Available online at www.sciencedirect.com
ScienceDirect
Vision Research 46 (2006) 4279–4293
Vision Research
www.elsevier.com/locate/visres

Separating transducer non-linearities and multiplicative noise in contrast discrimination

Stanley A. Klein

University of California, Berkeley, USA

Received 3 November 2005; received in revised form 16 March 2006

Abstract

It has been difficult to isolate the factors that limit contrast discrimination, one of the most fundamental aspects of the visual system. Kontsevich, Chen, and Tyler (2002) claim to have found a method that can answer the question of why discrimination thresholds increase with reference contrast. Is it because of a saturating contrast response function or because of increasing (multiplicative) noise? Based on four datasets they conclude that multiplicative noise is the controlling factor. Georgeson and Meese (2006) disagree and claim the jury is still out because only one of the four datasets has sufficiently good statistics to support the KCT claim. I reanalyze the KCT data and come to a different conclusion. I agree with GM that two of the four datasets have thresholds that are too low to be useful in discriminating models and that one dataset supports the KCT claim. The fourth dataset is the most interesting one in that it provides the strongest support for the KCT claim, but GM throw it out because the χ^2 of the best fit is high. The present paper makes a number of points: (1) two novel methods are used to fit the fourth dataset. One pair of models is based on the strong "finger error" asymmetry between the high and low contrast asymptotes of the psychometric function in the fourth dataset. I find that some version of multiplicative noise is needed. However, it may be multiplicative noise that depends on prior trials rather than just on the present trial. (2) Another model that allows the contrast response function to have maximal freedom fits the fourth dataset with a reasonable chi square, and with a need for multiplicative noise. (3) I examine alternative parameterizations of the model functions used by KCT and GM that provide a more intuitive interpretation of the parameters. In summary, although I find the data do support a generalized form of multiplicative noise, I agree with GM that the jury remains out about what are the factors that limit contrast discrimination.

Keywords: Human vision; Psychophysics; Constrast discrimination; Noise; Modelling; Statistics; Multiplicative noise

1. Introduction

Kontsevich et al. (2002) [KCT] measured contrast discrimination over a wide range of contrast levels in order to address the long-standing question of whether the threshold increase as a function of reference contrast is due to increased noise (multiplicative noise) or to a non-linear contrast response function. KCT claim to have answered the challenging question in favor of multiplicative noise, with no evidence of contrast saturation. This topic has been of great interest to me because it was the theme of Stromeyer and Klein (1974), my first vision paper. We used both a mildly saturating transducer function plus multiplicative noise (see Section 5.3) to account for our TvC (threshold vs. contrast) data whose shape was later called the contrast discrimination dipper function (Legge & Foley, 1980). The preceding paper by Georgeson and Meese (2006) [GM] questions some aspects of the KCT analysis. The question before us now is to what extent do the KCT data constrain models of contrast discrimination. This dataset and its analysis bring up a number of fascinating questions: (1) there is a peculiarity in the "finger errors" of the KCT data that may provide a key to understanding why the KCT multiplicative noise fit is so good. (2) Fitting the KCT data brings up a number of statistical issues, including optimal ways of parameterizing identical fits. (3) The KCT data forces us to be careful and thoughtful

E-mail address: sklein@socrates.berkeley.edu

0042-6989/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.visres.2006.03.032

about how to do χ^2 goodness-of-fit analyses. For linear regression there is a close connection between the z test or t test standard error of parameter estimates and the χ^2 or F test for goodness-of-fit. The non-linear models used to fit the KCT data provide a dramatic violation of that connection. (4) GM argue that the data from one of the four subjects (the subject providing strongest evidence for multiplicative noise) should be thrown out because of a high χ^2. Their argument is questioned. (5) Finally, on the original question of whether multiplicative noise is present, our answer surprisingly depends on the definition of multiplicative noise. If the definition allows perturbations from preceding trials, similar to Lu and Dosher (1999) then we agree with KCT that multiplicative noise is present. However, if one defines multiplicative noise as depending just on the stimuli in the current trial, we agree with GM that "the jury is still out", but for reasons related to "finger errors" whose source is actually something different.

1.1. The KCT data

KCT used a temporal 2AFC method to gather contrast discrimination data on two subjects, AK and SV. Each subject was tested using a sustained and a transient (16 Hz counterphase) condition. Thus there are effectively four subjects tested: AK-S, SV-S, AK-T, and SV-T. Three reference contrasts (15, 30, and 60%) and a large range of test contrasts were used as shown in Fig. 1. For reference contrasts of 15 and 60% the test contrast was always positive. For the 30% reference contrast both positive and negative test contrasts were used. During a run the reference contrast was fixed and all the different test contrasts were

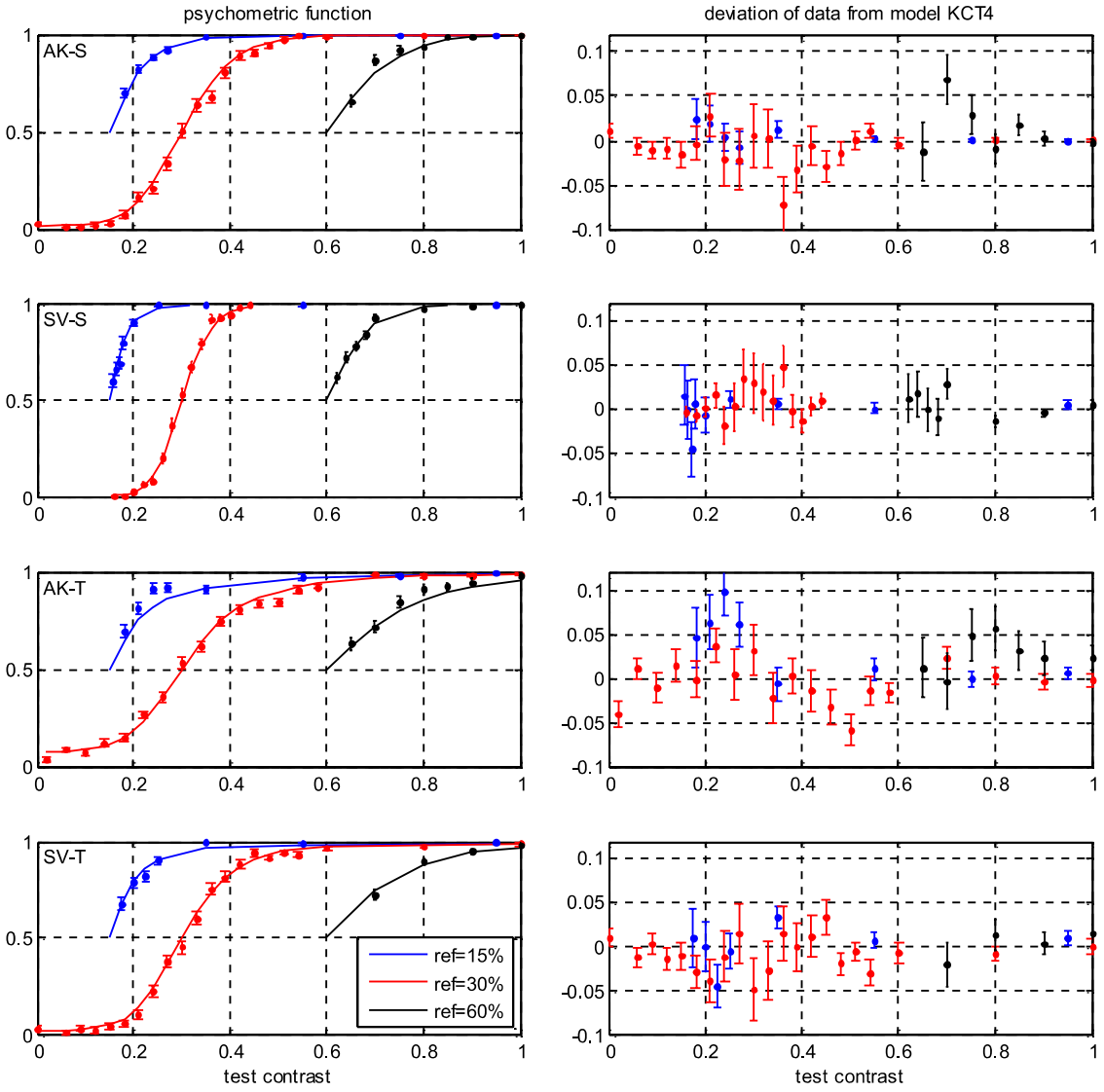

Fig. 1. The left panels present the KCT data for the four observers together with the best fitting model, KCT4 (solid lines). The contrast reference of 15, 30, and 60% are shown in blue, red and black respectively. The error bars are 1 SE, based on binomial statistics: SE $= \sqrt{(p(1-p)/N)}$. The abscissa is the total test contrast, c_{test}. The right panels are the identical data but now the difference between the data and the model are shown in order to provide an expanded view of the deviations of data from model. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this paper.)

randomly intermixed (personal communication). As will be discussed, this intermixing of test contrasts may have been responsible for producing results that look like multiplicative noise. Two hundred trials were presented at most test contrasts using the method of constant stimuli. For data that looked noisy further trials were presented up to 486 trials per test contrast. Further details regarding the data being fit can be found in the GM companion paper and in the original KCT paper.

