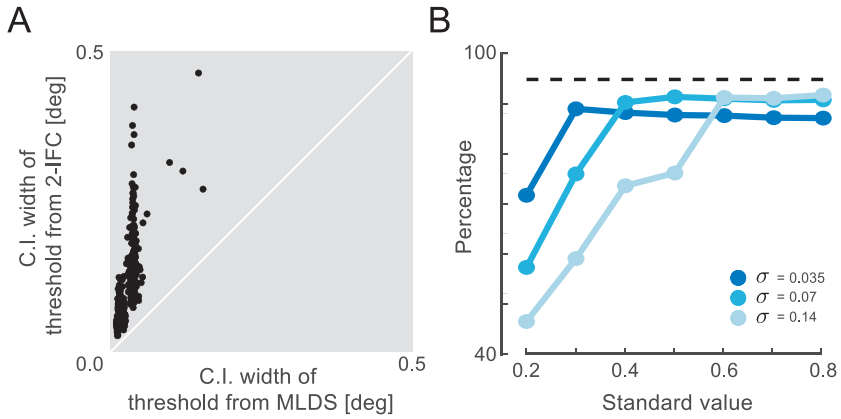

Figure 5. (A) Comparison of the variability in the threshold estimation. The width of the confidence intervals are plotted against each other for multiple simulations at one standard stimulus value $st = 0.4$ as example. (B) Coverage of threshold estimated at different standard levels, and for three different simulated noise levels (σ). Expected coverage of 95% is shown as a black dashed line.

was respectively lower or higher than in the equal-variance model. The reason for this is illustrated in Figure 3 which shows how threshold readout depends on noise on the sensory axis. Introducing correlations reduced the amount of agreement between thresholds independent of the standard value (Figure 4C). We observed the smallest agreements when both assumptions are violated (Figure 4D).

Variability of threshold estimates

To study the variability of threshold estimates, we made use of the bootstrap samples that are already generated by MLDS to calculate confidence intervals (CIs) for the scale values (error bars in Figure 3A; Knoblauch & Maloney, 2012). Bootstrap samples are generated from the response probabilities that are observed for each triad. Each bootstrap sample is a new perceptual scale and by default MLDS generates 1000 of these bootstrapped scales. To derive the bootstrap samples for a particular threshold value, we fitted a cubic spline to each bootstrap scale and determined the slant value corresponding to the threshold value (see Equation 1). From these bootstrap distributions, we obtained the 95% CI for each threshold (a detailed description of the procedure can be found in Appendix, Variability of the difference scale).

To compare the confidence intervals associated with each method, Figure 5A plots the widths of the respective CIs against each other, for one example standard. The main diagonal indicates equal width in the confidence intervals; data points above the main diagonal indicate that the width of the CIs for thresholds from MLDS were smaller than the width of the CIs for thresholds from 2-IFC. For all standard levels (Figure 5) and for all tested noise levels (Supplementary material) the majority of confidence intervals (99%) was smaller in MLDS than in 2-IFC. This is curious because MLDS requires a smaller amount of data than 2-IFC.

A smaller width in the confidence intervals could either be due to a truly more precise estimate (less underlying variability), or alternatively, it could result from an insufficient coverage of the confidence intervals. This is a common problem in derivations using bootstrap techniques (Wichmann & Hill, 2001) and we tested this with an analysis of the coverage of the scales and the derived thresholds. We calculated coverage by counting how many times the "true" value (as defined in the generative model) was contained in the estimated confidence interval of a scale or threshold. For confidence intervals to be credible, coverage across multiple simulations should reflect the confidence in the confidence interval, i.e., coverage should be 95% over multiple simulations for 95% CI.

Coverage of the scale estimates was adequate for the range of noise levels studied (Supplementary Figure S4). MLDS thus provides credible confidence intervals for the scale estimates that it was designed for. However, coverage for the threshold estimates was at best at 90% for nominal values of 95% (Figure 5B), and for stimulus values at shallow portions of the sensory function (e.g., smaller than 0.4) coverage was as low as 50%–60%. These results indicate that the confidence intervals for the thresholds derived with MLDS were indeed too narrow. Threshold variability might hence be underestimated when derived from MLDS in the way described above. This is an important caveat when using MLDS to estimate thresholds.

Upon suggestion of the reviewers, we performed a sanity check for the confidence intervals to test for their

stability and bias when trial numbers are high. We repeated the scale and threshold estimation procedure and calculated bias and coverage for increasingly large trial numbers (Supplementary Figures S8 and S9). For the scale estimation, the pattern of results indicates that coverage slightly improved with an increasing trial number. For the threshold estimation, coverage did not improve when the number of trials is tripled. This result suggests that low coverage was not due to a small sample size.

Summary and discussion

We simulated an observer model with a known sensory representation function. We compared thresholds derived from MLD scales with thresholds derived from a 2-IFC procedure at different standard values and performance levels. We found a high degree of consistency between thresholds obtained with each method, when all the assumptions are met (Figure 4A). The amount of agreement did not depend on the sensory noise level for the range of noise levels that was observed experimentally (see Supplementary material). The estimation procedure fails to obtain thresholds when sensitivity is low. In most of these cases both methods failed to estimate a threshold which is a further indication of consistency between them. The variability of threshold estimates, quantified as the width of their confidence intervals, is smaller for MLDS than for 2IFC thresholds (Figure 5A). This finding needs to be qualified by the coverage analysis which indicates that the bootstrapped confidence intervals for MLDS thresholds might be too small (Figure 5B).

The agreement between threshold estimates did not amount to the theoretically expected 100%. This might be due to the rather small number of simulated trials. However, the simulations should capture actual psychophysical experiments where it is not practicable to collect large numbers of trials. In addition, they capture the software pipeline of estimation and statistical inference, which could be prone to different kind of problems (e.g., numerical). Thus, the simulation results establish an upper bound for the agreement that is expected for a realistic amount of collected data and estimation procedures.

Finally, we found that violating the equal-variance assumption by MLDS might lead to disagreement between estimated thresholds. The disagreement is relevant because unequal variance models might fit behavioral data better when the equal variance assumption is violated in real data (e.g., Goris, Putzeys, Wagemans, & Wichmann, 2013).

Experiments

Observers

Six naïve (three male, three female; age range between 23 and 29) and two experienced observers (observers "O3" and "O6," two of the authors) participated in the study. All observers had normal or corrected-to-normal visual acuity. The participation of naïve observers was voluntary and financially compensated. Informed written consent was given by all observers prior to the experiment.

