Consciousness and Cognition 21 (2012) 422–430

Contents lists available at SciVerse ScienceDirect

Consciousness and Cognition

journal homepage: www.elsevier.com/locate/concog

Short Communication

A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings

Brian Maniscalco [a,*], Hakwan Lau [a,b]

[a] Department of Psychology, Columbia University, NY 10027, United States
[b] Donders Institute of Brain, Cognition and Behavior, Radboud University, Nijmegen, The Netherlands

ARTICLE INFO

Article history:
Received 23 May 2011
Available online 8 November 2011

Keywords:
Metacognition
Signal detection theory
Type 2 sensitivity
Confidence rating

ABSTRACT

How should we measure metacognitive ("type 2") sensitivity, i.e. the efficacy with which observers' confidence ratings discriminate between their own correct and incorrect stimulus classifications? We argue that currently available methods are inadequate because they are influenced by factors such as response bias and type 1 sensitivity (i.e. ability to distinguish stimuli). Extending the signal detection theory (SDT) approach of Galvin, Podd, Drga, and Whitmore (2003), we propose a method of measuring type 2 sensitivity that is free from these confounds. We call our measure meta-d', which reflects how much information, in signal-to-noise units, is available for metacognition. Applying this novel method in a 2-interval forced choice visual task, we found that subjects' metacognitive sensitivity was close to, but significantly below, optimality. We discuss the theoretical implications of these findings, as well as related computational issues of the method. We also provide free Matlab code for implementing the analysis.

1. Introduction

In psychological tasks, one measure of interest may be how well an observer's confidence ratings predict stimulus judgment accuracy. For instance, suppose that observers perform a discrimination task, and on every trial provides a judgment about how confident they are that their discrimination is correct. We may ask: to what extent are the observer's confidence judgments predictive of response accuracy? In the literature, the task of discriminating between one's own correct and incorrect responses with confidence judgments has been called the "type 2 task" (Clarke, Birdsall, & Tanner, 1959; Galvin et al., 2003), as opposed to the "type 1 task" of discriminating between stimulus alternatives.

There are several widely used measures of type 2 sensitivity. Assuming confidence judgments are characterized in a binary way (high or low), a straightforward way to measure type 2 performance is to measure how often confidence judgments are congruent with accuracy, i.e. the probability that correct and incorrect judgments are "correctly" endorsed with high and low confidence, respectively. (See e.g. the "advantageous wagering" measure in Persaud, McLeod, & Cowey, 2007). A related approach is to compute a correlation coefficient between accuracy and confidence (e.g. phi in Kornell, Son, & Terrace, 2007 and gamma in Nelson, 1984). However, while these approaches are simple conceptually and computationally, they do not model type 2 sensitivity and type 2 response bias as separate processes and thus risk confounding them. For instance, a difference in two observers' confidence-accuracy correlation coefficient may be due merely to a difference in overall likelihood to endorse responses with high confidence, rather than a true difference in type 2 sensitivity. For this reason, signal

* Corresponding author. Address: Department of Psychology, Columbia University, 1190 Amsterdam Ave., MC 5501, New York, NY 10027, United States.
E-mail address: brian@psych.columbia.edu (B. Maniscalco).

1053-8100/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.concog.2011.09.021