1.2. The KCT one parameter contrast response function

KCT and GM fit the dataset with a one parameter contrast response function, $r(c)$

$$r(c) = c^p \tag{1}$$

plus multiplicative noise, $\sigma(c)$, of the form

$$\sigma(c) = kr(c)^q. \tag{2}$$

The connection of these two functions to 2AFC contrast discrimination d'_{disc} is

$$d'_{\text{disc}} = (r(c_{\text{ref}} + \Delta c_{\text{test}}) - r(c_{\text{ref}}))/\sigma_{\text{2AFC}}(c_{\text{ref}}, \Delta c_{\text{test}}), \tag{3}$$

where c_{ref} is one of three reference contrast (0.15, 0.30, or 0.60), and Δc_{test} is the test increment that is added to the reference contrast in one of the two intervals. The internal noise in a 2AFC task, σ_{2AFC}, is the square root of the mean noise variance at the two contrasts being discriminated: $c_{\text{ref}} + \Delta\ c_{\text{test}}$ vs c_{ref}.

$$\sigma_{\text{2AFC}}(c_{\text{ref}}, \Delta c_{\text{test}}) = ((\sigma^2(c_{\text{ref}} + \Delta c_{\text{test}}) + \sigma^2(c_{\text{ref}}))/2)^{1/2}. \tag{4}$$

The contrast discrimination threshold, c_{th}, is defined as the value of the test increment, Δc_{test}, that gives $d'_{disc} = 1$. That is

$$1 = (r(c_{\text{ref}} + c_{\text{th}}) - r(c_{\text{ref}}))/\sigma_{\text{2AFC}}. \tag{5}$$

When $c_{\text{th}} \ll c_{\text{ref}}$ Eq. (5) can be approximated as

$$1 \approx \mathrm{d}r/\mathrm{d}c\ c_{\text{th}}/\sigma(c_{\text{ref}}) = p\ c_{\text{th}} c_{\text{ref}}^{p(1-q)-1}/k. \tag{6}$$

Eq. (6) can be rewritten in terms of the TvC (test vs. pedestal contrast) Weber fraction, W (KCT Eq. (13))

$$W(c_{\text{ref}}) = c_{\text{th}}/c_{\text{ref}} = k/pc_{\text{ref}}^{-p(1-q)}. \tag{7}$$

Two problems with the parameterization given by Eqs. (1) and (2) are: (1) The parameters p, q, and k are unnecessarily correlated. (2) Parameters p and k are not directly related to the raw data. In order to decorrelate the parameters we would like to transform the parameters so that W can be written as

$$W(c_{\text{ref}}) = b(c_{\text{ref}}/0.3)^{-w}, \tag{8}$$

where b is the Weber fraction in the middle of the contrast range ($c_{\text{ref}} = 0.3$) and the log–log slope of the TvC function, $\log(W)$ vs. $\log(c_{\text{ref}})$ at $c_{\text{ref}} = 0.3$ is $-w$. By comparing Eqs. (7) and (8) we see the new parameters are given in terms of the KCT parameters by

$$w = p(1-q) \tag{9}$$

and

$$b = W(0.3) = k * 0.3^{-w}/p. \tag{10}$$

In terms of the new parameters Eqs. (1) and (2) become

$$r_{\text{new}}(c) = (c/0.3)^{w/(1-q)}, \tag{11}$$

$$\sigma_{\text{new}}(c) = bw/(1-q) r_{\text{new}}(c)^q. \tag{12}$$

There are several advantages of using parameters like b and w that are closely tied to the TvC data:

(1) The values of b and w are robust to assumptions about the shape of the contrast response function and multiplicative noise, as we will see by fitting the KCT dataset using five different models.
(2) The values of b and w and their standard errors are directly interpretable in terms of familiar properties of the TvC function, namely the Weber fraction and the log–log slope of the TvC function.
(3) The parameters b and w are less correlated than the original k and p because they are defined in the middle of the data range ($c_{\text{ref}} = 0.3$) rather than $c_{\text{ref}} = 1.0$ for the original parameterization. The reduced correlation causes the standard errors to be less than they would be in the original parameterization.

1.3. Finger errors

An important aspect of the KCT data is that there are a substantial number of errors for test contrasts at which discrimination should have been easy. These are commonly called "finger errors". The fits to the data will be seen to exhibit a peculiar asymmetry in finger errors in the two tails of the probability correct of observer AK for both the sustained and transient (16 Hz) presentations at a reference contrast of $c_{\text{ref}} = 30\%$ (the only reference contrast with decrements as well as increments). Observers SV-S and SV-T on the other hand, showed no asymmetry. The Discussion considers how the asymptotic asymmetry in the raw data may be related to a form of multiplicative noise that is not captured by Eqs. (2) or (9).

In order to account for the finger error asymmetry a pair of finger error parameters, f_{low} and f_{high}, are introduced as follows

$$P = (1 - f_{\text{low}} - f_{\text{high}})P' + f_{\text{low}}, \tag{13}$$

P' is the idealized psychometric function that goes from 0 to 1. P is the actual psychometric function that goes from f_{low} to $1 - f_{\text{high}}$. KCT and GM, use only a single finger error parameter given by $g = f_{\text{low}} + f_{\text{high}}$. When we replicate the KCT and GM models with a single finger error parameter we use $f = f_{\text{low}} = f_{\text{high}} = g/2$ where g is the finger error parameter of KCT and GM.

2. The χ^2 goodness-of-fit metric

Six models are fit the KCT data by minimizing the Pearson X^2 statistic, the sum of squared deviations weighted by the inverse variance of each datum

$$X^2 = \Sigma_i\{(P_i - P_{\text{data}i})^2/(P_i(1-P_i)/N_i)\}, \quad (14)$$

where P_i is the predicted probability correct for each datum (Eq. (13)), $P_{\text{data}i}$ is the experimental data expressed as probability correct and N_i is the total number of trials at each datum. The denominator of Eq. (14) is the estimate of the variance of the datum based on the binomial distribution. It is standard in this type of probit analysis to use the model prediction P to estimate the variance because it is expected to have a lower variance than P_{data}. The Pearson X^2 statistic that is based on binomial variability, is typically very close to the chi-square (χ^2) statistic based on Gaussian variability. Fig. 3 of GM compares X^2 to χ^2. They find that the two statistics are well matched except for one observer who has several data points very close to unity as they discuss.

3. Six models for fitting the KCT contrast discrimination data

The first three models that we examine are based on the one parameter response function of Eq. (1) (or Eq. (11)).

Model 1 to be called GM3 (called FIX3 by GM) with three free parameters b, w and f with $q = 0$. The contrast Weber fraction is given by Eq. (8) to be $W = b\ cn^{-w}$ where the normalized reference contrast, cn is:

$$cn = c_{\text{ref}}/0.3 \quad (15)$$

Model 2 to be called KCT4 (called VAR4 by GM) with four free parameters: b, w, q, and f. The parameter q is a measure of multiplicative noise, with $q = 0$ representing zero multiplicative noise. Model 1 is a special case of Model 2 obtained by setting $q = 0$ in Eq. (11).

Model 3 to be called K4 adds an extra finger error parameter to GM3. The four parameters are w, b, f_{low}, and f_{high}. Model 1 is a special case of Model 3 with $f_{\text{low}} = f_{\text{high}}$ (symmetric finger errors). In addition to the above three models based on a one parameter contrast response function we will also consider two models with a two parameter contrast response function.

Model 4 to be called GM4 (called FIX4 by GM) adds an extra parameter to GM3 by using the Stromeyer–Foley (Stromeyer & Klein, 1974; Legge & Foley, 1980; Foley, 1994) contrast response function with the denominator exponent fixed at 2 (same choice as Stromeyer & Klein, 1974). The four parameters are w, b, ss, and f.

$$r(c) = (ss+1)\ cn^{w+2}/(ss + cn^2). \quad (16)$$

$r(c)$ is normalized so that its value is unity when the contrast is 30%, just like Eq. (8). The denominator of Eq. (12) uses the variable ss which is the square of the GM parameter s. We find it is better to use ss because of its better behavior near zero. For SV-S, the value of ss is slightly negative, making the GM4 parameter s imaginary. The value of the Weber fraction at $c = 0.30$ is b as before, however the log–log slope is slightly different from w, as will be mentioned in connection with Table 2. Model 1 is a special case of Model 4 with $ss = 0$.

Model 5 to be called K4.25 adds a quarter of a parameter to K4. Rather than using the Stromeyer–Foley function (Eq. (16)) we allow the exponent w to vary linearly with contrast. Thus, Eq. (11) becomes:

$$r(c) = cn^{w+w1^*(cn-1)}. \quad (17)$$

The 4 full parameters are, w, b, f_{low}, and f_{high}, the same as model 3. The extra parameter, $w1$ was fixed at $w1 = 0.01$ for all four observers. It was fixed because it is strongly correlated to other parameters and allowing it to float would have introduced large standard errors in other parameters. Since the same value of $w1$ was used for each of the four observers we count it as a 1/4 degree of freedom per observer. It turns out that the Weber fraction at $c = 0.3$ ($cn = 1$) is b and the log–log slope is $-w$, just as for Models 1, 2, and 3. Model 3 is a special case of Model 5 with $w1 = 0$. The final model fits the contrast response function with as many parameters as there are test levels.

Model 6 is called KTS because of its similarity to the data fitting approach of Katkov, Tsodyks, and Sagi (2006a, 2006b). Instead of parameterizing the contrast response function, $r(c)$ with a small number of parameter they allow it to be totally free so there are as many parameters as the are contrast levels being tested in the entire experiment, covering all three reference levels. In our fit, we constrained $r(c)$ to be monotonically increasing (no negative slope). The KCT dataset does not have sufficient data to allow σ to be similarly unconstrained, so we assumed the multiplicative noise had the form

$$\sigma(c) = r(c)^q, \quad (18)$$

similar to Eqs. (2) or (12) except that the normalization, k, is incorporated into $r(c)$ rather than into $\sigma(c)$, thereby making the free parameterization of $r(c)$ simpler. Because this model is so different from the other models further details regarding the fitting procedure will be discussed in the Discussion, in connection with claims made by Katkov et al. (2006a, 2006b).