Stimuli and apparatus

Stimuli were planes textured with a "polka dot" pattern and slanted about their horizontal axis. They were generated in two steps. First, the textures containing the "polka dot" pattern were generated as 2500×500 pixel images. The "polka dot" pattern is created using a *hard core point process*, which is a random spatial process that avoids dot superposition by applying an inhibition radius to each point. Using the R package *spatstat* (Baddeley & Turner, 2005), we generated fifteen samples of this process following specifications from previous work (Rosas et al., 2004). The textures consisted of black dots (0.4–0.6 cd/m^2, 12 pixels or 0.5° visual angle in diameter in the fronto-parallel plane) on a gray background area (48–52 cd/m^2, Figure 1).

In a second step the textured planes were rendered in 3-D using OpenGL (Shreiner, Woo, Neider, & Davis, 2005). The planes were slanted and perspectively projected into 2-D. The so-generated planes were viewed through simulated circular apertures that subtended 8.3° of visual angle and were added at the depth of the screen distance.

Stimuli were displayed on a 24.1-in. LCD monitor (Eizo CG243W 496×310 mm, 1920×1200 pixels, 60 Hz) located in a dark cabin. Observers viewed the stimuli monocularly with their dominant eye at a distance of 60 cm. Eye dominance for each observer was determined with the Miles test (Miles, 1930) prior to the start of the experiment. The nondominant eye was covered with an eye-patch and the head rested on a chin rest. Stimulus presentation was controlled by a computer (Apple Mac Pro QuadCore 2.66 with a graphic card Nvidia GeForce 7300GT) that was running custom-made software which was based on *python* and the visualization library *pyglet*. Observers' responses were registered via the keyboard.

Procedure experiment 1: MLDS

In each trial, three stimulus exemplars that varied in slant were presented next to each other. Each of the slanted surfaces was rendered independently and viewed through a different circular aperture (see Figure 1). Slant values (s) varied between 0 (fronto-parallel) to 70° in steps of 10°. This spacing results in $p = 8$ possible slant values and a total number of $n = p!/((p-3)! \times 3!) = 56$ unique triads.

By design each triad consists of stimuli that are slanted so that the two intervals enclosed by the three stimuli do not overlap. The stimuli in a triad were presented in either ascending ($s_1 < s_2 < s_3$) or descending ($s_1 > s_2 > s_3$) order, and the order was randomized across trials. Observers were asked to report which of the pairs, (s_1, s_2) or (s_2, s_3), contained the bigger perceived difference in slant. Observers viewed the stimulus configuration with no time limit for their response. They indicated their choice by pressing a keyboard button, and this triggered the next trial after a delay of one second. No feedback was given as to the correctness of the response.

The full set of unique triads was presented in one experimental block, and 15 such blocks were presented within one session. In total, each observer judged 840 triads. This was the same amount of trials used in the simulations. Observers could pause after each block. Before the experiment observers were shown two to five examples of extreme triads (0°, 10°, 70°) and (0°, 60°, 70°), together with the "correct" answers and the corresponding keyboard presses. We employed this instruction method to ensure that observers understood the task. Comparing stimulus intervals is not an obvious task, and in previous experiments we noted that, instead of reporting the pair with the biggest perceived *difference*, some observers reported the pair that included the most extreme slant.

Procedure experiment 2: 2-IFC

A standard 2-IFC procedure was employed in Experiment 2. A trial started with a fixation cross that appeared for 1000 ms in the center of the screen. Then the first stimulus was presented for 200 ms. Its contrast ramped on and off from zero to full contrast and back to zero within the first and last 50 ms of presentation so that the stimulus was seen at full contrast for 100 ms. After a blank interstimulus interval of 500 ms, the second stimulus was presented with temporal parameters identical to those of the first. After stimulus offset, observers had to report which of the two stimuli was more slanted using a keyboard button to indicate first or second. Observers did not receive feedback about their performance. Standard and comparison stimuli were randomly assigned to the first or the second interval.

Discrimination performance was measured for the same four standard slant values (26°, 37°, 53°, and 64°) for which MLDS thresholds were predicted. Each standard slant was compared with one of eight comparison slants (four below and four above the standard slant) in a method of constant stimuli procedure. In the first session the range of comparison stimuli for each standard slant was selected based on the point estimates corresponding to performance levels of $d' = 0.5$, 1, 2, and 3 that were derived from the MLD scale (see Simulations, MLDS thresholds section). After the first session the comparison values were adjusted so as to provide good coverage of the psychometric function (Wichmann & Hill, 2001). The full experimental design contained 4 standards × 8 comparison values (four above and four below the standard) × 80 repeats resulting in 2,560 trials in total. This amount was the same as in the simulations. The presentation was randomized and the total number of trials was subdivided into 40 blocks of 64 trials each. Observers completed all trials in three to four sessions of maximum one hour duration. Experiment 2 was run on a different day from Experiment 1 and subsequent to it.

There are obvious differences in stimulus spacing as well as in the number of trials between both methods, and both factors might affect the shape of the respective fitted functions, scales, or psychometric functions. However, there is no principled way to equate these aspects across the procedures, and we would argue that they had little effect on the present results. We performed goodness-of-fit analyses for both procedures which showed that the fitted functions captured the data, and which also indicate that the stimulus choice was reasonable.

Results

The objective of the experiments was to compare sensitivity estimates from a forced-choice and an MLDS procedure at different positions along the perceptual scale. Before we report these results, we will show that the thresholds from the forced-choice task were comparable to those reported in earlier studies of slant-from-texture discrimination (Rosas et al., 2004).

The procedure in the forced-choice task (Experiment 2) was identical to that employed by Rosas et al. (2004). To capture sensitivity they computed an "area" measure, which was defined as the region between the two psychometric functions fitted separately for smaller and larger comparison values enclosed by the 60% and 80% percent performance levels (see Figure 3B). This

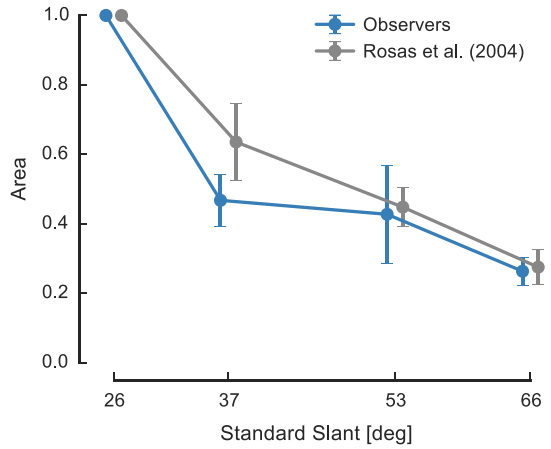

Figure 6. Sensitivity obtained from psychometric functions in Experiment 2. The "area" enclosed by the two psychometric functions and the 60% and 80% percentage correct (*y* axis, see Figure 3B for a depiction) is plotted for the different standards (*x* axis). Areas were normalized for each observer with respect to the maximum and aggregated across observers. Data from Rosas et al. (2004) is shown as reference (mean ± *SEM*).