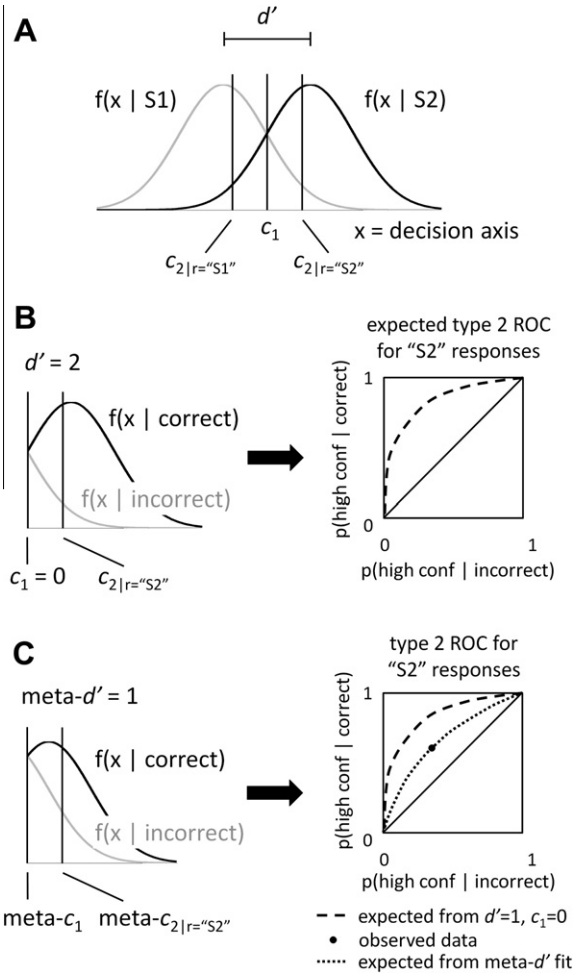

Fig. 1. Signal detection theory analysis of type 2 (metacognitive) performance. (A) The standard signal detection model. The observer must discriminate between stimulus classes S1 and S2. Each stimulus presentation generates a value on an internal decision axis, corresponding to the evidence in favor of S1 or S2. Evidence generated by each stimulus class is normally distributed across the decision axis, and the distance between these distributions in standard deviation units (d') measures how well the observer can discriminate S1 from S2. The observer sets a decision criterion c_1, such that all signals exceeding c_1 are labeled "S2" and all those failing to exceed c_1 are labeled "S1." The observer also sets criteria $c_{2|r="S1"}$ and $c_{2|r="S2"}$ to determine confidence ratings (higher ratings for signals farther from c_1). In this example, we set $d' = 2$ and $c_1 = 0$. (B) Expected type 2 sensitivity from d' and c_1. Consider only trials where the observer responds "S2," i.e. only the portion of the decision axis exceeding c_1. Then the S2 distribution corresponds to the distribution of evidence for correct responses (i.e. S2 stimuli classified as "S2"), and the S1 distribution corresponds to the distribution of evidence for incorrect responses (i.e. S1 stimuli classified as "S2"). All trials surpassing $c_{2|r="S2"}$ are endorsed with high confidence. Sweeping the $c_{2|r="S2"}$ criterion across the decision axis generates different values for type 2 false alarm rate (p(high confidence | incorrect)) and type 2 hit rate (p(high confidence | correct)), and thus generates a type 2 ROC (Receiver Operating Characteristic) curve. (Similar considerations hold for "S1" responses.) Thus, d' and c_1 are jointly sufficient to determine type 2 sensitivity for each response type, according to the standard signal detection model. (C) Characterizing type 2 sensitivity. Consider only trials where the observer responds "S2," i.e. only the portion of the decision axis exceeding c_1. The analysis from (A) and (B) can be inverted in order to characterize type 2 sensitivity. Suppose that the observer has $d' = 2$ and $c_1 = 0$, with "S2" responses having a type 2 hit rate = .64 and a type 2 false alarm rate = .41. We may characterize type 2 sensitivity as meta-d', i.e. the level of d' that would have been expected to have generated the observed type 2 data. In this example, meta-$d' = 1$ even though $d' = 2$, indicating type 2 sensitivity below expectation. Though not pictured here, this analysis can likewise be applied to "S1" responses.

detection theory (SDT) approaches (Macmillan & Creelman, 2005) should be preferred, because they allow one to separate the independent contributions of sensitivity and response bias in type 2 task performance.

Several SDT approaches to characterizing type 2 performance have been put forth. Kunimoto, Miller, and Pashler (2001) proposed modeling type 2 performance in the same way SDT models type 1 performance. In the simplest type 1 SDT model, we assume that two stimulus alternatives generate normal distributions of evidence along some internal decision axis (see e.g. Fig. 1A), with the normalized distance between the distributions, d', providing a measure of stimulus discrimination sensitivity. In Kunimoto et al.'s approach, we similarly assume that correct and incorrect judgments generate normal distributions of evidence along some decision axis, with the normalized distance between them, a', providing a measure of type 2

sensitivity. However, specifying the parameters of the standard SDT model[1] already places strong constraints on the distributions of evidence for correct and incorrect judgments (Fig. 1B; Galvin et al., 2003), and these distributions in general do not conform well to the model proposed by Kunimoto et al.; in a sense, their type 2 SDT model is thus inconsistent with the type 2 implications of the standard type 1 SDT model. Thus, *a*' does not satisfactorily separate type 2 sensitivity from type 2 response bias theoretically (Galvin et al., 2003) or empirically (Evans & Azzopardi, 2007).

Clarke et al. (1959) and more recently Galvin et al. (2003) discussed how distributions of evidence for correct and incorrect stimulus judgments could be derived from the type 1 SDT model. An important lesson from this work is that type 1 sensitivity (d') and response bias (c_1) influence the area under the type 2 ROC curve (Fig. 1B). This entails that two metacognitively optimal observers could differ on type 2 performance due only to differences in type 1 performance.[2]

This observation invites a distinction between what might be called "absolute" type 2 sensitivity and "relative" type 2 sensitivity.[3] Suppose observer A has $d' = 1$, $c_1 = 0$ and observer B has $d' = 2$, $c_1 = 0$, but that both observers make optimal use of the type 1 information available to them when performing the type 2 task. B will have greater area under her type 2 ROC curve than A, and in general her confidence ratings will be more predictive of accuracy. In this sense, B has greater "absolute" type 2 sensitivity than A. But by hypothesis, the difference in their metacognitive performance derives entirely from informational differences at the type 1 level, and so in a sense it is misleading to conclude that the metacognitive mechanisms of B are operating at a higher level of efficiency or sensitivity than those of A. The difference in their absolute type 2 sensitivity reflects the difference in the quality of type 1 information they are metacognitively evaluating, rather than in the quality of the evaluation itself. Once we take type 1 performance into account, we see that A and B are in fact equally effective at metacognitively evaluating the type 1 information available to them. In this sense, A and B have equivalent "relative" type 2 sensitivity, i.e. type 2 sensitivity relative to type 1 performance.

