Should metacognition be measured by logistic regression?

Manuel Rausch [1,2] and Michael Zehetleitner [1,2]

[1] Katholische Universität Eichstätt-Ingolstadt, Eichstätt, Germany

[2] Ludwig-Maximilians-Universität München, Munich, Germany

Correspondence should be addressed at:
Manuel Rausch
Katholische Univerität Eichstätt-Ingolstadt
Psychologie II
Ostenstraße 25, "Waisenhaus"
85072 Eichstätt
Germany
Phone: +49 8421 93 21639
Email: manuel.rausch@ku.de
http://www.ku.de/ppf/psychologie/psych2/mitarbeiter/m-rausch/

This manuscript was accepted for publication in Consciousness and Cognition. Please cite this work as: Rausch, M., Zehetleitner, M. (2017). Should metacognition be measured by logistic regression? *Consciousness and Cognition.* 49, 291-312. The final publication is available at http://dx.doi.org/10.1016/j.concog.2017.02.007

Abstract

Are logistic regression slopes suitable to quantify metacognitive sensitivity, i.e. the efficiency with which subjective reports differentiate between correct and incorrect task responses? We analytically show that logistic regression slopes are independent from rating criteria in one specific model of metacognition, which assumes (i) that rating decisions are based on sensory evidence generated independently of the sensory evidence used for primary task responses and (ii) that the distributions of evidence are logistic. Given a hierarchical model of metacognition, logistic regression slopes depend on rating criteria. According to all considered models, regression slopes depend on the primary task criterion. A reanalysis of previous data revealed that massive numbers of trials are required to distinguish between hierarchical and independent models with tolerable accuracy. It is argued that researchers who wish to use logistic regression as measure of metacognitive sensitivity need to control the primary task criterion and rating criteria.

Keywords: metacognition; metacognitive sensitivity, logistic regression; signal detection theory; type 2 signal detection theory; generalized linear regression, cognitive modelling

1 Introduction

Metacognitive sensitivity, also called type 2 sensitivity or resolution of confidence, refers to the efficiency with which participants' subjective reports during an experimental task discriminate between correct and incorrect responses in a primary task (Baranski & Petrusic, 1994; Fleming & Lau, 2014; Galvin, Podd, Drga, & Whitmore, 2003). It relates to a key aspect of metacognition: If participants possessed any knowledge about their performance in the task, their subjective reports about the task should differentiate between correct and erroneous trials. Consequently, measures of metacognitive sensitivity are relevant for all research areas where quantifying participants' insight into their task performance is of interest, including consciousness research (Dienes, 2004). Given the theoretical importance of metacognitive sensitivity, a universally accepted measure is desirable. However, various competing measures of metacognitive sensitivity were proposed in the literature:

- gamma correlation coefficients (Nelson, 1984),
- a'/type 2 d' (Kunimoto, Miller, & Pashler, 2001)
- type-2 receiver operating characteristic (Fleming, Weil, Nagy, Dolan, & Rees, 2010)
- meta-d' (Maniscalco & Lau, 2012)
- logistic regression analysis (Sandberg, Timmermans, Overgaard, & Cleeremans, 2010)

Logistic regression has been widely used in empirical studies as measure of the association between verbal reports and task accuracy (Rausch, Müller, & Zehetleitner, 2015; Rausch & Zehetleitner, 2014; Sandberg et al., 2010; Siedlecka, Paulewicz, & Wierzchoń, 2016; Wierzchoń, Asanowicz, Paulewicz, & Cleeremans, 2012; Wierzchoń, Paulewicz, Asanowicz, Timmermans, & Cleeremans, 2014). However, while gamma correlations, a', and meta-d' have been extensively examined using both empirical and analytical methods (Barrett, Dienes, & Seth, 2013; Evans & Azzopardi, 2007; Galvin et al., 2003; Masson & Rotello, 2009), the conditions for logistic regression to be an appropriate measure of metacognitive sensitivity have never been systematically investigated.