"area" is small when the psychometric functions are steep, i.e. when sensitivity to slant differences is high, and conversely, it is large when sensitivity is low. Thus, the calculated area measure is inversely related to the sensitivity at a particular standard slant.

We computed the area measure for each standard value and each observer. The results are shown in Figure 6. In order to average across observers, the area measure was normalized relative to the highest value for each observer individually, because observers had different overall sensitivity to slant (interobserver variability). In all observers the area measure was maximal for a standard slant of 26°, indicating lowest sensitivity. For comparison Figure 6 also shows the mean normalized area of the five observers reported in Rosas et al. (2004, p. 1523). Apart from the variability between observers, sensitivity increased with slant, which is in accordance with the data reported by Rosas et al. (2004).

Threshold comparison

Thresholds for MLDS and 2-IFC were obtained in the same way as in the simulations. Figure 7 shows the data of one single observer. Thresholds from both methods are plotted against each other for performance levels of $d' = \pm 0.5$, 1, 2 and for the four standard values tested (panels). Data points lying on the main diagonal indicate a quantitative agreement between thresholds. This was observed for thresholds obtained at standard slants of 37°, 53°, and 66°. For a standard slant of 26°, a correspondence between thresholds from both methods was observed for comparisons that were larger than the standard. For comparisons that were below the standard MLDS, thresholds were smaller than 2-IFC thresholds. For some combinations of performance levels and standard values thresholds from either or both methods could not be calculated (see Simulations, Thresholds that could not be obtained section).

As described for the simulations, we classified thresholds to be in agreement when one of the confidence intervals of either method crossed the identity line (as in Figure 3C). Observers differed substantially in their proportion of agreement between thresholds. We sorted them according to the amount of agreement in descending order (Figure 8). There was agreement in 15 out of 16 data points (94%) for observer O1 in Figure 7; 11 of 14 (79%) for observer O2; 15 of 22 (68%) for observer O3; 10 of 18 (56%) for observer O4; 10 of 19 (53%) for observer O5; 10 of 21 (48%) for observer O6; 5 of 20 (25%) for observer O7; and 2 of 14 (14%) for observer O8. For observers O7

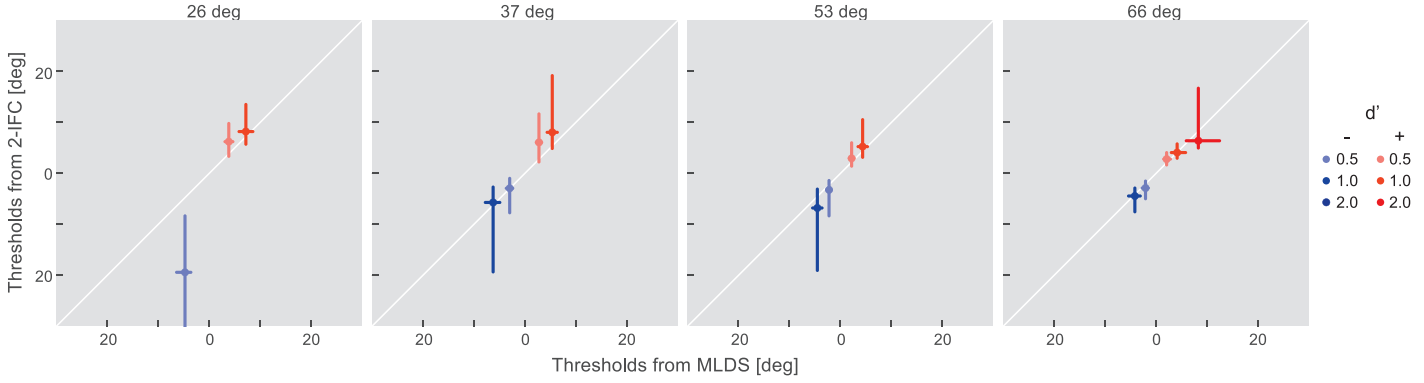

Figure 7. Threshold comparison for one observer (O1). Estimates of threshold from MLDS in Experiment 1 (*x* axis) and from the psychometric functions obtained in a 2-IFC procedure in Experiment 2 (*y* axis) are shown for each standard (different panels) and *d'* performance level, for comparisons above (+, warm colors) and below (−, cold colors) the standard. Thresholds are expressed as relative values to the standard. Error-bars denote the 95% CI of the point estimate.

and O8, the data points fell above the diagonal line for comparisons above the standard (Figure 8, red markers), and below the diagonal line for comparisons below the standard (blue markers). This pattern of results indicates that for these two observers' thresholds obtained with MLDS were consistently smaller than thresholds obtained with 2-IFC. In other words, MLDS estimated a higher sensitivity than the 2-IFC procedure; the opposite case did not occur. Taking all observers and standard levels together, 78 out of 144 (54%) estimated thresholds agreed between the two methods.

Thresholds that could not be obtained

Thresholds could not be obtained from either method for stimulus comparisons that involved the lowest standard value (26°) and/or comparison slant values below 30°. For example, for the observer depicted in Figure 7, thresholds from MLDS could not be obtained for performance levels of $d' = 1, 2$ for comparisons below the standard slant of 26°. The reason for this discrepancy was a shallow slope in the scale reflecting low sensitivity at that particular stimulus level.

As in the simulations, we counted the number of cases in which either one or both of the methods did not produce a threshold for our experimental results. A total of 22 cases occurred in which either one of both thresholds could not be obtained. Four out of the 22 cases were cases in which thresholds from MLDS were missing (at standard of 26° and 37°), eleven were cases in which thresholds from 2-IFC were missing (standard 26° and 37°) and seven were cases in which both thresholds were missing (all for a standard of 26°). So the methods consistently estimated low sensitivity in 32% of the cases for which thresholds could not be obtained.

Variability of threshold estimates

We also derived the variability of the threshold estimates for the experimental data. We found that the variability was lower for MLDS than for 2-IFC, consistent with the simulations. Figure 9 shows the widths of the confidence intervals for the thresholds obtained with each method from all observers. As in Figure 5A, confidence intervals were smaller for thresholds from MLDS than for thresholds from 2-IFC. Overall for 142 of the 144 threshold comparisons (98.6%), the width of the confidence interval was smaller for thresholds from MLDS (separate comparisons for each observer can be found in Supplementary Figure S7).