Note that absolute and relative type 2 sensitivity assess different aspects of metacognitive performance. Absolute type 2 sensitivity measures how much information confidence ratings carry about task performance. Relative type 2 sensitivity factors out the contribution of type 1 performance to absolute type 2 sensitivity, thus revealing the efficacy of metacognitive processing in and of itself. In other words, absolute type 2 sensitivity tells us how much we should trust an observer's confidence ratings, which depends on the quality of information being metacognitively evaluated as well as the quality of the metacognitive evaluation itself. Relative type 2 sensitivity separates these factors, providing a measure of the quality of the metacognitive evaluation itself.

For many research applications we are interested specifically in assessing the efficacy of metacognitive mechanisms in and of themselves. In such instances, measures of absolute type 2 sensitivity such as area under the type 2 ROC curve (e.g. Kolb & Braun, 1995; Wilimzig, Tsuchiya, Fahle, Einhäuser, & Koch, 2008) may not be appropriate, because such measures are likely to be influenced by both the efficacy of metacognitive function and the quality of information those mechanisms are evaluating.

How should we measure relative type 2 sensitivity? We endorse the proposal of Galvin et al. (2003) to evaluate *observed* type 2 sensitivity with reference to the type 2 sensitivity that would be *expected* to occur, given an SDT analysis of the observed type 1 performance (henceforth, "SDT-expected type 2 sensitivity"). Galvin et al. envisioned doing this comparison at the level of type 2 distributions of evidence, conditional on response accuracy. But this approach meets with several difficulties. It is difficult to compute *SDT-expected* type 2 sensitivity since it is difficult to derive general mathematical forms of the type 2 distributions from the type 1 model. And it is unclear how to compute *observed* type 2 sensitivity from observed type 2 ROC data in terms of parametric type 2 distributions, given their complexity and dependence on type 1 model parameters.

We observe that the spirit of Galvin et al.'s analysis can be retained while bypassing the difficulties of working directly with type 2 distributions (Fig. 2). Due to the theoretical link between type 1 and type 2 SDT models (Fig. 1B), type 2 sensitivity can be expressed at the level of type 1 distributions (Fig. 2A). That is, we can characterize *observed* type 2 sensitivity as the value of d' that a metacognitively optimal observer would have required to produce the empirically observed type 2 data. We call this measure "meta-d'" to reflect that it is a measure of type 2 sensitivity (meta-) expressed at the level of type 1 SDT (d'). One can think of meta-d' as a measure of the signal that is available for the subject to perform the type 2 task. While meta-d' measures *observed* type 2 sensitivity, its counterpart for *SDT-expected* type 2 sensitivity is simply the empirically observed value of d'. Importantly, since meta-d' is expressed in the same scale as the conventionally estimated d' value, the two can be compared directly. The comparison of meta-d' with d' achieves the comparison of *observed* type 2 sensitivity to *SDT-expected* type 2 sensitivity. In turn, this comparison gives us a measure of relative type 2 sensitivity.

Meta-d' has high interpretational value. If meta-$d' = d'$, then the observer exhibits type 2 sensitivity in agreement with what the standard SDT model would expect it to be, given the observed type 1 performance. In other words, on an SDT analysis we could say that the observer is metacognitively "ideal," making use of all the information available for the type 1 task

[1] In the following, we use the term "standard SDT model" to refer to the model depicted in Fig. 1A, where stimuli generate normal distributions of evidence on an internal decision axis, and observers set criteria to determine stimulus classification responses and confidence ratings. This model has had great success in capturing patterns in empirical data; see Macmillan and Creelman (2005).

[2] In the following, we use the term "type 1 performance" to refer jointly to the sensitivity and response bias of stimulus discrimination. We use "type 2 sensitivity" to refer to the efficacy with which confidence judgments discriminate between correct and incorrect responses. "Type 2 response bias" refers to the overall propensity to endorse type 1 decisions with high confidence.

[3] Note that absolute and relative type 2 sensitivity are not defined analogously to previously used metacognitive measures, absolute and relative accuracy.

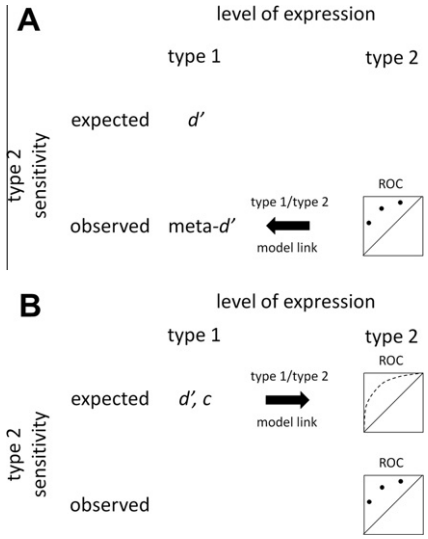

Fig. 2. Schematic representation of two ways to quantify absolute and relative type 2 sensitivity in an SDT framework. Absolute type 2 sensitivity is a quantification of how well confidence ratings distinguish between correct and incorrect type 1 decisions. Absolute type 2 sensitivity can be expressed at either the level of type 1 or type 2 processes due to the theoretical relationship between type 1 and type 2 processes in SDT (Fig 1). Relative type 2 sensitivity measures how well metacognitive mechanisms function. It is derived by a quantitative comparison between observed and SDT-expected values of absolute type 2 sensitivity occurring on the same level of expression. (A) From type 2 to type 1. The observed type 2 ROC can be expressed in terms of the type 1 SDT parameter meta-d' that would make the observed type 2 data most likely to have occurred. Meta-d' can be compared to d' to yield a measure of relative type 2 sensitivity. (B) From type 1 to type 2. An estimate of the area under the observed type 2 ROC can serve as an absolute measure of type 2 sensitivity. The observed values of type 1 parameters d' and c_1 can be used to generate an expectation for the type 2 ROC. The areas underneath the observed and expected curves can be compared to yield a measure of relative type 2 sensitivity.