1.1 Logistic regression as measure of metacognitive sensitivity

Logistic regression is a specific case of a generalized linear regression model (GLM). In general, it is a method to quantify the relationship between a binary outcome variable and one or several dichotomous or continuous predictors. The standard approach to quantify metacognitive sensitivity by means of logistic regression is to model the probability of being correct in the primary task $P(T)$ as a linear function of a subjective report C, e.g. a confidence judgment or a visibility rating. A linear relationship between predictors and outcome is obtained by transforming the probability of being correct into the logarithm of the odds of the primary response being correct to being incorrect:

$$\log\left(\frac{P(T)}{1-P(T)}\right) = a + b * C \qquad (1)$$

As can be seen from Fig. 1, metacognitive sensitivity is indexed by the slope b of the regression line: the steeper the regression line, the stronger are subjective reports associated with the probability of being correct (Sandberg et al., 2010). Logistic regression is also used to quantify the minimal criteria participants apply when they make a subjective report: The more conservative participants' reporting strategy is, the better they perform while still giving the lowest possible subjective report. As the intercept a is just the transformed accuracy when the subjective report is zero, it is interpreted as measure of criterion (Wierzchoń et al., 2012).

Quantifying metacognition by logistic regression is tempting due to three reasons: First, the hierarchical structure of the data often found in behavioral experiments can be explicitly included into the model by using nested random effects (Sandberg, Bibby, & Overgaard, 2013; Siedlecka et al., 2016): For example, two experimental groups with several participants each contributing a number of trials can be described by a random effect of trial nested within a random effect of participant nested within groups. As such an analysis can be conducted on a single trial level without the need for summary statistics to be computed for each participant, logistic regression may also be a promising way to increase statistical power (Sandberg et al., 2010). Second, using random effects allows the data to be unbalanced, i.e. the number of observations can vary between conditions or there can be empty cells in the design matrix (Rausch et al., 2015; Siedlecka et al., 2016). Consequently, slopes on the group level can be obtained even when not all participants made errors in all experimental conditions. This is particularly useful in studies of metacognitive sensitivity because the number of errors may vary heavily between participants and conditions. Finally, it has been argued that logistic regression, unlike SDT-measures, does not make any assumptions about the sources of the evidence involved in making subjective reports (Siedlecka et al., 2016).

However, logistic regression as measure of metacognition may also suffer from at least two drawbacks: First, it depends on the assumption that the relation between subjective reports and transformed accuracy is linear. A non-linear relationship implies that there is no single slope of the regression line that could be interpreted as measure of metacognitive sensitivity. A previous analysis suggested non-linear trends between subjective reports and logit-transformed accuracy occur relatively frequently, although there were also some data sets where the assumption of a linear relationship appeared to be justified (Rausch et al., 2015). It should be noted that this critique only applies to rating scales with more than two response options because two data from two response options always can be connected by a straight line.

The present analysis explores a potential second and more principal problem of logistic regression: An adequate measure of metacognitive sensitivity should depend exclusively on the amount of evidence available for subjective reports, and should not be confounded by other factors such as rating criteria and response biases (Barrett et al., 2013). Although the distinction between slopes and intercepts appears superficially similar to a separation between metacognitive sensitivity and criteria, it has never been systematically investigated what exactly are the assumptions that have to be fulfilled so that slopes are independent from criteria.