General discussion

The goal of the present study was to test whether judgments of stimulus appearance and judgments of stimulus discriminability are mutually consistent, which would suggest that both types of judgments rely on a common perceptual representation of the stimulus dimension under study. The evidence on this question is mixed (e.g., Hillis & Brainard, 2007; Krueger, 1989; Ross, 1997), but comparing suprathreshold judgments in a MLDS procedure and near-threshold judgments in a forced-choice procedure, Devinck and Knoblauch (2012) have reported that the two can be linked within a common signal detection framework. Using slant-from-texture stimuli, we conducted two experiments that independently measured the sensitivity to differences in slant. In the first experiment observers judged suprathreshold stimulus differences, and we derived thresholds from perceptual scales using the MLDS framework (Maloney & Yang, 2003). In the second experiment we measured sensitivity in a conventional two-alternative, forced-choice procedure and we derived thresholds from psychometric functions. For some observers there was agreement between thresholds obtained with both methods, but across observers the methods agreed in only 54% of the cases. For two observers (O7 and O8), sensitivity estimates from the MLDS procedure were consistently higher than those from the forced-choice procedure.

The observed lack of correspondence between the estimates could imply that the two tasks do indeed probe different perceptual representations of a stimulus. Alternatively, the lack of correspondence might result from violations of the model assumptions, and hence would not be informative about the relationship between appearance and discrimination tasks.

Violations of model assumptions

The equivalence between MLD scales and the 2-IFC procedure used in the present work relies on a number of theoretical assumptions concerning the sensory representation and the decision model. In the following we will describe the effect of violations of one or more of these assumptions on the estimated scales and the consequences for the estimation procedure.

Goodness of fit

The MLDS framework provides goodness of fit procedures that test the plausibility of the data being produced by a difference scaling model (Knoblauch & Maloney, 2008). In our data, the goodness of fit of the difference scales was insufficient for five out of eight

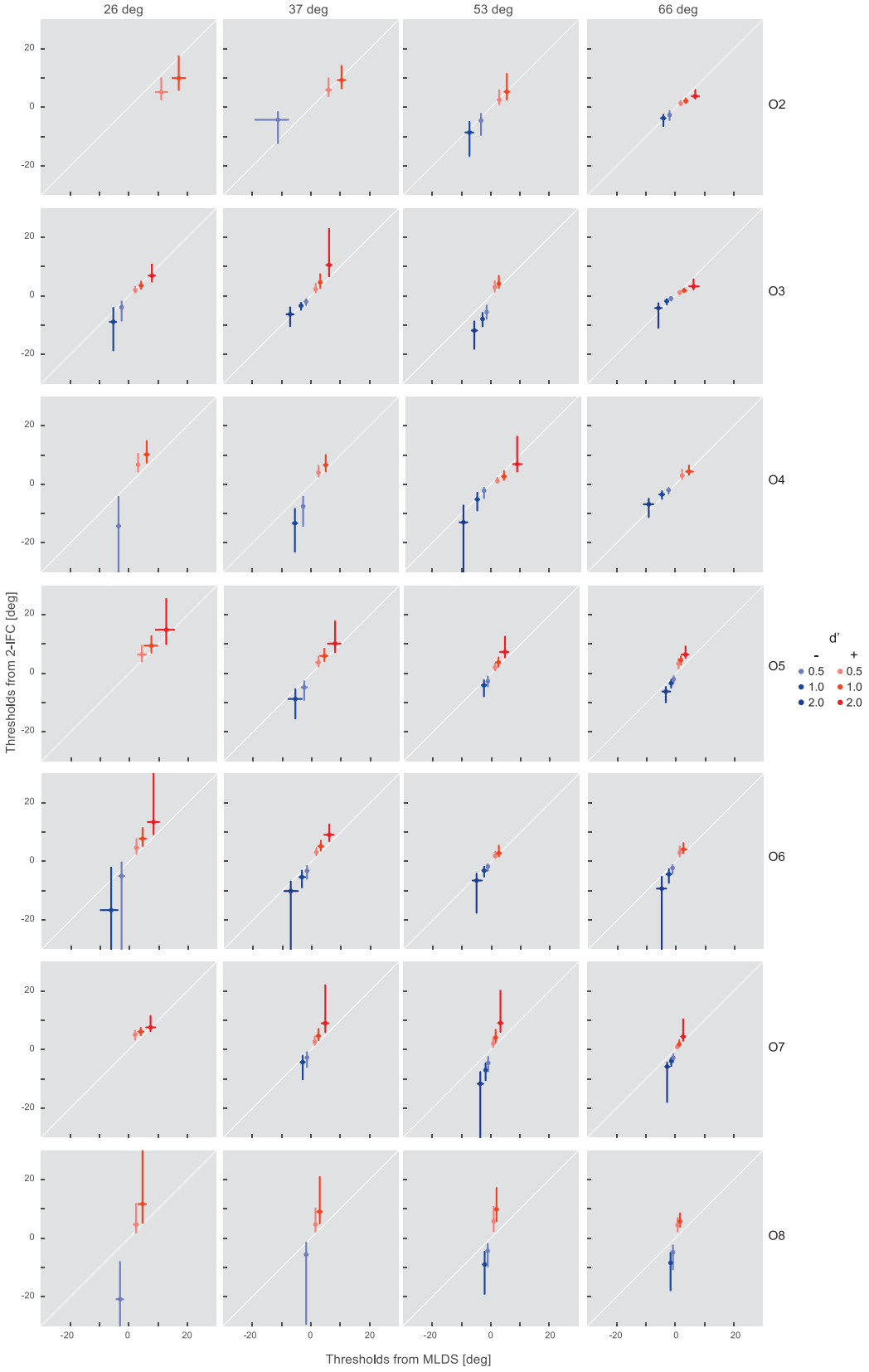

Figure 8. Threshold comparison for seven observers (O2–O8). Similar to Figure 7, thresholds relative to the standard obtained from the two methods are compared by observer (rows), standard (columns), and performance level (d'), for comparison values above (+, warm colors) and below (−, cold colors) the standard. Observers are sorted by percentage of threshold agreement between the two methods in descending order.