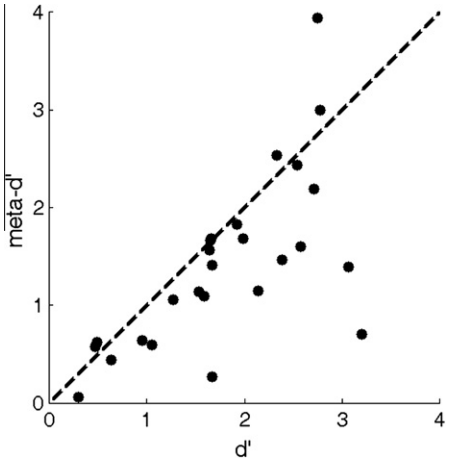

Fig. 3. Observed vs. expected type 2 sensitivity in a spatial 2IFC task using meta-d'. 27 subjects performed a spatial 2IFC task and rated confidence in response accuracy (Study 1 from main text). Their absolute type 2 sensitivity was characterized using meta-d' (Fig1; Fig 2A). Dots represent individual subject data and the dashed line shows the curve expected from SDT, meta-d' = d'. There is a strong correlation between d' and meta-d' (r = .68, p = .0001), such that d' serves as an approximate upper bound on meta-d′, in agreement with the theoretical link between type 1 and type 2 SDT models. Yet there is substantial between-subject variance in the relationship between meta-d' and d', such that several subjects fall well below the level of expected type 2 sensitivity and the group average for meta-d' is significantly lower than d' (paired t-test, p = .009). This suggests that we cannot take it for granted in empirical data that d' = meta-d', and thus meta-d' has meaningful empirical work to do in contributing to absolute and relative measures of type 2 sensitivity.

when performing the type 2 task. If meta-$d' \neq d'$, then the observer's type 2 sensitivity either outperforms or underperforms expectation (Fig. 1C). Typically, one would expect meta-$d' \leqslant d'$, on the assumption that the information available for the type 1 task is *exhaustive* of the information available for the type 2 task. In this case, the degree to which meta-d' is smaller than d' reflects the degree to which the observer is metacognitively inefficient.

Because d' has ratio scaling properties (Macmillan & Creelman, 2005), differences and ratios of d' values are meaningful; for instance, if observer A has d' = 2 and observer B has d' = 1, it is meaningful to say that A has twice the sensitivity of B.

Since meta-d' is expressed on the same scale as d', numerical comparisons between d' and meta-d' are also yield meaningful quantities. Thus, we are not limited to testing the null hypothesis that d' = meta-d', but can make graded assessments of relative type 2 sensitivity based on the outcome of differences or ratios. For instance, one can meaningfully state that a certain psychophysical manipulation changed an observer's metacognitive capacity from 100% to 70%, or that a certain drug reduced the observer's metacognitive capacity by 0.3 signal-to-noise ratio units (since d' and meta-d' are expressed in signal-to-noise ratio units).

The detailed method for estimating meta-d' is described in Supplementary Online Materials. In brief, the central idea of the estimation is that the type 1 SDT model entails what the type 2 ROC curves for each type 1 response should be (Fig. 1B). Thus, we can directly fit the parameters of a type 1 SDT model so as to optimize the fit of the type 1 SDT model's *predicted* type 2 ROC curves to the *observed* type 2 ROC data (Fig. 1C). Meta-d' is the d' of the type 1 SDT model that maximizes the likelihood of the observed type 2 ROC data (given a response bias similar to that observed in the empirical data; see Supplementary Online Materials). So for instance, if an observer has d' = 2 and meta-d' = 1, we could say that although their *actual* d' is 2, their response-conditional type 2 ROC curves behave *as if* their d' were only 1.

For implementation, we provide free Matlab code for easy estimation of meta-d' (http://www.columbia.edu/~bsm2105/type2sdt/). On that website we also provide further documentation that gives full technical treatment to several theoretical and computational issues.

In the present study, we applied this new analysis approach to estimate subjects' metacognitive sensitivity in a spatial 2IFC (2-interval forced-choice) visual task. Specifically, we tested how far they deviated from optimal metacognitive sensitivity given their type 1 performance.

2. Methods

Thirty participants performed a spatial 2IFC task. In each trial, participants distinguished between 2 spatial arrangements of visual stimuli and then rated their confidence in the accuracy of their responses on a four-point scale. Details are reported in Supplementary Online Materials.