4. Results

The main results of this study are presented in Fig. 1 and in five tables. The left panels of Fig. 1 show the data for the four subjects with the blue, red and black data and curves corresponding to $c_{\text{ref}} = 0.15$, 0.30, and 0.60. The abscissa is the total contrast of the test pattern $c_{\text{test}} = c_{\text{ref}} + \Delta c_{\text{test}}$. The ordinate is the percentage of times that c_{test} is judged to be

greater than c_{ref}. The solid curve is the KCT4 prediction that does the best job of fitting the data. The right panels show the difference between the data and the curve. It is provided to show a magnified view of the quality of the fit of the KCT4 model to the data.

The four columns on the left side of Table 1 presents b, the contrast Weber fraction (in %) at $c_{ref} = 0.3$ (see Eq. (8)), for the five models and four subjects. For a Weber fraction of $b = 21\%$, as found in subjects AK-S and SV-T the contrast discrimination threshold is 0.063 (21% * 0.3). The bottom two rows of Table 1 are the medians and standard deviations across the models for each of the subjects. We report the median rather than the mean in this and subsequent tables because of the presence of outliers (see AK-T's results). The standard deviation (bottom row of Table 1) is calculated across the five models without weighing. Except for the third subject (AK-T), the standard deviations across the five models (0.8, 0.3, 3.8, and 1.0%) are compatible with the estimated SEs of b (the number to the right of ± in the tables).

The right four columns of Table 1 present the Weber fraction of the same five models for a reference contrast of 100%. The advantage of presenting the reference contrast at 100% is that the Weber fraction is equal to k/p, where k and p are two of the parameters used by KCT and by GM. The discussion of Table 2 discusses how to connect p to our parameter w, and then the right half of Table 1 provides the information for obtaining k. Whereas k varies across the models by more than a factor of 500% for observer SV-S (Georgeson & Meese, 2006), the ratio k/p is relatively constant (about 2% in Table 1 for SV-S, our particularly striking example). Table 1 shows the dramatic advantage of using a parameter (b rather than k) that has a direct connection to the raw data.

The parameter values and X^2 values agree with the values found by KCT and GM for models in common (KCT4, GM3, and GM4). The standard errors of the parameter estimates were calculated from the inverse of the Jacobian matrix (Press, Flannery, Teukolsky, & Vetterling, 1992, Numerical Recipes) that is output by Matlab's non-linear regression program "lsqnonlin". The standard errors (SEs) are close to the KCT4 values that were obtained by KCT using Monte Carlo simulations (GM did not calculate SEs). The question of whether the SEs calculated in this manner are trustworthy (not always, will be the answer) is one of the important questions to be taken up in the Discussion. The relatively small standard deviation of the mean across the five models (bottom row of Table 1) for all subjects provides a validation for the robustness of the contrast Weber fraction in the mid-contrast range.

Table 2
w (log–log Weber slope at $c_{ref} = 30\%$)

Model	AK-S	SV-S	AK-T	SV-T
GM3	.64 ± .05	.44 ± .05	.63 ± .06	.34 ± .06
KCT4	.56 ± .05	.40 ± .05	.42 ± .06	.30 ± .05
K4	.61 ± .05	.43 ± .05	.46 ± .07	.35 ± .06
GM4	.74 ± .38	.63 ± .04	.81 ± .05	.56 ± .04
K4.25	.53 ± .05	.36 ± .07	.050 ± .010	.18 ± .18
Median	0.61	0.43	0.46	0.34

Table 2 is similar to Table 1 but for the log–log Weber slope, $-w$. Note that the log–log slope of the contrast response function, r, is $1-w$. If contrast discrimination followed a perfect Weber's law then $w = 0$. The value of w for models GM3 and K4 (the simple power law models) is in good agreement with the prediction from Table 1 that presents the Weber fractions at reference contrasts of 30 and 100%. For example, a two-fold decrease in Weber fraction over a 3.3-fold increase in reference contrast gives a log–log slope of $w = \log(2)/\log(3.3) = 0.58$, in approximate agreement with the data of AK-S. Similarly the Weber ratio from Table 1 of $21.2/14 = 1.5$ for SV-T corresponds to $w = \log(1.5)/\log(3.3) = 0.34$, in agreement with the Table 2 value. The first three models (GM3, KCT4, and K4) with power law contrast response functions, have log–log slopes, w, that are within the error bars of each other. The last two models, with non-constant log–log slopes have values of w that are model dependent. For the GM4 and K4.25 models the parameter w provides an imperfect estimate of the log–log Weber slope at $c_{ref} = 0.3$. That is because we did not introduce a correction for the deviation of the function $r(c)$ from a power function. This explains the downward bias of w for K4.25 and the upward bias for GM4. The slope parameter for observer AK-T is especially sensitive to the model differences, even for the relatively stable results for b in Table 1, for reasons to be taken up in Section 5.

Table 1
The contrast discrimination Weber fraction in % is given for the four observers and five models

	b (percent Weber Fraction at $c_{ref} = 30\%$)				k/p (percent Weber Fraction at $c_{ref} = 100\%$)			
Model	AK-S	SV-S	AK-T	SV-T	AK-S	SV-S	AK-T	SV-T
GM3	20.6 ± 0.5	11.6 ± 0.3	26.8 ± 1.0	21.3 ± 0.7	9.5 ± 0.6	6.8 ± 0.4	12.5 ± 1.0	14.1 ± 1.0
KCT4	19.4 ± 0.7	11.3 ± 0.5	25.3 ± 1.0	20.0 ± 0.9	10.0 ± 0.5	6.9 ± 0.4	15.3 ± 1.0	14.0 ± 0.8
K4	21.1 ± 0.5	11.6 ± 0.4	28.1 ± 1.0	21.0 ± 0.7	10.1 ± 0.7	6.9 ± 0.6	16.0 ± 1.3	13.8 ± 1.1
GM4	20.8 ± 0.6	11.5 ± 0.4	29.3 ± 1.2	21.2 ± 1.0	9.3 ± 0.6	6.9 ± 0.6	13.0 ± 1.4	14.3 ± 1.5
K4.25	21.7 ± 0.6	12.1 ± 0.4	35.1 ± 1.2	22.8 ± 1.7	10.3 ± 0.7	6.9 ± 0.5	16.3 ± 1.3	14.0 ± 1.1
Median	20.8	11.5	28.1	21.2	10.0	6.9	15.3	14.0
Stand dev	0.8	0.3	3.8	1.0	0.42	0.04	1.7	0.18

The left half of the table is for a reference contrast of 30% (parameter b) and the right half is for $c_{ref} = 100\%$ (parameter ratio k/p). The median and standard deviation of the mean of the five models are shown at the bottom.

Remarkably, the values of w, defined at $c_{ref} = 0.30$ have a simple connection to the slope parameter, p, used by KCT and GM. The connection of our parameter w to the KCT4 exponent p is given by Eq. (9) to be $p = w/(1-q)$ where q is the multiplicative noise exponent to be shown in Table 4. The connection between w and the GM4 exponent, p can be seen from Eq. (16) to be $p = w + 2$, using the approximation that ss is very small. Comparing these estimates of p with the values reported by GM shows that the approximations are excellent except for SV-S whose value of ss is the largest of the observers (see Table 4).

Table 3 presents the finger error parameter, f_h. For the models used by GM (GM3, KCT4, and GM4) this parameter specifies the percentage of errors at the low and high asymptotes. A value of 2% implies that for 200 trials, 4 finger errors would be expected for discriminating 30% and 0% and also for discriminating 30 and 100%. For these three models the finger error parameter is half the parameter, g, reported by KCT and GM, since their parameter specifies the estimated percentage of times at which the subject makes a totally random guess. The actual errors are half the guessing rate given that the correct answer would be obtained 50% of the time by chance. The values of g, reported by GM is constrained to have an upper limit of 5% (personal communication from Mark Georgeson) corresponding to $f_h = f_l = 2.5\%$. The results for AK-T should surprise the reader. Model KCT4 shows 0.2% finger errors whereas model GM3 shows a 4.6% finger error rate. The result is surprising because one would think that finger errors are determined by the raw data in a model independent manner. This clue to funny-business in the AK-T data set will become important to our analysis of the KCT data. For models K4 and K4.25 the finger errors for comparing the 30% reference data to very low and very high contrasts are allowed to differ. For these two models, parameter 3 (f_h) specifies the finger errors at the upper asymptote and parameter 4 (f_l) (see Table 4) will represent the low asymptote finger errors.

Table 3
f_h (percent finger error, high tail for K4 and K4.25 and symmetric for others)

Model	AK-S	SV-S	AK-T	SV-T
GM3	0.9 ± 0.2	0.7 ± 0.2	4.6 ± 0.5	2.2 ± 0.4
KCT4	0.0 ± 0.2	0.6 ± 0.3	0.2 ± 1.0	1.4 ± 0.6
K4	0.3 ± 0.2	0.7 ± 0.3	2.7 ± 0.5	2.5 ± 0.5
GM4	0.9 ± 0.2	0.7 ± 0.2	3.3 ± 0.5	2.1 ± 0.4
K4.25	0.3 ± 0.2	0.7 ± 0.3	2.5 ± 0.5	2.3 ± 0.5
KTS	$0 \pm ??$	0.2 ± 0.4	$1.2 \pm ??$	1.4 ± 0.4
Median	0.3	0.7	2.6	2.2

Table 3 includes information on a sixth model, based on the Katkov et al. (2006a, 2006b) approach of letting the contrast response function be totally free, except for a monotonicity constraint. The finger error parameter is taken to be symmetric for the low and high asymptotes. A singular matrix precluded estimation of the SE of the finger error for AK-S and AK-T. Otherwise the fit was excellent as will be discussed.