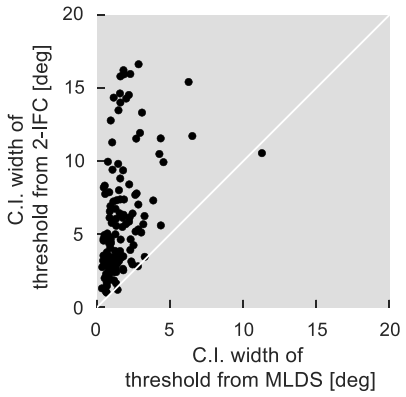

Figure 9. Comparison of the variability in the threshold estimation. The width of the confidence intervals from Figure 7 and Figure 8 are plotted against each other, for all observers and standard values.

observers when the default parameter values were used. We followed the refitting procedure suggested by Knoblauch and Maloney (2008, pp. 219–222) for these cases to modify the model specification. The procedure includes the estimation of "guess" and "lapse" rates, and a split of the raw data into two parts that were evaluated separately (detailed description of the goodness of fit procedure is provided in the Supplementary material). After the refitting procedure we obtained an appropriate goodness of fit for all observers, and we derived thresholds from the refitted scales. The thresholds that were derived from the adjusted scales were not markedly different from the original ones. In particular, the disagreement between thresholds that we observed in three observers was present with or without the goodness of fit adjustment. Thus, the model violations that were detected by the goodness of fit routines did not have much of an effect on the shape of the scale, at least for the present data.

Reconciling MLDS with 2-IFC thresholds

The assumption of independence between different levels of the sensory representation (assumption 2) is not and cannot be tested by the goodness of fit routine. If this assumption is violated, it would affect the noise and it would require an adjustment of the scaling factor that transforms the original MLDS scale into a scale in units of d'. The independence assumption would be violated when the sensory representations cannot be characterized as independent realizations of a Gaussian random variable but are instead correlated with each other. We tested the effect of these kinds of correlations in the sensory representation in simulations. Correlated sensory variables do indeed affect the threshold estimates. To illustrate the effect, we show that the magnitude of the correlation can be chosen so as to elicit a correspondence between thresholds derived from MLD scales and from 2-IFC. Figure 10 shows the thresholds for observer O8 for a simulated case in which the sensory representations are highly correlated ($\rho = 0.9$). As a consequence of this correlation, we give up the independence assumption and would have to rescale the perceptual scale by a factor of 0.6 (instead of the theoretical factor of two). In this scenario the resulting thresholds from MLDS correspond better with the thresholds from 2-IFC. Thus, an alternative transformation that accounts for a model violation can "produce" a higher agreement between the two types of thresholds. We are not aware of any method to test the assumption of independence empirically, and it is therefore not possible to evaluate which of the many possible transformations is closest to the true sensory representation.

A similar issue arises when we scrutinize the effect of violating the assumed decision rule (assumption 4). Based on the assumption that the sensory representation function is monotonically increasing, the decision rule can be expressed as a double difference operation (Δ variable in Figure 2B) instead of an absolute value operation (Δ variable in Figure 2A). This change from an absolute to a relative difference operation can have noticeable effects when the sensory representations are random variables (as assumed here) instead of fixed values. To explore the effect of the differencing operation, we simulated an observer that judged the triads by using either one of the two decision rules. To analyze the effect on the estimated noise, we applied MLDS to each of the two types of simulated responses, and found that the absolute difference operation produced higher noise estimates than the double difference operation. This difference increases progressively as the underlying sensory noise increases. Thus, the two decision rules can produce different results (see Supplementary material for simulations and details).

It is not possible to determine empirically whether or when observers apply an absolute or a relative differencing rule, they might even change the rule with varying difficulty of the judgment. One should be aware that a deviation from the assumed decision rule or a violation of the independence assumption may both affect the noise estimate although in opposite directions. Thus, the combined contributions of both factors can produce various types of deviation from the true scaling factor, and this deviation affects the scale and the derived sensitivity estimate.

Variability of thresholds estimated by MLDS

As in the simulations, we observed that the variability of thresholds derived from MLDS was smaller than the variability of thresholds derived from 2-IFC (Figure 5 and Figure 9). Again, this is

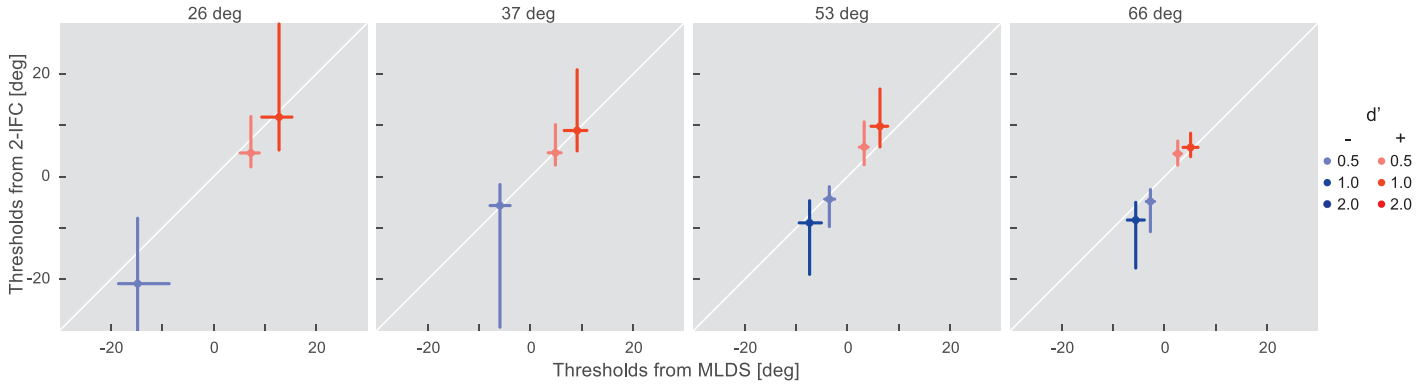

Figure 10. Threshold comparison for observer O8 when the difference scales are rescaled by a different factor to account for a possible dependence between different realizations of the random variable that characterizes the sensory representation (a factor of 0.6 corresponding to a correlation coefficient of 0.9).

counterintuitive because MLDS requires a smaller amount of data for the threshold derivation. However, the smaller variability must be interpreted with care, because our simulations revealed that the coverage of MLDS-derived threshold might be insufficient and the width of the confidence intervals might be underestimated. This should be considered for hypotheses tests as it may lead to Type-I errors.

However, apart from the coverage problem associated with our threshold derivation, MLDS provides an efficient method to acquire sensitivity estimates. In the present Experiment 1 we ran the MLDS procedure with 840 trials which took about 45 min per observer. In contrast, the 2-IFC procedure in Experiment 2 required 2,560 trials and lasted 3 hr. Thus, the difference in both the amount of data and the required acquisition time might be up to three to four times more for 2-IFC than for MLDS. MLDS thus provides an efficient alternative to forced-choice procedures to obtain a rough estimate of sensitivity.