3. Results

Despite the fact that we tried to titrate the stimulus contrast to control for type 1 performance level, there was substantial between-subject variation in (type 1) d'. Nonetheless, we can use this to our advantage by observing the results of the meta-d' estimation across a range of d' values. In Fig. 3, we plotted meta-d' vs d' for every subject. Note that there was a substantial positive relationship between these variables with most data points clustering near the line meta-d' = d', in line with the SDT prediction (Pearson's r = .68, p = .0001). Further, while very few subjects outperformed the SDT expectation, several fell below expectation (i.e. below the line meta-d' = d'). Indeed, a paired t-test reveals that the mean level of meta-d' (1.37) was significantly lower than the mean d' (1.78), $t(26) = 3.0$, p = .006. The mean value of meta-d'/d' was 0.77, indicating that on average subjects in this task exhibited absolute type 2 sensitivity at only about 77% of what would have been expected from their type 1 task performance.

The log likelihood of the meta-d' fits (see Supplementary Online Materials) did not correlate with meta-d' (p = .2) or meta-d'–d' (p = .8), suggesting that measured variation in absolute and relative type 2 sensitivity in this data set cannot be attributed merely to variation in the quality of data fitting.

4. Discussion

In this study we have employed a novel method for isolating and measuring the sensitivity with which metacognitive mechanisms differentiate between correct and incorrect decisions. One of the primary strengths of SDT is that we can use it to calculate d', a measure of stimulus classification sensitivity independent from the influence of response bias. In a similar spirit, we have demonstrated a method for extending the standard SDT model in order to estimate meta-d', a measure of type 2 sensitivity. Unlike most previously proposed measures of type 2 sensitivity, meta-d' is not confounded with type 1 sensitivity (d'), type 1 response bias (propensity to use one stimulus classification response more than another), or type 2 response bias (propensity to give high confidence ratings).

Although the potential confounding of sensitivity and response bias is familiar from standard SDT, another potential confound looms in any analysis of type 2 data. This potential confound arises from the insight that, according to the standard SDT model, type 1 task performance influences the area under the type 2 ROC curve. Thus, apparent differences in observers' metacognitive capabilities may in fact be attributable merely to differences in how they perform on the primary stimulus classification task. Our approach characterizes type 2 sensitivity in terms of meta-d', the d' one would expect to have generated the observed type 2 hit rates and false alarm rates. In the spirit of Galvin et al. (2003), meta-d' can then be compared to the actual d' exhibited by the observer in order to quantify how well observed type 2 sensitivity compares to the type 2 sensitivity that would be expected to result from the observed type 1 performance. Meta-d' measures type 2 sensitivity independently from type 2 response bias, and evaluating meta-d' with respect to d' (e.g. calculating meta-d'–d' or meta-d'/d') takes into account the effect of type 1 performance on the type 2 data.

One potential source of concern in our analysis is the assumption that type 1 evidence is distributed normally along an internal decision axis. If the assumption of normality does not hold then meta-d' is not an appropriate measure of type 2 sensitivity. However, historically the assumption of normality has been observed to fit a wide range of psychological data well (Macmillan & Creelman, 2005); it is for precisely this reason that SDT has been so successful. We note also that the general approach we describe can be readily generalized to any model where two stimulus classes generate distributions of evidence along a type 1 decision axis; that is, the assumption of normality per se is not critical to the spirit of the kind of analysis we propose.

Finally, we note that the majority of supposedly "non-parametric" measures do not offer a satisfactory alternative. Although measures such as advantageous wagering and trial-by-trial correlation coefficients between confidence and accuracy are not explicitly built upon parametric assumptions, they nonetheless imply curves of constant sensitivity on an ROC plot, which in turn entails the values of the likelihood ratios of the underlying distributions of evidence at every point on the decision axis (Macmillan & Creelman, 2005). Thus, although these measures are not explicitly parametric, they nonetheless tacitly imply parametric assumptions. In fact, this is true for any measure that assigns a sensitivity value to a single hit rate and false alarm rate. Alternatively, one may characterize absolute type 2 sensitivity in a non-parametric way by empirically estimating the area under the type 2 ROC curve. However, it seems that converting this into a *relative* measure of type 2 sensitivity by correcting for the influence of type 1 performance requires parametric assumptions after all, in order to generate an SDT-expected type 2 ROC based on type 1 performance (Figs. 1B and 3B).

4.1. Sources of variation in type 2 sensitivity

Our results indicate that subjects' type 2 sensitivity was below the point of optimality and exhibited substantial inter-subject variability. What mechanisms might be responsible for an observed level of type 2 sensitivity that deviates from SDT-expectation? We briefly consider two general kinds of mechanisms that could account for such deviations, although this treatment is not intended to be exhaustive.

One way of interpreting a finding that meta-$d' = d'$ is that the cognitive mechanisms responsible for making type 1 and type 2 decisions access the same source of information. However, it is possible that in some situations, the different decision making mechanisms do not access the same information. If type 1 and type 2 mechanisms access information that differs in quality, quantity, source, or type, such differential access could entail a relative difference in sensitivity such that meta-$d' \neq d'$.

For instance, on some models, type 2 decisions are based upon further processing of the same information used to make type 1 decisions (Lau & Rosenthal, 2011; Maniscalco & Lau, 2010; Pleskac & Busemeyer, 2010). In such cases, information may decay or accrue internal noise subsequent to the type 1 decision. If so, then although type 1 and type 2 mechanisms access the same *source* of information, the *quality* of the information available for each level may differ.