Table 4 presents parameter 4, the extra parameter that is added to the three parameter model GM3. For the KCT4 and KTS models the extra parameter is q, the multiplicative noise exponent (Eqs. (2) and (9)). For model GM4 is it ss, the Stromeyer–Foley function saturation parameter (Eq. (14)). For models K4 and K4.25 it is a second finger error parameter.

KCT4 and KTS are the only models for which q differs from zero. The main goal of the KCT, GM, and present articles is to determine the value of q. In particular, we seek to determine whether the null hypothesis of $q = 0$ can be rejected with confidence. Table 4 presents KCT4 z-score values for q ($z_q = q/SE_q$) in order to test the null hypothesis. The values of z_q range from 10 to 38. The weakest rejection of the null hypothesis is for SV-S whose z-score is $z_q = 0.83/0.08 = 10$ corresponding to a χ^2 (z^2) of 100 with 1 degree of freedom. If this extremely powerful rejection of the null hypothesis ($p < 10^{-20}$) were believable, the KCT claim that multiplicative noise is present in the JND task would be strongly confirmed and the GM claim (''jury still out'') would be incorrect. However, the Discussion will discuss why the small SEs for q in Table 4 are not to be trusted and are not relevant for testing the null hypotheses. This turns out to be an example where non-linear regression differs strongly from linear regression. It should be noted that the error bars for b, w, and f_h in Tables 1–3 are trustworthy since those parameters are directly connected

Table 4
The last parameter of the five models other than GM3 are presented

Model	Parameter name	AK-S	SV-S	AK-T	SV-T
GM3	(there is no fourth parameter)				
KCT4	q (noise exponent)	$.76 \pm .03$	$.83 \pm .08$	$.84 \pm .03$	$.85 \pm .04$
	z-test $= q/SE_q$	25	10	38	21
K4	f_l (% finger error, low tail)	1.4 ± 0.4	0.9 ± 0.5	6.5 ± 0.7	2.0 ± 0.4
GM4	ss (saturation parameter)	$.05 \pm .43$	$.17 \pm .10$	$.16 \pm .10$	$.12 \pm .12$
K4.25	f_l (% finger error, low tail)	1.4 ± 0.4	0.8 ± 0.5	6.6 ± 0.7	2.0 ± 0.4
KTS	q (noise exponent) range	.15–.7	.15–.75	.5–.85	.1–.75

The parameters are: the multiplicative noise, q, for models KCT4 and KTS, the low tail 'finger error' parameter, f_l, for models K4 and K4.25, and the saturation parameter, ss, for model GM4. The z-score test for multiplicative noise in KCT4 is presented below the q values for that model. The standard errors are obtained from the output of the nonlinear regression fit.

to the data in a relatively model-free manner, as was discussed.

For model KTS, the q parameter was tested in a grid search rather than as part of the non-linear regression search that was done for the other parameters. Table 4 reports the range of q for which the X^2 of the fit is less than $X^2_{\min} + 4$, corresponding to a 95% confidence interval if the fit had been a linear regression.

For model GM4, parameter 4 is ss, the parameter that controls the saturation point of the Stromeyer–Foley contrast response function (see Eq. (14)). The third from-last row of Table 4 shows that *ss* deviates from 0 by less than two standard errors for all four subjects. Thus parameter 4 is not expected to have a large impact on the goodness-of-fit for model GM4. In any case, the Stromeyer–Foley function used in GM4, does not deviate strongly from a power function (the local power is always in the range of w and $w + 2$), so it does not provide a broad test of alternatives to multiplicative noise.

A statistically stronger alternative to multiplicative noise than model GM4 are the models K4 and K4.25 where the extra parameter is a second finger error parameter, needed to decouple the errors at the low and high asymptotes. The discussion leading up to Eq. (13) points out that subject AK's raw data (both sustained and transient conditions) have many more "finger errors" in discriminating a 30% reference contrast from very low contrasts than from very high contrasts. For models K4 and K4.25 the high and low contrast finger errors are 2.7% and 6.5%. A finger error rate of 6.5% implies AK-T is randomly guessing 13% of the time when discriminating 30% contrast from much lower contrasts. The Discussion will introduce explanations other than asymmetric random guessing for the asymmetric "finger errors", including a type of multiplicative noise.

Table 5 presents the X^2 values given by Eq. (14) as a goodness-of-fit measure for each observer and model. The values in parentheses are the relevant degrees of freedom (*df*). The *df* on the right half of Table 5 will be clarified in the Discussion. Since the X^2 (binomial noise) and χ^2 (Gaussian noise) distributions are almost identical (see Klein (2001) for where differences are expected) the mean and SE of the expected X^2 for a good fit are approximately: $df \pm \text{sqrt}(2\ df)$. Four different patterns are exhibited by the four observers. Consider first the X^2 values for model GM3, the 3 parameter reference model. Observer SV-S has X^2 almost identical to the degrees of freedom. This may have been expected from Table 1 that shows SV-S has a very low threshold with a Weber fraction about half of the other observers. Observer AK-T is at the other extreme with $X^2 \gg$ df. GM consider these large X^2 values to be a sufficient reason to disqualify that observer's data from consideration. We take the opposite position and conclude from the nature of AK-T's errors that AK-T is highly informative about the underlying model. In general, the parameter values found for AK-T are highly sensitive to the nonlinear regression initial conditions. The two other observers, AK-S and SV-T have X^2 values that are about $df + 13$ making the GM3 model slightly outside the 95% confidence zone. The X^2 values for SV-T are changed very little across the five models whereas AK-S has large changes when an extra parameter is added.

The KTS model (bottom row of Table 5) has far fewer degrees of freedom than the other models because the contrast response function had as many parameters as there were levels tested. The KTS fits to the four subjects are shown in Fig. 3. The format of Fig. 3 is the same as Fig. 1. In order to clarify the number of parameters going into the fits consider subject AK-T. With fit KCT4 subject AK-T has 34 data points with 4 parameters, giving $34 - 4 = 30$ degrees of freedom. With fit KTS there are 24 parameters and thus $34 - 24 = 10$ degrees of freedom. The 24 parameters consist of 22 effective test levels, one finger error parameter and one multiplicative noise parameter, q, that was determined in a grid search. There were 25 actual test and reference levels used in the fit, but the contrast response function had a negative slope at three of the 25 points so 3 parameters were frozen out, leaving 22 levels. The X^2 fits indicate that the unconstrained model provides a good fit to the data. This is not surprising for AK-S, SV-S, and SV-T, that were already well fit by a four parameter model. What is gratifying is that when the contrast response function is unconstrained (except for monotonicity) $X^2 = 19.0$ with 10 degrees of freedom, corresponding to $p = 0.04$. This is a respectable p value that

Table 5
The Pearson X^2 values are given on the left half of the Table for the fits of the six models for the four subjects

	X^2 (*df*)				$X^2(q=0) - X^2(q>0)$ (*df*) *p*-value			
Model	AK-S	SV-S	AK-T	SV-T	AK-S	SV-S	AK-T	SV-T
GM3	46.9 (33)	29.7 (29)	121.0 (31)	43.0 (29)	11.7(1) 0.0006	1.2 (1) 0.26	45.0 (1) 0.000	3.1 (1) .08
KCT4	35.3 (32)	28.5 (28)	76.1 (30)	39.4 (28)	0 (0)	0 (0)	0 (0)	0 (0)
K4	38.9 (32)	29.6 (28)	96.4 (30)	42.0 (28)	3.7 (1) 0.05	1.2 (1) 0.281	20.3 (1) 0.000	2.5 (1) 0.11
GM4	43.8 (32)	29.6 (28)	114.9 (30)	40.8 (28)	8.5 (1) 0.003	1.7 (1) 0.19	38.8 (1) 0.000	1.6 (1) 0.21
K4.25	36.5 (31.75)	30.7 (27.75)	77.0 (29.75)	41.1 (27.75)	1.3 (1.25) 0.32	2.2 (1.25) 0.19	1.9 (1.25) 0.23	1.7 (1.25) 0.26
KTS	3.1 (12)	4.0 (4)	19.0 (10)	13.2 (10)	5.6 (1) 0.0180	0 (1)	7.3 (1) 0.0069	4.8 (1) 0.0285

The degrees of freedom are in parentheses. The right half of the Table is the X^2 improvement obtained by allowing the multiplicative noise parameter, q, to float. For models GM3, K4, M4 and K4.25 the improvement is calculated as the X^2 value relative to KCT4. For KTS the improvement is obtained by comparing a fit with fixed q vs. a floating q. The p values of the significance of the improvement with 1 degree of freedom are underlined.

does not justify rejection of the fit, nor rejection of AK-T as a credible observer.