1.2 Models of metacognition

A systematic investigation of the impact of rating criteria and primary task criterion on logistic regression models requires a mathematically formulated model that accounts for both task responses as well as subjective reports. One of the most prominent models of perceptual decision making under uncertainty is signal detection theory (SDT) (Green & Swets, 1966; Macmillan & Creelman, 2005; Wickens, 2002). Standard SDT applies to tasks where participants are instructed to correctly classify a binary stimulation S. For the purpose of present analysis, it is not relevant whether the two variants of the stimulation are interpreted as signal and noise or as two different stimuli. Each stimulus provides the participants with sensory evidence which of the two response options he or she should select. Participants select their responses based on a comparison of the sensory evidence with a primary task criterion θ. θ represents the degree to which observers tend towards one response option independent of the sensory evidence. Participants respond 0 if the sensory evidence is smaller than θ and 1 otherwise. As there is noise in the system, the sensory evidence is not always the same at each presentation of the stimulus, but instead is modeled as a random sample out of two distributions, one for each variant of S. If the observer's perceptual system was unable to differentiate between the variants of S, the two distributions would be identical. The more sensitive the observer is to the stimulus, the greater is the distance d between the centers of the two distributions. This distance d can therefore be interpreted as the ability of the observer's perceptual system to differentiate between the two kinds of S. The SDT model can be extended to include subjective reports by assuming that task responses and ratings are considered to form an ordered set of responses such as ''I'm sure it's A'', ''I guess A'', ''I guess B'', ''I'm sure it's B''. The different response options are delineated by a series of criteria, c_1, c_2, …, c_n. Participants select one response out of the set of responses by comparing the sensory evidence against the set of criteria. For example, they respond "I guess A" when the sensory evidence falls between that criterion separating the response ''I'm sure it's A'' from ''I guess A" and that criterion separating "I guess A" from "I guess B". An important implication of the SDT model is that the full evidence available to task responses is also available to subjective reports (Macmillan & Creelman, 2005), which is why the model is sometimes referred to as "ideal observer model" (Barrett et al., 2013). While the SDT model has been successfully applied to a vast number of different experiments over the last decades (Macmillan & Creelman, 2005; Wickens, 2002), in recent years, more recent experiments both found support for the SDT model (Peters & Lau, 2015), but also situations where subjective reports were not as optimal as expected from SDT (Maniscalco & Lau, 2012, 2016) or even better than expected (e.g. Rausch & Zehetleitner, 2016).

1.2.1 Hierarchical model

Measuring metacognition only makes sense in models where metacognition is not necessarily perfect. For the purpose of the present analysis, we consider two models of how the SDT model can be extended to account for imperfect metacognition, the *hierarchical model* and the *independent model*. In both of these two models, the task response is selected by a comparison between sensory evidence and the task criterion, just as in the SDT model. A summary of all free parameters of the two models is found in Table 1.

TABLE 1 ABOUT HERE

The hierarchical model assumes that the sensory evidence involved in selecting the task response is also involved in selecting a rating category (see Fig. 2a). However, in contrast to standard SDT, it is not assumed that the sensory evidence involved in the task response completely determines the subjective report as well. Instead, the sensory evidence is read out by metacognitive processes, whereas the read-out can be incomplete or distorted (Maniscalco & Lau, 2016). One way to express this mathematically is by assuming that the sensory evidence is overlaid by random additive noise, characterized by its standard deviation, σ (cf. Maniscalco & Lau, 2016). The additive noise σ can interpreted as the degree of distortion between the task process and the metacognitive processes. When the additive noise is absent, the model is identical to the SDT model. For simplicity, we also assume that two rating criteria are placed symmetrically around the primary task criterion θ. The distances of the rating criteria to the task criterion are controlled by the same parameter τ. Conceptually, the spread of rating criteria τ expresses whether subjective reports are made more liberally or more conservatively: When τ is large, the rating criteria are further away from the center of the distribution of sensory evidence. Thus, it is rather unlikely that the sensory evidence is more extreme than the rating criteria; therefore, participants will not often report high confidence.