Alternatively, type 1 decision making mechanisms may have access to information that is entirely unavailable to type 2 mechanisms, or vice versa. For instance, an 'unconscious' processing stream may contribute to type 1, but not type 2, decisions (Del Cul, Dehaene, Reyes, Bravo, & Slachevsky, 2009). More complicated processing structures implying an asymmetry of information access are also possible (e.g. Pasquali, Timmermans, & Cleeremans, 2010).

A third possibility is that type 1 and type 2 decisions are made based upon different representations or transformations of the same underlying information. In SDT terms, the observer may apply different decision axes in internal evidence space for type 1 and type 2 decisions, where each decision axis embodies a particular way of summarizing and evaluating the available information (Macmillan & Creelman, 2005). One decision axis may represent a more optimal decision rule than the other, which could entail differences in decision sensitivity.

A second general kind of mechanism that could account for unexpected type 2 sensitivity is differential noise in the type 1 and type 2 decision process, rather than differential access to the information being evaluated. When applying SDT to data, researchers usually model the decision criteria as if they are constant from trial to trial. However, it may be the case that criterion setting is actually a noisy process in which criteria are placed at differing locations of the decision axis from trial to trial (Benjamin, Diaz, & Wee, 2009; Mueller & Weidemann, 2008). Increasing criterion variability has the effect of decreasing measured sensitivity. Thus, if type 2 criterion setting is noisier than type 1 criterion setting, we should expect that meta-$d' < d'$. Likewise, if the type 1 criterion is more variable than the type 2 criteria, we should expect that meta-$d' > d'$.

4.2. Alternative approaches to estimating type 2 sensitivity

It is popular practice to estimate multiple pairs of false alarm rate (FAR) and hit rate (HR) for a single observer by asking the observer to provide a stimulus classification and confidence rating on every trial. When the observer has two possible stimulus classification responses "S1" and "S2" and N possible confidence ratings, there are $2 * N$ possible responses an observer may give on any trial. If these responses are ordered from "highest confidence S1" to "highest confidence S2", there are $2 * N - 1$ ways of partitioning the total response scale into two sets. Each such partition represents a way of combining the observer's classification and confidence ratings into two categories of responses, one favoring "S1" and the other favoring "S2," and these combined response categories can be used to calculate a (FAR, HR) pair. Thus, we can calculate $2 * N - 1$ (FAR, HR) pairs and thereby create an empirical ROC curve using rating data.

One interesting feature of type 1 ROC constructed from rating data is that it depends jointly on type 1 performance and type 2 sensitivity (Clifford, Arabzadeh, & Harris 2008). For instance, suppose $d' > 0$ but meta-$d' = 0$. In this case, each rating-constructed type 1 ROC point for "S2" responses will lie on the line between (0,0) and (type 1 FAR, type 1 HR). Likewise, each rating-constructed type 1 ROC point for "S1" responses will lie on the line between (type 1 FAR, type 1 HR) and (1,1). Thus, the shape of the type 1 ROC curve will be piecewise linear. When meta-$d' = d'$, the shape of the type 1 ROC curve will be identical to the smooth, continuous shape predicted from standard SDT. When 0 < meta-$d' < d'$, the type 1 ROC curve will lie between these two extremes.

In fact, Clifford et al. (2008) argued that the shape of the rating-constructed type 1 ROC curve could be used to assess type 2 sensitivity. In particular, they argued that if rating-constructed type 1 ROC points are better fit by a piecewise linear ROC curve than the ROC curve predicted by SDT, then one can infer that metacognitive sensitivity is suboptimal. This approach has the virtue of assessing type 2 sensitivity independently from type 1 performance (i.e. the location of "true" type 1 ROC point in ROC-space) and type 2 response bias. It thus provides an assessment of relative type 2 sensitivity.

However, the approach of Clifford et al. (2008) is limited in several ways. Historically, it has been difficult to establish empirical deviations from expectation in rating-constructed type 1 ROC shape (Macmillan & Creelman, 2005). One reason for this difficulty may be that the area under the rating-constructed type 1 ROC curve does not change drastically even for extreme differences in type 2 sensitivity. This is because the area under the rating-constructed type 1 ROC curve is predominately determined by the "true" type 1 ROC point, with only marginal contributions from variation in type 2 sensitivity. For instance, when $d' = 2$ and $c_1 = 0$, the area under the rating-constructed type 1 ROC is .921 if meta-$d' = d' = 2$ and .841 if meta-$d' = 0$. Thus in this example, a 100% difference in relative type 2 sensitivity generates only an 8.7% difference in area under the type 1 ROC curve. Recent studies such as Mueller and Weidemann (2008) have used the approach of assessing the shape of the rating-constructed type 1 ROC curve to make inferences about type 2 sensitivity, but these approaches seem to require manipulating the type 1 criterion in extreme ways via strong base rate or response incentive manipulations in order to detect a difference between conditions in type 1 ROC shape. This is because increasing the bias of the type 1 criterion tends to increase the influence of type 2 sensitivity on the area under the rating-constructed type 1 ROC curve.

Another weakness of this approach is that it makes graded assessment of type 2 sensitivity difficult. The approach Clifford et al. (2008) endorse essentially amounts to a binary test of the null hypothesis that meta-$d' = d'$. By contrast, the meta-d' approach we propose in this paper allows one to assess relative type 2 sensitivity on a graded scale and make graded comparisons to expectation. Likewise, our approach allows one to detect cases where type 2 sensitivity outperforms expectation, a possibility not accounted for by the approach of Clifford et al.