Conclusions

In the present experiment we investigated the question of equivalence of thresholds derived from an MLD scale and thresholds derived from a forced-choice procedure. Using simulations, we established upper bounds for a possible agreement considering the theoretical model assumptions, the finite amount of collected data and the necessary software pipeline. Experimentally, we found varying degrees of correspondence between the methods for different observers. Out of a total of 144 threshold estimates, the methods' sensitivity estimates differed in 66 cases. We discuss that the equivalence of thresholds (or lack thereof) might either indicate a corresponding equivalence between the underlying perceptual representations (or lack thereof) as has been argued by Devinck and Knoblauch (2012), or alternatively, it might result from violations of the model assumptions.

An important point that has been made by one of the reviewers is that we gave the 2-AFC method the benefit of history. In the present analysis we used the 2-AFC method as a standard of reference against which we compared the sensitivity estimates derived with MLDS. Accordingly, we tested the effect of model violations on threshold agreement only for the assumptions underlying MLDS. However, considering the present data and the numerous benefits associated with the experimental procedures of MLDS, it might be warranted to try to elaborate the first principles case of which of the two methods we would trust more if we started out de novo.

Our positive evaluation of the MLDS method is corroborated by recent results from Kingdom (2016) who used MLDS to decide between competing theories of internal noise in contrast transduction. In summary, we conclude that MLDS, as state-of-the-art scaling method, seems to have great potential to be used beyond the purpose that it was originally designed for.

Keywords: MLDS, appearance, discrimination, signal detection theory, slant from texture, near-threshold performance, psychometric function

Acknowledgments

This work has been supported by an Emmy-Noether research grant of the German Research Foundation to Marianne Maertens (DFG MA5127/1-1) and by the GRK Graduate School "Sensory computation in Neural Systems" from the German Research Foundation to Guillermo Aguilar (GRK1589/1). We would like to thank David Brainard, Oliver Henaff, Michael Kubovy, Laurence Maloney, and one anonymous

reviewer for their constructive suggestions which helped improve this manuscript.

Commercial relationships: none.
Corresponding author: Guillermo Aguilar.
Email: guillermo@bccn-berlin.de.
Address: Modelling of Cognitive Processes Group, Department of Software Engineering and Theoretical Computer Science, Technische Universität Berlin, Berlin, Germany; Bernstein Center for Computational Neuroscience, Berlin, Germany.

Footnotes

[1] Which, in turn, is analogous to Thurstone's case V of the Law of Comparative Judgment (Thurstone, 1927).

[2] Although MLDS assumes a monotonically increasing function for the decision model, there is no restriction of monotononicity imposed for the coefficients found by the GLM solver. Thus, nonmonotonic scales from MLDS are possible outcomes (see O2 and O3 in Supplemental Figure S3).

[3] This restricted case can be derived from the MLDS model only because: (i) the decision variable after the differencing is assumed to be Gaussian, for which the simplest case is when it is produced by underlying Gaussian distributed representations; and (ii) equal-variance is the simplest case to relate the variance of each representation with the variance of the decision variable. Other models (e.g. unequal-variance) cannot be derived as they would be underconstrained.

References

Baddeley, A., & Turner, R. (2005). spatstat: An r package for analyzing spatial point patterns. *Journal of Statistical Software*, *12*(6), 1–42.

Baird, J. (1978). *Fundamentals of scaling and psychophysics*. New York: Wiley.

Baird, J. (1989). The fickle measuring instrument. *Behavioral and Brain Sciences*, *12*(02), 269–270, doi:10.1017/S0140525X00048585.

Charrier, C., Maloney, L. T., Cherifi, H., & Knoblauch, K. (2007). Maximum likelihood difference scaling of image quality in compression-degraded images. Journal of the Optical Society of America. *A, Optics, Image Science, and Vision*, *24*(11), 3418–3426.

Devinck, F., & Knoblauch, K. (2012). A common signal detection model accounts for both perception and discrimination of the watercolor effect. *Journal of Vision*, *12*(3):19, 1–14, doi:10.1167/12.3.19. [PubMed] [Article]

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman and Hall.

Emrith, K., Chantler, M. J., Green, P. R., Maloney, L. T., & Clarke, A. D. F. (2010). Measuring perceived differences in surface texture due to changes in higher order statistics. *Journal of the Optical Society of America. A, Optics, Image Science, and Vision*, *27*(5), 1232–1244.

Fechner, G. (1860). *Elemente der psychophysik* [Translation: *Elements of psychophysics*]. Leipzig, Germany: Breitkopf und Hartel.

Fleming, R. W., Jäkel, F., & Maloney, L. T. (2011). Visual perception of thick transparent materials. *Psychological Science*, *22*(6), 812–820, doi:10.1177/0956797611408734.

Gescheider, G. (1997). *Psychophysics the fundamentals* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Goris, R. L. T., Putzeys, T., Wagemans, J., & Wichmann, F. A. (2013). A neural population model for visual pattern detection. *Psychological Review*, *120*(3), 472–496, doi:10.1037/a0033136.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Hillis, J. M., & Brainard, D. H. (2007). Do common mechanisms of adaptation mediate color discrimination and appearance? Contrast adaptation. *Journal of the Optical Society of America. A, Optics, Image Science, and Vision*, *24*(8), 2122–2133, doi:10.1364/JOSAA.24.002122.

Kingdom, F. A. (2016). Fixed versus variable internal noise in contrast transduction: The significance of Whittle's data. *Vision Research*, *128*, 1–5, doi:10.1016/j.visres.2016.09.004.

Knill, D. C. (1998). Discrimination of planar surface slant from texture: Human and ideal observers compared. *Vision Research*, *38*(11), 1683–1711, doi:10.1016/S0042-6989(97)00415-X.

Knoblauch, K., & Maloney, L. T. (2008). MLDS : Maximum likelihood difference scaling in R. *Journal of Statistical Software*, *25*, 1–26, doi:10.1.1.204.8835.

Knoblauch, K., & Maloney, L. T. (2012). *Modeling psychophysical data in R*. New York, NY: Springer.

Krueger, L. E. (1989). Reconciling Fechner and Stevens: Toward a unified psychophysical law. *Behavioral and Brain Sciences*, *12*, 251–320, doi:10.1017/S0140525X0004855X.

Kuss, M., Jäkel, F., & Wichmann, F. A. (2005).