1.2.2 Independent model

The present study will propose that logistic regression is closely related to another model of metacognition, the *independent model*. The independent model is a new variant of so-called dual channel models (cf. Cul, Dehaene, Reyes, Bravo, & Slachevsky, 2009; Maniscalco & Lau, 2016; Rahnev, Maniscalco, Luber, Lau, & Lisanby, 2012). Dual channel models assume that the rating decision does not depend on the sensory evidence considered for the task decision. Instead, rating decisions are based on a second sample of evidence (see Fig. 2b). In contrast to previous flavors of the dual channel model, the independent model discussed here assumes that there is no interaction between the two samples of evidence except that observers know which of two the task responses they choose. As a consequence, they respond high confidence or high visibility when the evidence sampled in parallel confirms the response option they decide for. It is assumed that the distributions from which the parallel samples of evidence are drawn are characterized by the same shape as those distributions of evidence involved in the primary task, except that the distance between the two distributions may deviate from the distance of distributions in the primary task and is denoted by the rating sensitivity parameter d_m. Conceptually, the rating sensitivity parameter d_m can be interpreted as the amount of evidence available to metacognitive processes that predict whether a decision is correct. Again, we assume that rating criteria are placed symmetrically around the primary task criterion θ, with the distances of the rating criteria to the task criterion controlled by the same parameter τ. The independent model is able to accommodate patterns of data that the hierarchical model struggles to explain: First, the hierarchical model cannot account for participants successfully detecting their own errors (Yeung & Summerfield, 2012). Second, the independent model is able to account for blind insight, e.g. cases when participants perform at chance, but their confidence responses are able to differentiate between correct and incorrect trials (Scott, Dienes, Barrett, Bor, & Seth, 2014). While the hierarchical model achieved better fits to the data than several variants of dual channel models in a metacontrast masking task (Maniscalco & Lau, 2016), the independent model has never been formally assessed with empirical data.

1.2.3 Rationale of the present study

The present analysis was performed to investigate the eligibility of logistic regression as a measure of metacognition. For this purpose, we computed analytically whether logistic regression slopes depend on parameters conceptually associated with metacognition, i.e. the internal noise σ in the hierarchical model and the distance between distributions d_m in the independent model. To test whether logistic regression slopes is biased by task and rating criteria, we varied task bias θ and the spread of rating criteria τ in both models, and investigated the effects on logistic regression slopes. If logistic regression slopes were suitable measures of metacognitive sensitivity, they should be associated with internal noise in the hierarchical model and rating sensitivity in the independent model, and also be independent from task bias and report criteria. To examine the generality of these effects, we also varied the shape of the distributions of evidence generated by the two stimuli. In addition, we varied the link functions, i.e. the transformations to relate subjective reports and task performance. In addition, we performed an analogous analysis to investigate if rating criteria can be assessed by regression intercepts. Finally, data obtained in a low-contrast orientation discrimination task (Rausch & Zehetleitner, 2016) was reanalyzed to investigate if it is possible to differentiate between the hierarchical and the independent model empirically.

2 Logistic regression in a hierarchical model of metacognition

All analyses were conducted using the free software R (R Core Team, 2014). The analysis code and all reported results are freely available at the Open Science Framework (https://osf.io/72aqe/) to facilitate reproduction of the present study and replication of its results (Ince, Hatton, & Graham-Cumming, 2012; Morin et al., 2012).

2.1 Calculation of GLM slopes

Analytical closed-form solutions exist for the coefficients of logistic regression when there is one categorical predictor (Lipovetsky, 2015), but this approach generalizes to other GLMs as well. In case of metacognition, the GLM is given by the formula

$$g(P(T = 1|C)) = a + b * C \tag{2}$$

with g denoting the link function, which describes the transformation used to relate accuracy and predictors, $P(T = 1|C)$ the probability of being correct in the primary task conditioned on participant's subjective report, a as intercept, b as slope, and C ∈ {0, 1} as subjective report. The intercept is

$$a = g(P(T = 1|C = 0) \tag{3}$$

and the slope, which is indicative of metacognitive sensitivity, is given by

$$b = g(P(T = 1|C = 1)) - g(P(T = 1|C = 0)) \tag{4}$$

Consequently, GLM slopes can be computed analytically in all situations when $P(T = 1|C = 0)$ and $P(T = 1|C = 1)$ are known.