Yet another approach for comparing observed with SDT-expected type 2 sensitivity can be achieved at the level of the type 2 ROC (Fig. 2B). In this case, SDT-expected area under the type 2 ROC curve (AUC) can be directly estimated from the type 1 model (Fig. 1B), and observed type 2 AUC can be estimated from the empirical data using the non-parametric measure A_g (Pollack & Hsieh, 1969). This method of comparing observed and SDT-expected type 2 sensitivity is closely related to the meta-d' analysis discussed thus far, providing a viable alternative method that potentially provides greater conceptual and computational simplicity.

The primary disadvantage of the type 2 AUC approach is the meaning of differences and ratios between two AUC values are more difficult to interpret. For instance, it is presumably easier for an observer to improve AUC from 0.5 (chance performance) to 0.6 than it is to improve from 0.9 to 1.0 (perfect performance). (Type 1 AUC values of 0.5, 0.6, 0.9, and 1 correspond to d' values of 0, 0.35, 1.85, and infinity, respectively.) The same difference score seems to correspond to different changes in actual performance level, depending on the magnitude of the values being compared. Thus, while this method allows us to statistically assess the null hypothesis that observed and SDT-expected type 2 sensitivity are equal, it does not seem ideally suited for making graded quantitative assessments of relative type 2 sensitivity.

Another complication is that the non-parametric measure of type 2 AUC, A_g, has some non-ideal properties. A_g essentially takes the area under the ROC curve formed by drawing straight lines between observed (FAR, HR) pairs and the ROC endpoints (0,0) and (1,1). As a consequence, it potentially underestimates AUC, since (FAR, HR) pairs on an ROC curve tend to be connected by concave down lines, not straight lines. The degree of this underestimation is potentially exacerbated if strong response biases exist in the data, since such biases make (FAR, HR) pairs cluster near the endpoints (0,0) and/or (1,1).

Unlike the observed type 2 ROC data, the SDT-expected type 2 ROC curves are fully continuous. We thus recommend sampling points from the continuous SDT-expected curve (for instance, those (FAR, HR) pairs whose FAR values match the empirically observed type 2 FAR values) and computing A_g from these pairs. The comparison of observed type 2 A_g with type 2 A_g from the sampled points of the SDT-expected type 2 ROC curve mitigates the underestimation problem by making both A_g quantities similarly underestimated.

4.3. Relevance to conscious awareness

How should we conceptualize the cognitive phenomenon that type 2 sensitivity measures? Some (e.g. Kunimoto et al., 2001; Persaud et al., 2007) have argued that type 2 sensitivity provides an "objective" measure of subjective awareness. That is, an observer can be said to be aware of a stimulus if his confidence ratings distinguish well between correct and incorrect stimulus classifications. Likewise, if an observer's confidence ratings are uninformative regarding the correctness of his classifications, then he can be said to be unaware of the stimuli, even if he can classify them at above chance levels.

Although this interpretation may have some heuristic value, we do not endorse it as a rigid criterion for measuring awareness. For instance, imagine that subjects perform a visual detection task and rate decision confidence. In condition A, subjects use a confidence scale with two options; in condition B, they use a scale with 20 options. Even if type 1 task performance is the same for both conditions, we will likely find that the subjects in condition B exhibit lower type 2 sensitivity than those in condition A (Wickelgren, 1968). A natural explanation is that the cognitive demand of maintaining so many type 2 decision criteria in condition B caused elevated variability in type 2 criterion setting. But if so, it seems counterintuitive to say that the subjects in condition B were *less aware* of the visual stimuli. After all, their deficit in type 2 sensitivity was likely an artifact of the task demands of judging and reporting confidence, rather than reflecting a deficit in the underlying information on which those confidence reports were based.

Additionally, it has been shown that blindsight patient GY can place wagers on perceptual discriminations in his blind visual field with an above chance level of type 2 sensitivity (Persaud, McLeod, et al., 2007; Persaud, Davidson, et al., 2011). Thus, it seems that awareness is not necessary for above-chance levels of type 2 sensitivity. Even when an observer reports no direct phenomenal representation of the stimulus he is judging, his confidence in decision accuracy may nonetheless be somewhat diagnostic of actual type 1 performance.

Consideration of these arguments suggests a double dissociation between type 2 sensitivity and the contents of awareness. Reductions in type 2 sensitivity do not necessarily reflect reductions in phenomenological stimulus awareness, and above-chance levels of type 2 sensitivity do not necessarily imply the presence of phenomenological stimulus awareness. Thus it does not seem tenable to use type 2 sensitivity as a hard and fast measure of awareness.

Similar arguments have been put forth by Dienes and Perner (2004) and Dienes (2004), who highlight the importance of placing type 2 sensitivity in the context of the actual content of the type 2 reports themselves when making inferences about awareness. For instance, in the above example, GY's direct denials of having any visual experience in his blind field (e.g. Persaud & Lau, 2008) are what cause us to doubt that the above-chance type 2 sensitivity he can sometimes exhibit in his blind field is indicative of visual awareness. Similarly, suppose an observer consistently endorses his decisions about a stimulus with either "high" or "very high" confidence, but suppose also that the observer's type 2 sensitivity is at chance. It would seem odd to infer that the observer has no awareness of the stimulus, given his consistently high ratings of confidence. A more careful interpretation is just that the observer's distinction between "high" and "very high" confidence is not a distinction that carries any useful information about task performance. The observer may be quite aware of the stimuli he is observing in several meaningful respects, despite being undiscerning in his distinction between "high" and "very high" confidence (Dienes, 2004).