The test for whether multiplicative noise is needed to account for the KCT data (is q significantly greater than 0) can be seen in the values in the right four columns of Table 5 that give the change in X^2 between the four models without multiplicative noise relative to the KCT4 model with multiplicative noise. The first row on the right side compares GM3 to KCT4, namely the change in X^2 when the multiplicative noise parameter, q, changes from $q = 0.0$ to 0.8. The non-significant change in X^2 for SV-S and SV-T (1.2 and 3.1 respectively) means that these two observers were compatible with the null hypothesis (no multiplicative noise is needed). The Discussion will consider the question of whether for these two observers: (1) the experimental data had insufficient power to detect the presence of multiplicative noise, (2) there was very little multiplicative noise, or (3) the high sensitivity (low Weber fraction) of SV-S and SV-T made them poor observers for the KCT method of measuring multiplicative noise. The finding from Table 5 that SV-S and SV-T show no multiplicative noise contradicts the highly significant z-test value of the multiplicative noise parameter, q, that was shown in Table 4. This contradiction will be considered in the Discussion. On the other hand, observers AK-S and AK-T showed dramatic reductions in X^2 (11.7 and 45.0, respectively) when multiplicative noise was introduced, seeming to provide strong evidence that multiplicative noise is present. It's not that simple however, in that there is a model (K4.25) without explicit multiplicative noise that has an X^2 similar to KCT4. The presence of multiplicative noise for AK-S and AK-T depends on how multiplicative noise is defined, as will be taken up in the Discussion.

The p-values presented on the right side of Table 5 (below the X^2 values) are tests of the multiplicative noise hypothesis. Consider, AK-S. With $q = 0.76 \pm .03$, $X^2 = 35.3$ (model KCT4), and with $q = 0$, $X^2 = 46.9$ (model GM3) giving a difference of $\Delta X^2 = 11.7$ with $df = 1$, corresponding to a probability of $p = 0.0006$ of having such a large X^2 difference by chance (without multiplicative noise). GM introduced an extra parameter with their Stromeyer–Foley fit (GM4) that reduced ΔX^2 to 8.5 with a probability of $p = 0.003$. The comparison of models KCT4 and GM4 is unconventional since one is not embedded in the other. They both have four parameters, with three being shared. The logic of the comparison is to make-believe GM4 only had 3 parameters when calculating the degrees of freedom and that even with an extra (not counted) parameter, the X^2 value for GM4 is still 8.5 larger than that of KCT4. If we considered the comparison to have $df = 1$ (the value shown in Table 5) rather than $df = 0$, it would have a probability of $p = 0.003$. That is, the models suggested by GM are not able to provide as good a fit as the multiplicative noise fit. Models K4 and K4.25 provide acceptable fits to the AK-S data with ΔX^2 values of 3.7 and 1.3, respectively. The latter model, in particular has $p = 0.32$ for $df = 1.25$.

The extra quarter parameter fixed at $w1 = 0.01$ for model K4.25 gave a dramatic reduction in X^2 for AK-T that had $p \approx 0$ for all the other models. The K4.25 fit has no multiplicative noise according to the conventional definition of multiplicative noise (but see Discussion for an extended definition), yet it produces X^2 values that are very close to KCT's value, $\Delta X^2 = 1.9$, giving $p = 0.23$ for $df = 1.25$. This fit to all the subjects' results (albeit with a high X^2) without conventionally defined multiplicative noise provides a strong argument in favor of "the jury is still out". However, in order to achieve the K4 and K4.25 fits we used a pair of finger error parameters. The Discussion will take up the question of whether introducing a second finger error parameter is a subtle way of introducing a novel form of multiplicative noise.

All entries on the right half of Table 5 represent the X^2 difference ($X^2(q = 0) - X^2(q > 0)$) due to allowing the noise to vary. For the first five models this was accomplished by comparing the models to KCT4, the model with multiplicative noise. For KTS, this was accomplished by comparing the fit with $q = 0$ to the fit with $q > 0$ that had minimum X^2. The bottom row shows that the optimal q is significantly greater than zero ($p < .05$) for AK-S, AK-T and SV-T. For SV-S, $X^2 = 4$ independent of q, because AK-S's thresholds were too low for the non-linearities needed by the KCT method to work. The AK-T data had $X^2(q = 0.8) = 19.0$ and $X^2(q = 0) = 26.3$ ($\Delta X^2 = 7.3$) which is able to reject $q = 0$ at the $p = .01$ level for 1 degree of freedom. Of equal interest is that even though only symmetric finger errors were put into model KTS, Fig. 3 shows that the predicted psychometric functions have asymmetric asymptotes. This is partly due to the KCT mechanism of multiplicative noise, present in the optimal KTS fit, and partly due to the asymmetric contrast response function.

5. Discussion

5.1. Misleading z-test or t-test for hypothesis testing

One of the lessons to be learned from fitting the KCT data is that one must be careful when testing hypotheses based on non-linear fits to data. The second row of Table 4 seems to unambiguously say that the no-multiplicative noise null hypothesis of $q = 0$ can be resoundingly rejected. The values of q are 0.76 ± 0.03, 0.83 ± 0.08, 0.84 ± 0.03, and 0.85 ± 0.04 for the four observers. Even when a heterogeneity factor (see the forthcoming section "Dealing with outlying data") is used to convert the z (χ^2) test to a $t(F)$ test, the $q = 0$ hypothesis is soundly rejected. However, when one actually does the fit with $q = 0$ (model GM3) one finds reasonable X^2 values, especially when some flexibility is given to the contrast response function shape. How is this possible? The answer, given by GM in their Fig. 2, is that because of the strong non-linearity of the fitting function, except for AK-T, only when q gets close to unity does q have much of an effect on the value of X^2. The implication is that with non-linear regression one must always be prepared for surprises regarding the t test for the significance of parameters. One should explore how X^2

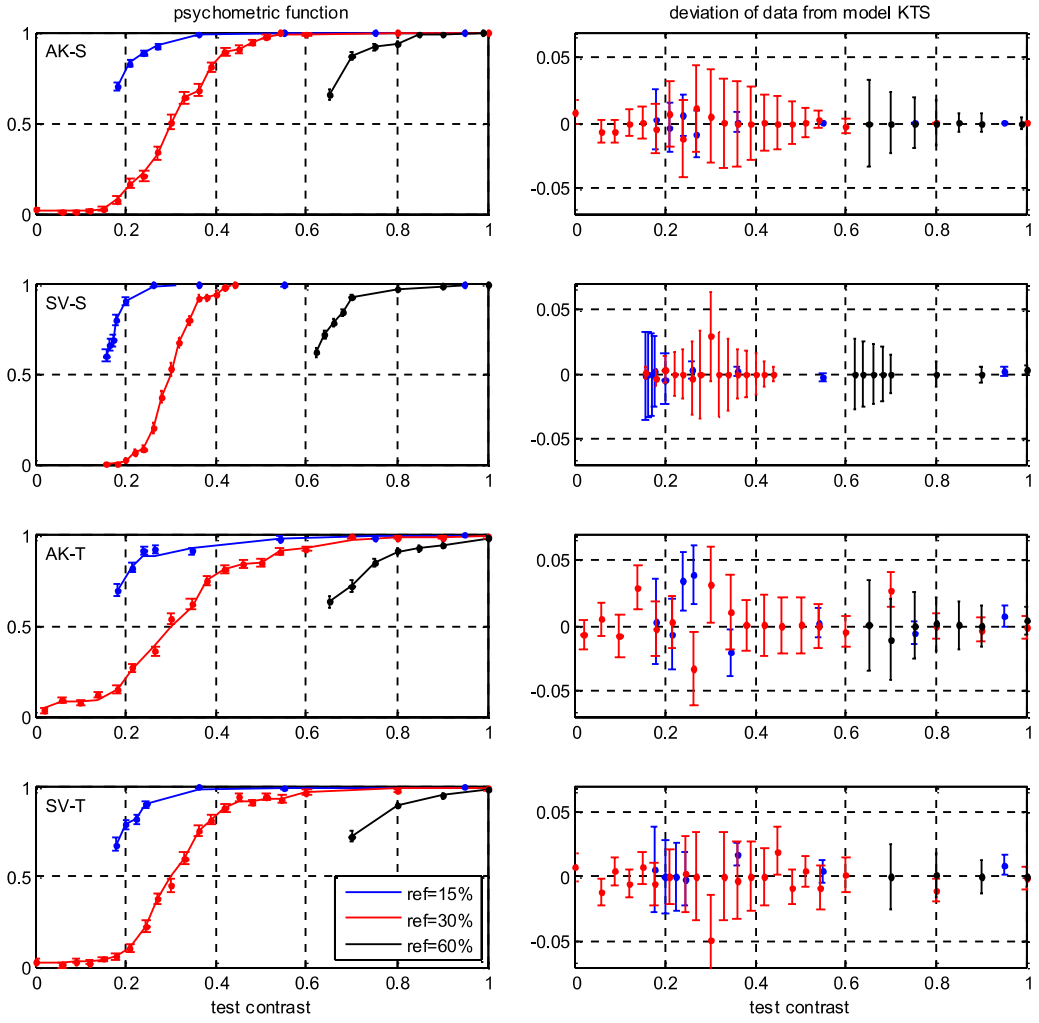

Fig. 2. The panels are the same as Fig. 1 except that instead of the KCT4 as the model, the non-parametric KTS model is used. The model with best fitting value of q (see Table 4) is shown in the left panels. The right panels show the deviations between data and model. The KTS model has almost as many parameters as data (see Table 5) so many of the deviations between data and model are zero.

depends on parameters using a grid search before making strong conclusions regarding hypothesis testing!