2.2 Model description

A graphical model of the hierarchical model is found in Fig. 3. We assume that in each trial of the experiment, participants are presented with one out of two manifestations of the stimulus S ∈ {0, 1} which both occur with the same probability. Participants select a response R ∈ {0, 1} which of the two stimuli was presented. Accuracy of the response T is 1 when $S = R$, and $T = 0$ otherwise. In each trial, participants select a response based on a single sample of sensory evidence x. The sensory evidence x is a random sample out of a distribution that depends on the stimulus. The two distributions corresponding to the two stimulus variants have the same standard deviation of 1. However, their location depends on the sensitivity parameter d: When $S = 0$, the mean of the distribution is -0.5 d, and when $S = 1$, then the mean of the distribution of evidence is 0.5 d. Participants' response R is 0 when x is smaller than the primary task criterion θ, and 1 otherwise. The sensory evidence x is overlaid by internal noise, which is randomly sampled out of a distribution with a mean of 0 and a standard deviation determined by the rating noise parameter σ. The sum of x and internal noise is called decision variable z and determines subjective reports by a comparison with rating criteria c_0 and c_1: When $R = 1$ and the z is greater than the rating criterion c_1, then the subjective report C is 1. When $R = 1$ and the z is smaller than the rating criterion c_1, then the subjective report C is 0. Likewise, when $R = 0$, than C is 1 if z is smaller than the rating criterion c_0, and C is 0 if z is greater than the rating criterion c_0. For simplicity, we assume that the distance between θ and c_0 is the same as the distance between θ and c_1. This distance is controlled by the parameter τ, which reflects the conservativeness of rating criteria. The formulae for computing $P(T = 1|C = 0)$ and $P(T = 1|C = 1)$ in the hierarchical model can be found in the Appendix A.

We considered two different distributions of evidence, the Gaussian distribution and the logistic distribution. The Gaussian distribution is often motivated by the central limits theorem and the averaging of many events (DeCarlo, 1998). The logistic distribution can be motivated from Choice Theory (Luce & Suppes, 1965; Macmillan & Creelman, 2005). For most SDT applications, results obtained based on logistic and Gaussian distributions are very similar (DeCarlo, 1998; Wickens, 2002).

2.3 Link functions

The link function that transforms the probability of being correct to the logarithm of the odds of being correct to being incorrect is called the logit link and is the defining feature of logistic regression. In the present analysis, three different link functions are examined:

1) the logit link
2) the probit link
3) the half-logit link.

The logit link was the default choice in previous studies. The probit link was included into the analysis because SDT models can be seen as a subclass of generalized linear regression models where the inverse link function corresponds to the cumulative distribution function of the evidence (Brockhoff & Christensen, 2010; DeCarlo, 1998). Logistic regression assumes a logistic distribution of errors, while probit regression assumes a standard normal distribution of errors. Consequently, it appears necessary to examine if logistic regression is only valid in logistic SDT models, and probit regression in Gaussian SDT models. The adjusted link was proposed as an adjustment in tasks with a finite guessing probability (Brockhoff & Müller, 1997). The use of the logit and probit transform implies that the probability of being correct varies between 0 and 1; however, in binary tasks, the probability of being correct is bounded by the guessing probability 0.5 (Rausch et al., 2015). To account for the guessing probability, a link function can be used to ensure that the transformed accuracy is free to vary in the full range between -∞ and ∞. In tasks with two choices, this can be achieved by the adjusted link function $\mathrm{g(x)} = \log((\mathrm{x} - 0.5)/(1 - \mathrm{x}))$. This function was referred to as half-logit link (cf. Williams, Ramaswamy, & Oulhaj, 2006).

2.4 Results

To investigate if GLM slopes are sensitive to metacognition and unbiased by primary task criterion and by rating criteria assuming the hierarchical model, we calculated logistic regression slopes as a function of internal noise σ as well as primary task criterion θ (Fig. 4) and spread of rating criteria τ (Fig. 5). These calculations were repeated using Gaussian and logistic distributions of evidence and with logit, probit, and half-logit link functions.

2.4.1 Are GLM slopes sensitive to metacognition?

In Fig. 4 and 5, separate lines indicate different degrees of internal noise σ. Greater amounts of internal noise are associated with lower regression slopes at each level of primary task criterion (Fig. 4), and at each level of rating criteria spread (Fig. 5). This pattern holds independently from distributions of evidence and link functions (separate panels of Fig. 4 and

5). When the amounts of noise are extreme, i.e. when metacognition is effectively absent, regression slopes also tend towards zero. Overall, GLM analysis is sensitive to metacognition according to the hierarchical model.