Bayesian inference for psychometric functions. *Journal of Vision*, *5*(5):8, 478–492, doi:10:1167/5.5.8. [PubMed] [Article]

Maloney, L. T., & Yang, J. N. (2003). Maximum likelihood difference scaling. *Journal of Vision*, *3*(8):5, 573–585, doi:10:1167/3.8.5. [PubMed] [Article]

Miles, W. R. (1930). Ocular dominance in human adults. *The Journal of General Psychology*, *3*(3), 412–430, doi:10.1080/00221309.1930.9918218.

Obein, G., Knoblauch, K., & Viénot, F. (2004). Difference scaling of gloss: Nonlinearity, binocularity, and constancy. *Journal of Vision*, *4*(9):4, 711–720, doi:10:1167/4.9.4. [PubMed] [Article]

Paulun, V. C., Kawabe, T., Nishida, S., & Fleming, R. W. (2015). Seeing liquids from static snapshots. *Vision Research*, 1–12, doi:10.1016/j.visres.2015.01.023.

Rosas, P., Wichmann, F. A., & Wagemans, J. (2004). Some observations on the effects of slant and texture type on slant-from-texture. *Vision Research*, *44*(13), 1511–1535, doi:10.1016/j.visres.2004.01.013.

Ross, H. E. (1997). On the possible relations between discriminability and apparent magnitude. *British Journal of Mathematical and Statistical Psychology*, *50*(2), 187–203, doi:10.1111/j.2044-8317.1997.tb01140.x.

Saunders, J. A. (2003). The effect of texture relief on perception of slant from texture. *Perception*, *32*(2), 211–233, doi:10.1068/p5012.

Saunders, J. A., & Backus, B. T. (2006). Perception of surface slant from oriented textures. *Journal of Vision*, *6*(9):3, 882–897, doi:10.1167/6.9.3. [PubMed] [Article]

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

Shreiner, D., Woo, M., Neider, J., & Davis, T. (2005). *Opengl programming guide: The official guide to learning opengl, version 2* (5th edition). Upper Saddle River, NJ: Addison-Wesley.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review*, *64*(3), 153–181.

Stevens, S. S. (1975). *Psychophysics: Introduction to its perceptual, neural, and social prospects*. New York: Wiley.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychology Review*, 273–286.

Todd, J. T., Christensen, J. C., & Guckes, K. M. (2010). Are discrimination thresholds a valid measure of variance for judgments of slant from texture? *Journal of Vision*, *10*(3):22, 1–18, doi:10.1167/10.3.22. [PubMed] [Article]

Todd, J. T., Thaler, L., & Dijkstra, T. M. H. (2005). The effects of field of view on the perception of 3D slant from texture. *Vision Research*, *45*(12), 1501–17, doi:10.1016/j.visres.2005.01.003

Treisman, M. (1964a). Sensory scaling and the psychophysical law. *Quarterly Journal of Experimental Psychology*, *16*(1), 11–22, doi:10.1080/17470216408416341.

Treisman, M. (1964b). What do sensory scales measure? *Quarterly Journal of Experimental Psychology*, *16*(4), 387–391, doi: 10.1080/17470216408416400.

Velisavljević, L., & Elder, J. H. (2006). Texture properties affecting the accuracy of surface attitude judgements. *Vision Research*, *46*(14), 2166–2191, doi:10.1016/j.visres.2006.01.010.

Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: II. Bootstrap-based confidence intervals and sampling. *Perception & Psychophysics*, *63*(8), 1314–1329.

Appendix

MLDS with the method of triads

In the method of triads variant of MLDS (Maloney & Yang, 2003), three stimuli are presented for comparison in one triad (s_1, s_2, s_3). The task is to decide which of the two adjoining pairs, (s_1, s_2) or (s_2, s_3), comprises the larger interval. The stimuli are assumed to produce discrete responses on a singular sensory representation (Ψ_s), and these sensory responses are used to compute a decision variable Δ. It is further assumed that the decision variable Δ is corrupted by additive noise $\varepsilon \sim N(0, \sigma^2)$.

$$\Delta = |\Psi_{s_3} - \Psi_{s_2}| - |\Psi_{s_2} - \Psi_{s_1}| + \epsilon \quad \text{(A1)}$$

When $\Delta > 0$, the model evaluates the interval in pair (s_2, s_3) as larger, (s_1, s_2) otherwise. The p different stimuli are chosen from the stimulus dimension, giving a total number of $n = p!/((p-3)! \times 3!)$ unique triads to be judged.

The current implementation of MLDS performs the estimation of the perceptual scale using a generalized linear model (GLM) for logistic regression (Knoblauch & Maloney, 2008, 2012; Maloney & Yang, 2003). The method assumes that the sensory representation function is monotonically increasing, thus avoiding the

absolute value operation in Equation A1 and leading to a decision variable that can be rewritten as a linear combination of the sensory variables

$$\begin{aligned}\Delta &= (\Psi_{s_3} - \Psi_{s_2}) - (\Psi_{s_2} - \Psi_{s_1}) + \epsilon \\ &= \Psi_{s_3} - 2\Psi_{s_2} + \Psi_{s_1} + \epsilon \end{aligned} \quad \text{(A2)}$$

We simulated the output of an absolute (Equation A1) and a simple difference (Equation A2) decision rule for different noise levels and found that the respective estimates differed only in any relevant way for noise levels that were much higher than what we observed experimentally, thus allowing us to work with the much easier to handle difference decision rule (see Supplementary material for details)

The GLM takes the responses of an observer ($\boldsymbol{Y}$) from a triad experiment and the stimulus design matrix ($\boldsymbol{X}$), and it estimates a set of coefficients $\boldsymbol{\beta}$ that best account for the data. Formally, this model is described by

$$g(\mathrm{E}[\boldsymbol{Y}]) = \boldsymbol{X}\boldsymbol{\beta} \quad \text{(A3)}$$

where $\boldsymbol{Y}$ is a vector of length n with entries 0 or 1, indicating the observer's response (for first vs. second pair, respectively). $\boldsymbol{X}$ is the design matrix of size $n \times p$, whereby n is the total number of triads and p is the number of stimulus levels sampled as well as the number of estimated points on the perceptual scale. Each row in matrix $\boldsymbol{X}$ contains nonzero entries (1,-2, and 1) in the columns corresponding to the stimulus values for the presented triad values (s_1, s_2, and s_3), and zero entries in the remaining $p - 3$ columns. The coefficient vector $\hat{\boldsymbol{\beta}}$ is of length p, and it contains the scale estimates.