5.2. Asymmetric finger errors

A significant asymmetry was found in fitting the low and high contrast tail of AK-S's and AK-T's data for the 30% reference contrast. The K4 fit for AK-T has a lower asymptote (Table 4) of $f_l = 6.5 \pm 0.7\%$ and an upper asymptote (Table 3) of $97.3 \pm 0.5\%$ ($f_h = 2.7$). For AK-S the finger error asymmetry is: $f_l = 1.4 \pm 0.4\%$ and $f_h = 0.34 \pm 0.4\%$. What might be causing this asymmetry? It cannot be random errors since true finger errors would be approximately equal at the low and high branches of the discrimination curve. In order to get insight into what might contribute to the large number of errors at the low asymptote, I replicated that condition on myself. I did a 2AFC discrimination of 30% from 10% contrast in the transient condition (16 Hz counterphase flicker of a 3 c/deg grating). I made zero errors out of 400 trials. It was a very easy task. AK-T on the other hand made 51 errors out of 772 trials for distinguishing 30% from <10%. From that experience I concluded that what we have been calling 'finger errors' probably has a more interesting explanation than simply being 'errors'.

I can think of four underlying reasons for the asymmetric asymptotes:

Reason #1. *Multiplicative noise.* It is the nature of the 2AFC multiplicative noise mathematics that there are asymmetric asymptotes. The KCT4 fit with multiplicative noise shows that one can have a relatively steep psychometric function associated with asymmetric asymptotes, even though the explicit finger errors in the parametric fit were symmetric. The shape of the psychometric function at low contrasts is flat, thus appearing to be asymptotic, because of the accelerating contrast response function near detection threshold.

Reason #2. *Shallow psychometric function.* The GM3 and GM4 fits are able to fit the elevated low contrast

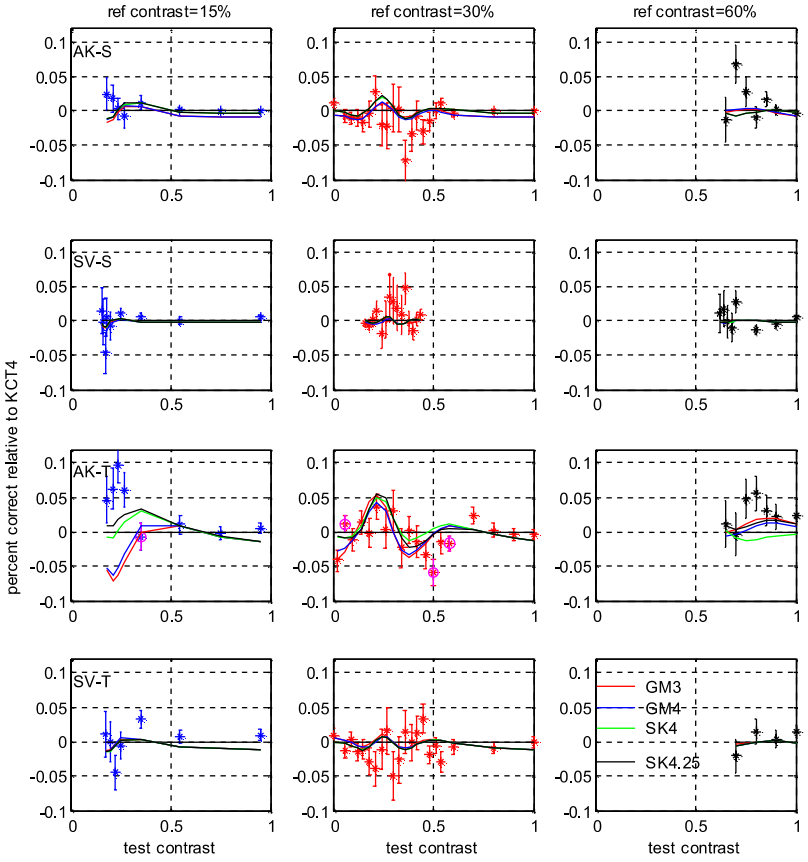

Fig. 3. The panels are similar to the right panels of Fig. 1 with the data minus the KCT4 fit being plotted. For clarity of viewing, the data for the three reference contrasts have been plotted separately in the three columns. The solid lines are the difference of the four models minus the KCT4 model, with the color given by the legend. One of the most striking things shown by the plots are that the five models are much closer to each other than they are to the data. That finding suggests that there are alternative models, such as KTS, that can do a better job of fitting the data. In exploring possible reasons for the less than perfect fit to AK-Ts data we explored the consequences of removing four data points where there was a big jump from one datum to the next. Those four data are shown in magenta, using a circle symbol on top of the asterisk. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this paper.)

asymptote by a different strategy. Fig. 1 of Georgeson and Meese (2006) shows that the GM3 fit (red dashed line) has a shallower slope than the KCT4 fit. The same is true for the GM4 fit as can be seen in our Table 1 and our Fig. 3. Fig. 3 is similar to the right panels of Fig. 1 except that the panels are separated into the three reference contrasts and the four parametric fits are plotted relative to the KCT4 fit. Table 1 shows that the KCT4 fit has the smallest Weber fraction and thus the steepest psychometric function among the parametric fits. The shallower psychometric functions of GM3 and GM4 are able to produce an asymmetry because on the low side the psychometric function has not yet reached its asympotote. Fig. 3 shows the asymmetry graphically. The data points and curves in Fig. 3 are plotted relative to the KCT4 fit, the same as in the right panels of Fig. 1. The panel for AK-T with $c_{ref} = 0.3$ is especially interesting. Near $c_{test} = 0.3$ the curves have a negative slope indicating that the four models other than KCT4 have a shallower slope (larger Weber fraction) than the KCT4 fit. For $c_{test}<0.2$ however, the slope is positive, indicating that the asymptotic region has not yet been reached Although this broader psychometric function does a reasonable job of fitting the $c_{ref} = 0.3$ data, the increased Weber fraction causes the $c_{ref} = 0.15$ data to be poorly fit as can be seen in both Figs. 1 and 3. The $c_{ref} = 0.6$ data also is not well fit because the psychometric function is not sufficiently steep. That is, the poor GM3 and GM4 fits for $c_{ref} = 0.15$ and 0.60 are due to the need to fit the $c_{ref} = 0.30$ data with a relatively shallow function to account for the asymmetric asymptotes.

Reason #3. *Heterogeneous intermixing of test contrasts.* I suspect that subject AK made a large number of errors at low test contrasts because of the non-blocked nature of the KCT paradigm. The experiment I did on myself mentioned at the beginning of this subsection had only a single 10% vs 30% condition. Rather than this blocked design of discriminating 30% contrast from a single low contrast, KCT also intermixed trials in which 30% had to be discriminated from very high contrasts. Two possibilities for how the presence of high contrasts in preceding trials could produce errors in discriminating 30% from low contrasts are:

(a) The high contrasts could adapt the visual system making the 30% reference appear less visible. A Personal Communication from L. Kontsevich regarding this hypothesis is as follows:

"During a lengthy discussion with the author a few years ago I ran a control experiment (with myself as an observer) that resolved the issue of asymmetric finger errors. In one condition I measured contrast detection threshold with blank field between the trials. In the other condition I presented on every other trial a 90% contrast stimulus in one of the intervals (which produced 100% correct response rate). The finding was that the presence of the high-contrast stimuli had negligible effect on the threshold. In the data presented in KCT the effect (if there was any) should be even smaller. The implication from this experiment is that the conjecture made by the author that the asymmetric finger errors are due to effect of high-contrast stimuli onto low-contrast ones across trials is most like false."

One possibility is that even though the intermixed high contrast stimulus had no effect on detection threshold it might have reduced the perceived contrast of the 30% reference. Another possibility is that there are individual differences, since observer SV showed no finger error asymmetry.

(b) The high contrasts could produce afterimages making low contrast test patterns appear to be of higher contrast, thus getting confused with the 30% flickering reference.

The possibility that these two factors could produce more errors at the low asymptote than at the high asymptote provides a rationale for our introduction of an extra 'finger error' parameter in models K4 and K4.25. Intermixing very different test contrasts can also have an effect on the shape of the contrast response function. One could worry that the increasing transducer exponent in the KCT4.25 fit produces an unreasonable shape of the TvC function whereby at contrasts above 50%, the TvC function saturates sharply. Past findings (Legge & Foley, 1980; Foley & Legge, 1981), on the other hand, found TvC functions with relatively constant log–log slopes. However, TvC functions are normally measured in a blocked design with a narrow range of contrasts being discriminated. It is possible that under conditions where the test contrast is roving trial to trial, from above 90% to below 10%, such as the test contrasts in the KCT experiments, the TvC shape will depart from that measured in a blocked design, making the psychometric function shapes of KCT4.25 and KTS possible.

Reason #4. *Attentional factors.* Intermixing the high and low test contrasts could cause the subject to pay less attention to low and moderate contrasts. That could cause the 30% reference to sometimes be missed, which would produce "finger errors" of the sort demonstrated by AK. It could also be that when a high contrast stimulus is present in one of the intervals the subject "wakes up", pays attention and makes less finger errors. This effect could also be present in blocked presentations in that discriminating barely visible stimuli might be less attention grabbing than discriminating high contrast stimuli. The last part of this Section 5 will point out that finger errors contingent on the contrast of the stimulus can be considered to be a form of multiplicative noise.