We advocate an interpretation of type 2 sensitivity more strictly in line with its operational definition. Absolute type 2 sensitivity tells us how well an observer can distinguish between his own correct and incorrect decisions. Relative type 2 sensitivity more specifically isolates the sensitivity with which metacognitive mechanisms operate. These measures do not necessarily inform us about the *source* of an observer's confidence ratings (e.g. about states of phenomenal awareness), but rather they inform us about the informational *relationship* between confidence ratings and task performance. In particular, relative type 2 sensitivity measures how well an observer can evaluate his own first-order stimulus evaluations.

Acknowledgment

This work is supported by internal funding from Columbia University (to HL). We thank Dobromir Rahnev and Steve Fleming for their helpful comments on the manuscript.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.concog.2011.09.021. The supplementary material contains details on the experimental procedures for the experiment reported in the main text. It also contains a discussion on the methodology for estimating meta-d'.

References

Benjamin, A. S., Diaz, M., & Wee, S. (2009). Signal detection with criterion noise: Applications to recognition memory. *Psychological Review, 116*(1), 84–115.

Clarke, F. R., Birdsall, T. G., & Tanner, J. (1959). Two types of ROC curves and definitions of parameters. *The Journal of the Acoustical Society of America, 31*(5), 629–630.

Clifford, C. W. G., Arabzadeh, E., & Harris, J. A. (2008). Getting technical about awareness. *Trends in Cognitive Sciences, 12*(2), 54–58.

Del Cul, A., Dehaene, S., Reyes, P., Bravo, E., & Slachevsky, A. (2009). Causal role of prefrontal cortex in the threshold for access to consciousness. *Brain, 132*(9), 2531–2540.

Dienes, Z. (2004). Assumptions of subjective measures of unconscious mental states: Higher order thoughts and bias. *Journal of Consciousness Studies, 11*, 25–45.

Dienes, Z., & Perner, J. (2004). Assumptions of a subjective measure of consciousness: Three mappings. In R. Gennaro (Ed.), *Higher order theories of consciousness* (pp. 173–199). Amsterdam: John Benjamins Publishers.

Evans, S., & Azzopardi, P. (2007). Evaluation of a 'bias-free' measure of awareness. *Spatial Vision, 20*(1–2), 61–77.

Galvin, S. J., Podd, J. V., Drga, V., & Whitmore, J. (2003). Type 2 tasks in the theory of signal detectability: discrimination between correct and incorrect decisions. *Psychonomic Bulletin & Review, 10*(4), 843–876.

Kolb, F. C., & Braun, J. (1995). Blindsight in normal observers. *Nature, 377*(6547), 336–338.

Kornell, N., Son, L. K., & Terrace, H. S. (2007). Transfer of metacognitive skills and hint seeking in monkeys. *Psychological Science. A Journal of the American Psychological Society/APS, 18*(1), 64–71.
Kunimoto, C., Miller, J., & Pashler, H. (2001). Confidence and accuracy of near-threshold discrimination responses. *Consciousness and Cognition, 10*(3), 294–340.
Lau, H., & Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends in Cognitive Sciences, 15*(8), 365–373.
Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Lawrence Erlbaum, p. 512.
Maniscalco, B., & Lau, H. (2010). Comparing signal detection models of perceptual decision confidence. *Journal of Vision, 10*(7), 213.
Mueller, S. T., & Weidemann, C. T. (2008). Decision noise: An explanation for observed violations of signal detection theory. *Psychonomic Bulletin & Review, 15*(3), 465–494.
Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95*(1), 109–133.
Pasquali, A., Timmermans, B., & Cleeremans, A. (2010). Know thyself: metacognitive networks and measures of consciousness. *Cognition, 117*(2), 182–190.
Persaud, N., McLeod, P., & Cowey, A. (2007). Post-decision wagering objectively measures awareness. *Nature Neuroscience, 10*(2), 257–261.
Persaud, N., & Lau, H. (2008). Direct assessment of qualia in a blindsight participant. *Consciousness and Cognition, 17*(3), 1046–1049.
Persaud, N., Davidson, M., Maniscalco, B., Mobbs, D., Passingham, R. E., Cowey, A., et al (2011). Awareness-related activity in prefrontal and parietal cortices in blindsight reflects more than superior visual performance. *NeuroImage, 58*(2), 605–611.
Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review, 117*(3), 864–901.
Pollack, I., & Hsieh, R. (1969). Sampling variability of the area under the ROC-curve and of *d*'e. *Psychological Bulletin, 71*(3), 161–173.
Wickelgren, W. A. (1968). Unidimensional strength theory and component analysis of noise in absolute and comparative judgments. *Journal of Mathematical Psychology, 5*(1), 102–122.
Wilimzig, C., Tsuchiya, N., Fahle, M., Einhäuser, W., & Koch, C. (2008). Spatial attention increases performance but not subjective confidence in a discrimination task. *Journal of Vision, 8*(5), 7.1–10.