The link function $g()$ is required to establish the relationship between the *linear predictors* $\boldsymbol{X}\boldsymbol{\beta}$ and the mean response variable $\mathrm{E}[\boldsymbol{Y}]$, where $\boldsymbol{Y}$ is binomially distributed with $n = 1$ (also known as a Bernoulli process). We used the default link function for MLDS, that is, the inverse of the Gaussian cumulative distribution function (Φ^{-1}), as it has been shown to be robust against distribution changes and deviations from the equal variance assumption (Maloney & Yang, 2003). The coefficients $\hat{\boldsymbol{\beta}}$ are estimated by maximum likelihood using standard GLM solvers (Knoblauch & Maloney, 2008).[2]

Difference scales and signal detection theory

The difference scale that is estimated by MLDS using the GLM implementation consists of the coefficients $\hat{\boldsymbol{\beta}}$ in Equation A3. They constitute an "unconstrained scale" (Knoblauch & Maloney, 2012) which—by design has a lowest value of zero ($\hat{\beta}_1 = 0$) and a highest value ($\hat{\beta}_p$) that is equal to the inverse of the noise parameter (ε) in the decision process,

$$\hat{\beta}_p = \frac{1}{\hat{\sigma}} \quad \text{(A4)}$$

whereby $\hat{\sigma}$ is the estimate of σ in Equation A2.

A difference scale with that type of parametrization can be converted to a normed scale that is defined in units of d', when the following assumptions are met. First, it is assumed that the decision process is not stochastic but deterministic. This would attribute all of the observed noise to the sensory representation, ψ_s, which is a Gaussian random variable with mean Ψ_s. Second, it is assumed that the noise is constant, i.e. independent of the stimulus level. Finally, it is assumed that the sensory representations are independent of each other. It follows from these assumptions that the ψ_s are independent Gaussian random variables with equal variance.[3] Then, the noise parameters can be "carried" to the sensory representation, by rewriting the decision model (Equation A2) in this way

$$\psi_{s_i} \sim \mathcal{N}\left(\Psi_{s_i}, \frac{\sigma^2}{4}\right) \quad \text{(A5)}$$

$$\Delta = (\psi_{s_3} - \psi_{s_2}) - (\psi_{s_2} - \psi_{s_1}) \quad \text{(A6)}$$

The variance of the decision variable Δ is σ^2 (Equation A2). When rewriting the model equations, the variance of each sensory representation ψ_s must be adjusted so that Equation A2 still holds. Because the decision variable Δ is computed as a linear combination of four independent, Gaussian random variables, its variance is four times the variance of each individual variable ψ_{s_i}. Therefore each individual variance in the sensory representation needs to be "corrected" by a factor of 1/4.

Yet, MLDS provides the noise estimate $\hat{\sigma}$ (Equation A4) as an estimate of parameter σ of the decision variable and not of the sensory representation directly. By knowing the above explained relationship between the variance in the sensory representation and in the decision variable, the difference scale can be adjusted so as to represent the variance in the sensory representation. The $\hat{\sigma}$ estimated by MLDS corresponds to the double of the variance present in the sensory representation ψ_s (Equation A5). Thus, the conversion is accomplished by multiplying the original scale by a factor of two (see also Devinck & Knoblauch, 2012). Formally, the new transformed scale maximum is two times the maximum of the original scale

$$\hat{\beta}_p{}' = \frac{1}{\frac{\hat{\sigma}}{2}} = 2\frac{1}{\hat{\sigma}} = 2\hat{\beta}_p$$

This new scale $\hat{\beta}'$ is in "d' units," i.e., an interval difference of one in the scale dimension should represent a performance of d' of one, when all assumptions are met.

Estimation of variability in MLDS

Variability of the difference scale

The variability of the scale estimation (see Appendix, MLDS with the method of triads section) can be studied using bootstrapping techniques (Knoblauch & Maloney, 2008, 2012), with the goal of estimating the variability of the coefficient $\hat{\boldsymbol{\beta}}$. For that purpose, Equation A3 is rearranged in order to compute the mean response probability for each triad from the fitted data $\mathrm{E}[\boldsymbol{Y}] = g^{-1}(\mathbf{X}\hat{\beta})$. The obtained vector $\mathrm{E}[\boldsymbol{Y}]$ contains the expected probability of a Bernoulli variable (Y) for each triad, in other words, the mean probability of binary responses given the presented stimulus values in each triad. These probabilities are used to simulate a Bernoulli response in each triad, which is in turn used to estimate a new set of coefficients $\hat{\beta}^*_j, j = 1...p$, i.e., $\hat{scale}^*_i$ using the same GLM procedure described above. The coefficients $\hat{\beta}^*_{ij}, j = 1...p$ are the *i-th* bootstrap sample, and many bootstrap samples are drawn by repeating the simulation procedure many times ($N_s =$ 10,000), obtaining a matrix $\boldsymbol{S}$ of $N_s \times p$ entries.

$$S_{i,j} = \begin{pmatrix} \hat{\beta}^*_{1,1} & \cdots & \hat{\beta}^*_{1,p} \\ \vdots & \ddots & \vdots \\ \hat{\beta}^*_{N_s,1} & \cdots & \hat{\beta}^*_{N_s,p} \end{pmatrix} = \begin{pmatrix} \hat{scale}^*_1 \\ \vdots \\ \hat{scale}^*_{N_s} \end{pmatrix}$$

Confidence intervals for the *j-th* scale value $[\hat{\beta}^{(\alpha)}_j, \hat{\beta}^{(1-\alpha)}_j]$, $j = 1...p$ were obtained by taking the "bias-corrected and accelerated" (BCa; Efron & Tibshirani, 1993) percentiles corresponding to a 95% CI.

In other words, the confidence intervals for each scale estimate were obtained from the distribution of bootstrap samples along each individual column of the matrix $\boldsymbol{S}$.

Variability of threshold estimation

The same matrix $\boldsymbol{S}$ can be used to obtain the variability of the threshold estimation. The same fitting and readout procedure applied to the point estimate of the scale (see Simulations, MLDS thresholds section) was applied to each bootstrap sample $\hat{\beta}^*_{ij}, j = 1...p$. We fitted a spline to each scale bootstrap sample, i.e. to each row in matrix $\boldsymbol{S}$, and from this scale we readout a bootstrap threshold. By repeating to all $i = 1.N_s$ bootstrap samples, a distribution of thresholds bootstrap samples is calculated, and confidence intervals $[\hat{\theta}^{(\alpha)}, \hat{\theta}^{(1-\alpha)}]_{st,d|\mathrm{MLDS}}$ were obtained by taking the "bias-corrected and accelerated" (BCa; Efron & Tibshirani, 1993) percentiles corresponding to a 95% confidence interval.