5.3. Alternative shapes of the contrast response function: The Stromeyer–Foley function

Stromeyer and Klein (1974) introduced a set of equations for fitting contrast discrimination data that were almost identical to those used by GM:

$$d'(c) = (a+1)(c/\mathrm{th})^{t+2}/(a+(c/\mathrm{th})^2) \tag{19}$$

and the multiplicative noise standard deviation was assumed to have the form (Stromeyer & Klein, 1974):

$$\sigma(c) = 1 + kd'(c)^q \tag{20}$$

Eqs. (19) and (20) are a combination of models KCT4 and GM4 with both multiplicative noise and also a non-polynomial response function. The only difference between Eqs. (19) and (16) is that the former is normalized to unity at the detection threshold (th = detection threshold), the GM4 normalization has th = 100%, and our implementation of GM4 uses a 30% reference contrast (th = 0.3). By normalizing the contrast response function to unity at the detection threshold Stromeyer and Klein (1974) were able to refer to it as $d'(c)$ rather than $r(c)$. Eq. (20) for the noise standard deviation $\sigma(c)$ differs from Eq. (2) by having a constant of unity in front so that at the lowest contrasts the noise variance is a constant value of unity rather than going to zero at zero contrast. Eqs. (19) and (20) have the advantage that they can be used for detection as well as for contrast discrimination. Foley (1994) and Foley and Legge (1981) have made extensive use of the form in Eq. (19) so that function has been called the Stromeyer–Foley function (Klein, 2001). Stromeyer and Klein (1974) found a good fit to their data with $t = 2$, $a = 1.5$, $k = 0.25$, $q = 1$ and the threshold, th, depended on the observer. Parameters t and q are remarkably close to the parameters of the KCT4 fit. Stromeyer and Klein used multiplicative noise rather than a saturating response function because multiplicative noise near threshold with $k = 0.25$ was known from other experiments (Nachmias & Kocher, 1970). It may well be that when improved methods are developed for identifying the mechanisms underlying contrast discrimination, it will be found that both multiplicative noise and contrast response saturation will be important factors for placing limits on contrast discrimination, as was anticipated by Stromeyer and Klein (1974).

5.4. The Katkov, Tsodyks and Sagi analysis of 2AFC. Monotonic contrast response function

Katkov et al. (2006a, 2006b) fit 2AFC contrast discrimination data by allowing both the contrast response function, $r(c)$, and the noise, $\sigma(c)$, to have arbitrary shapes rather than the parametric power function shapes such as

seen in Eq. (19). KTS report that when fitting 2AFC contrast discrimination data with general $r(c)$ and $\sigma(c)$, there is often a singularity whereby a wide variety of functions $r(c)$ and $\sigma(c)$ can fit the same data. KTS imply that the possibility of a singularity makes it difficult to use 2AFC to distinguish multiplicative from constant noise. Two topics will be discussed here: (a) is there an intrinsic singularity for near-constant noise models? (b) How does a KTS type fit work for the KCT data?

The issue of a singularity is somewhat tricky in that it is well known (see KCT) that there is a fundamental trade-off between saturation of the contrast response function and multiplicative noise. That is why so much time is being spent on this topic by multiple authors. For *steep* psychometric functions, as KCT and GM (see discussion associated with Fig. 6 of Georgeson & Meese, 2006) point out, there is indeed a "singularity" making it impossible to isolate the cause of the rising TvC function. However, when the psychometric function is less steep, KCT and GM (see their Fig. 6 of Georgeson & Meese, 2006) note that it is feasible to determine the amount of multiplicative noise. The analyses in the Appendices of Katkov et al. (2006b) involve throwing away higher order terms of the Taylor's series expansion of the fundamental equation (Eq. (3))

$$d'_{2\mathrm{AFC}} = (r(c_1) - r(c_2))/((\sigma(c_1)^2 + \sigma(c_2)^2)/2)^{1/2}.$$

Discarding those higher order terms is responsible for the singularity, and is equivalent to being in the regime of steep psychometric functions.

In order to check the KTS claim of a singularity (see Katkov et al. (2006b) Table 1, line 2) for constant noise models I carried out a number of simulations of constant noise models using idealized data with three reference stimuli similar to the KCT data. Contrary to the KCT data, for each of the three reference contrasts there were matching test contrasts. Rather than fitting the data with parametric functions I allowed the contrast response function to have as many parameters as there were contrast levels. Thus the contrast response function was totally unrestrained. I examined a variety of noise functions (sigma) including sigma constrained to be (a) monotonic, (b) monotonic except near the ends of the range and (c) totally unconstrained. Noise was added to the idealized data assuming standard errors of sqrt(p(1-p)/N) with N either 200 or 2000. I used reasonably shallow psychometric function shapes similar to that of AK-T. I found no evidence for a fundamental singularity in the case of a constant noise model. That is, the randomness produced by the binomial noise in the data is no different for a constant noise model or a multiplicative noise model and no different from what is expected based on our χ^2 analysis. One possibility for the difference between our claims and those of Katkov et al. (2006a, 2006b) is that they used a sum of square error as their cost function rather than a cost function with binomial error weighting. Another possibility is that their psychometric functions were too steep, similar to SV-S as discussed in the preceding paragraph. The issues brought up by Katkov et al. (2006a, 2006b) are important for fitting this type of data. It is important to be aware of conditions like steep psychometric functions that can produce singularities in parameter estimation.

As was discussed earlier, I was not deterred by the possibility of singularities so I fit the KCT data using a KTS type model where r(c) was allowed to be an arbitrary monotonically increasing function. The KCT data have the problem that only three reference contrasts were used (15, 30, and 60%) and very few test contrasts were duplicated, as would be needed for constraining both $r(c)$ and $\sigma(c)$. I handled this insufficiency of data relative to parameters by two means:

(i) Choose $\sigma(c)$ to be a parametric function as given by Eq. (18) : $\sigma(c) = r(c)^q$.
(ii) A few test contrast levels were shifted very slightly to bring them into alignment across different reference contrasts. For example, for AK-T for the 15% reference, test contrasts were shifted from 18, 27, and 55%, to 18.5, 27.5, and 54%, while for the 30% reference, test contrasts were shifted from 19, 28, and 54% to 18.5, 27.5, and 54%. The alignment at 54% was put in the datum with the 15% reference contrast rather than going to 54.5%, because it was fully in the asymptotic regime so a shift there is inconsequential. The left half of Table 5 and its associated discussion shows that the KTS type of unconstrained fit to $r(c)$ does a good job of fitting the data, including AK-T data where the five parametric fits had unreasonably high X^2. This good fit was achieved even with the constraint of $\sigma(c) = r(c)^q$. As was stated above we were unable to let $\sigma(c)$ be free since in that case there would be more parameters than data. The right half of Table 5 shows that for both AK-S and AK-T the $q = 0$ constraint does a significantly worse job of fitting the data than does allowing for multiplicative noise ($q > 0$). The KCT dataset had insufficient overlap of test contrast levels for the three reference contrasts to be able to pin down the exponent q with any precision when $r(c)$ is allowed to be limited just by a monotonicity constraint. So on that score, the jury is still out.

5.5. Dealing with outlying data

It sometimes happens that parametric models do not adequately fit data, as was the case with the AK-T data (3rd column of Table 5). This section explores the reasons for the poor fit. It is worth examining the details of the AK-T fit since it has the shallowest psychometric function and thereby, according to GM's analysis, it would be the most sensitive of the four datasets for determining the presence of multiplicative noise.

The next to last column of Table 5 shows that the strongest parametric models evidence in favor of multiplicative noise comes from observer AK-T because of the dramatic decrease in X^2 of 38.8 between GM's best model (GM4) and the multiplicative noise model (KCT4). A similar decrease in X^2 was found for model K4.25 model without explicit multiplicative noise, but with a strong "finger error" asymmetry that may itself be an indication of a type of multiplicative noise, as will be discussed in the last section of this paper. In order to maintain their position that "the jury is still out" GM argue that the results of subject AK-T should not be given much weight because the X^2 values are too high for all five parametric models ($X^2 > 76$ for 30 degrees of freedom). We present three approaches to counteract GM's rejection of the AK-T dataset.

The first approach was to abandon using a parametric fit for $r(c)$, as discussed in the preceding section on the KTS fit. We found that by removing constraints on $r(c)$ other than monotonicity, the AK-T dataset was able to be fit with a reasonable X^2 value, while still providing evidence for multiplicative noise. The remainder of this section will discuss issues concerning the parametric fits of $r(c)$.

The second approach comes from looking closely at the AK-T data in Fig. 3 and noticing that there are both systematic and random sources of the large X^2 values. Fig. 3 shows the deviations of the data and models from the KCT4 model. The *systematic* deviations are best exemplified in the first four data points for the 15% reference contrast (left panel of the AK-T data). These four data are close to each other and strongly deviate from the KCT4 model prediction (the best of the 5 predictions in terms of X^2). The closeness of the data to each other suggests that the data are valid and should not be down-weighted in comparing models. The fifth datum, on the other hand is suspicious. It deviates from the neighboring data by several standard errors. The same is true for the 2nd, 13th, and 15th data for the 30% reference contrast (AK-T, middle panel). These four deviant four data with large random errors are circled and in magenta. When these four aberrant data are removed from the fit, the X^2 value for the K4.25 fit is reduced from 77.0 to 42.9 corresponding to a p value of $p = 0.009$, for 23.75 degrees of freedom. Further, removal of datum 3 of the 15% reference contrast produces a further reduction to $X^2 = 34.4$ with 22.75 *df* corresponding to $p = 0.05$. We are aware of the dangers of trimming data in this manner, which is why we made the distinction between systematic errors and random errors. The importance of aberrant data points should not be underestimated. The relative goodness of the K4.25 fit is largely dependent on how well it fits the 2nd datum of the 30% reference contrast. That is one of the magenta data in Fig. 3 and it is an aberrant datum with a very small error bar. We do not advocate removing systematic errors since those errors provide hints for improving the model (see next section). Trimming off random errors, where randomness is judged by the raggedness of adjacent data points, on the other hand, is less bothersome. KCT's subject AK-T devoted 8,890 trials to that dataset so it would be a shame to throw out that entire dataset and ignore what those data tell us. GM devote a large section of their discussion to showing that the most relevant data for distinguishing multiplicative noise from static noise are data with the shallowest psychometric function (largest Weber fraction). By that criterion Table 1 shows that the AK-T data is the most valuable, not the least valuable, of the four datasets.

The third approach is to invoke the methodology used in many statistics texts directed at the social sciences and medical fields. Rather than using a χ^2 test we can be more conservative and use an F test. Bevington and Robinson, 1992 justify this approach by saying that when one suspects that there is some randomness or non-stationarity that goes beyond the binomial variability, one can introduce a heterogeneity factor that converts the X^2statistic to an F-statistic. That is, rather than trusting the binomial statistics prediction of the standard error of the data, one uses the data itself for estimating the SEs. The penalty of this common approach is that one forgoes assessing the goodness-of-fit, that GM emphasize with respect to AK-T. The heterogeneity factor is the reduced X^2 of the best model, which is $X^2_{\text{red}} = 76.1/30 = 2.54$ for AK-T with the KCT4 model. With this heterogeneity correction, the deviation of the GM4 model (GM's non-multiplicative noise model with lowest X^2) from the KCT4 model has an F test of $F = 38.8/2.54 = 15.3$. The p value is $p = .00047$. This p-value is above zero (the original χ^2 test was zero according to Matlab's double precision calculation) but still small enough to reject the possibility that the GM4 fit was as good as the KCT4 fit. It is worth commenting again on our unconventional p value calculation of comparing GM4 to KCT4, two non-embedded models with four parameters, three of which are shared. We argued earlier that this comparison was conservative in that for purposes of computing a p value we treated GM4 as if it were GM3, ignoring the extra parameter (the Stromeyer–Foley parameter), so that GM4 could be considered to be embedded in KCT4 with $df = 1$.

5.6. The KCT dataset as part of the Modelfest project

Why has this paper devoted so much attention to what may seem like subtle, nit-picking statistical topics. Readers with fond memories of the spatial frequency revolution of the 1970s will recall that the single channel models of previous generations were overturned by subthreshold summation data that rested on 25% effects. Ensuing generations grappled with question of whether these 25% effects could be explained by attention or uncertainty. In order to deal with 25% effects one must be very careful about all the statistical issues being discussed in the present paper. Dealing with these types of issues has been a motivating force behind the Modelfest project (Carney, Klein, & Tyler, 1999; Watson & Ahumada, 2005). The idea behind Modelfest is to have multiple groups develop models for a common dataset, a practice that is all too rare in

vision science, a fragmented field where researchers tend to collect their own data and competition among models fitting the same data is uncommon. The obvious advantage of a common dataset is that better and better models will evolve that can cover a wider range of data. In addition, by having researchers analyze the same dataset, all the statistical issues discussed in the present paper will need to be discussed openly. I would like to propose that the KCT data become part of the Modelfest data bank. There are several reasons for this suggestion.

First, the vision science community needs to develop approaches for comparing incommensurate models. There are well-accepted methods for comparing embedded models (e.g., GM3 can be obtained from the other four parametric models by setting one or more parameters equal to zero). But what standards and statistical tests should be used for comparing incommensurate models such as KCT4, GM4, K4, and K4.25?

A second reason is that the Modelfest process with its focus on a common dataset forces us to consider "throwing out rules" have been considered in the present paper, but that are not commonly discussed: what are acceptable standards for ignoring data from a particular subject or for ignoring particular data points of a given subject.

A third reason is that the Modelfest process has already begun. There have already been four published analyses of the KCT data (Georgeson & Meese, 2006; Katkov et al., 2006a; Kontsevich et al., 2002; and present paper). It is expected that there will be further data on this topic that can be included in the Modelfest process.

In order for the KCT dataset to be available to future modelers, the data (with KCT's permission) will be available at http://cornea.berkeley.edu/KCTmodelfest or at http://neurometrics.com/KCTmodelfest.

5.7. What does one mean by multiplicative noise?

Finally, it is important to note that an important reason that the KCT4 fit does so well is that it is able to account for the asymmetric asymptotes of the AK-S and AK-T data. The K4.25 model, without explicit multiplicative noise, is able to fit the data as well the KCT4 model with multiplicative noise. K4.25 achieved the good fit by introducing an extra parameter to account for the asymptotic asymmetry. KTS achieved the good fit by a minimally constrained, unusual contrast response function with a shallow slope at low contrasts. One could argue that by introducing the "finger error" asymmetry parameter or the unusual contrast response function, one is in effect providing a multiplicative noise parameterization. Thus the improved X^2 values of models K4 and K4.25 and KTS do not provide evidence against a form of multiplicative noise.

Earlier in the Discussion three hypotheses were proposed for how intermixing of very high contrast trials with low contrast trials could produce errors in the low contrast discriminations: (a) adaptation, (b) afterimages, or (c) attentional factors. This brings up the issue that there are two sorts of multiplicative noise that need to be distinguished. In Eqs. (2), (9), and (19) the formula for the noise standard deviation is given as just depending on the contrast of the present stimulus, with no memory for preceding stimuli. That definition of multiplicative noise is widespread. However, Lu and Dosher (1999) have introduced a generalized notion of multiplicative noise whereby the noise variance depends not only on the present stimulus but also on preceding stimuli. It is quite possible that the asymmetry in the tails of the psychometric function are explainable by this generalized notion of multiplicative noise that includes the effects of adaptation, afterimages and attentional factors. It is easy to design experiments to distinguish between multiplicative noise that depends only on the current stimulus from noise produced by adaptation or afterimages due to preceding stimuli. The simplest approach would be to compare runs where high and low contrasts are intermixed (KCT data) to blocked runs that avoid intermixing. An alternative approach is to measure order effects. Levi, Klein, and Chen (2005) have an extended discussion on order effects in connection with the definition of internal noise that can be measured with a multi-pass classification image methodology. One can tease out order effects by presenting identical stimuli in either the identical order across runs or in scrambled order. The KCT data do not provide order effect information so more experiments are needed to discriminate current stimulus-dependent multiplicative noise from history-dependent multiplicative noise.

Observers AK-S and AK-T provide evidence that some sort of multiplicative noise is present using the KCT methodology. However, if the experiments had been done in a blocked manner, without intermixing high and low contrasts, it is quite possible that the evidence for multiplicative noise would vanish. Thus, I come to the same conclusion, but for very different reasons, as the final sentence of Georgeson and Meese (2006): "However much we might like to see this important issue resolved, the question remains open".

Acknowledgments

I thank Lenny Kontsevich and Christopher Tyler for discussions of their data over several years and Mark Georgeson for his useful input. This work was supported by NEI Grant EY 04776.

References

Bevington, P. R., & Robinson, D. K. (1992). *Data reduction and error analysis for the physical sciences*. New York: McGraw Hill.

Carney, T., Klein, S.A., & Tyler, C.W., et al. (1999). The development of an image/threshold database for designing and testing human vision models. In Rogowitz, & Pappas (Eds.), *Human vision and electronic imaging*, Vol. IV, Proceedings SPIE 3644, pp. 542–551.

Foley, J. M. (1994). Human luminance pattern vision mechanisms: masking experiments require a new model. *Journal of the Optical Society of America A, 11*, 1710–1719.

Foley, J. M., & Legge, G. E. (1981). Contrast detection and near-threshold discrimination in human vision. *Vision Research, 21*, 1041–1053.

Georgeson, M., & Meese, T. M. (2006). Fixed or variable noise in contrast discrimination? The jury's still out... Preceding paper. *Vision Research, 46*, 4294–4303.

Katkov, M., Tsodyks, M., & Sagi, D. (2006a). Singularities in the inverse modeling of 2AFC contrast discrimination data. *Vision Research, 46*, 259–266.

Katkov, M., Tsodyks, M., & Sagi, D. (2006b). Analysis of two-alternative forced-choice Signal Detection Model. *Journal of Mathematical Psychology, 50*, 411–420.

Klein, S. A. (2001). Measuring, estimating and understanding the psychometric function: a commentary. *Perception & Psychophysics, 63*, 1421–1455.

Kontsevich, L. L., Chen, C. C., & Tyler, C. W. (2002). Separating the effects of response nonlinearity and internal noise psychophysically. *Vision Research, 42*, 1771–1784.

Legge, G. E., & Foley, J. M. (1980). Contrast masking in human vision. *Journal of the Optical Society of America, 70*, 458–471.

Levi, D. M., Klein, S. A., & Chen, I. (2005). What is the signal in noise? *Vision Research, 45*, 1835–1846.

Lu, Z. L., & Dosher, B. A. (1999). Characterizing human perceptual inefficiencies with equivalent internal noise. *Journal of the Optical Society of America A, 11*, 764–778.

Nachmias, J., & Kocher, E. C. (1970). Visual detection and discrimination of luminance increments. *Journal of the Optical Society of America, 60*, 382–389.

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1992). *Numerical recipes in C: The art of scientific computing*. Cambridge: Cambridge University Press.

Stromeyer, C. F., III, & Klein, S. A. (1974). Spatial frequency channels in human vision as asymmetric (edge) mechanisms. *Vision Research, 14*, 1409–1420.

Watson, A. B., & Ahumada, A. J. Jr., (2005). A standard model for foveal detection of spatial contrast. *Journal of Vision, 5*, 717–740.