2.4.2 Do GLM slopes depend on the primary task criterion?

As can be seen from each panel of Fig. 4, regression slopes depend heavily on the primary task criterion according to the hierarchical model. The precise form of the relationship between primary task criterion and slopes depends on a complex interaction between the amount internal noise, the shape of the distributions of evidence, and the link functions. When the distributions are logistic, when the half-logit transform is used, or when the internal noise is not small, logistic regression slopes increase monotonously with primary task criterion: Consequently, greater regression slopes are not necessarily due to metacognition, but could also be due to a stronger bias towards one of the task alternatives. However, when Gaussian distributions are assumed and the amount of noise is small, the relationship between regression slopes and primary task criterion is u-shaped: Slopes are maximal when observers are either not biased at all or extremely biased towards one of the task alternatives. Therefore, a primary task criterion may not only increase, but also decrease regression slopes. Overall, regression slopes depend on the primary task criterion, but the direction and magnitude of the effect is strongly dependent on the other model parameters of the hierarchical model.

FIG.4 ABOUT HERE

2.4.3 Do GLM slopes depend on rating criteria?

As can be seen from Fig. 5, logistic and probit regression slopes increase with rating criteria spread τ, i.e. when more conservative rating criteria are set (top and central panels). The relationship between half-logit regression slopes and rating criteria can be u-shaped or decreasing (bottom panels). However, the relationships between slopes and rating criteria are moderated by internal noise as well the shape of the distributions: When the distributions are Gaussian, the effect imposed by rating criteria will be smaller the more internal noise is superimposed on the sensory evidence. When the distributions are logistic, the effect imposed by rating criteria is maximal at medium level of internal noise. Overall, according to hierarchical models, differences between regression slopes can not only be caused by metacognition and task criterion, but also by the way participants set rating criteria. Again, the effect is strongly dependent on the other model parameters of the hierarchical model.

FIG. 5 ABOUT HERE

2.4.4 Can rating criteria be assessed by regression intercepts?

To investigate if regression intercepts are sensitive to rating criteria and independent from metacognition and primary task criterion in the hierarchical model, we calculated intercepts as a function of internal noise σ, primary task criterion θ, rating criteria spread τ, shape of the distributions, and link functions. Consistent across distributions and link functions, intercepts were negatively related to the primary task criterion θ and positively correlated to the internal noise sigma σ (see Supplementary Fig. 1). However, intercepts were only sensitive to the spread of rating criteria τ when the amount of internal noise was low (see Supplementary Fig. 2). This means that intercept effects cannot uniquely be attributed to rating criteria, but also to metacognition or due to primary task criterion. Even more, when rating criteria are different between two conditions, it will not be possible to detect the effect using intercepts when the amount of internal noise is high.

3 Logistic regression in the independent model of metacognition

3.1 Calculation of GLM slopes and link functions

Calculation of GLM slopes and link functions were identical to the hierarchical model.

3.2 Model description

The independent model can be expressed by the graphical model in Fig. 6. It is analogous to the hierarchical model with the following differences: Participants select only the primary task response based on the sensory evidence x. For subjective reports, a second sample of sensory evidence y is created, which is stochastically independent from x. The standard deviation of the distribution of y is assumed to be 1. The location of the distribution of y depends on the stimulus in the current trial as well as on the rating sensitivity parameter d_m: When $S = 0$, the mean of the distribution is -0.5 d_m, and when $S = 1$, then the mean of the distribution of evidence is 0.5 d_m. When $R = 1$ and y is greater than the rating criterion c_1, then the participant's subjective report C is 1. When $R = 1$ and y is smaller than the rating criterion c_1, then the subjective report C is 0. Likewise, when $R = 0$, the C is 1 if y is smaller than the rating criterion c_0, and C is 0 if y is greater than the rating criterion c_0. Again, it is assumed for simplicity that the distance between θ and c_0 is the same as the distance between θ and c_1, controlled by the parameter τ reflecting the conservativeness of rating criteria. The formulae for computing $P(T = 1|C = 0)$ and $P(T = 1|C = 1)$ in the independent model can be found in the Appendix B.

3.3 Results

To investigate if GLM slopes are sensitive to metacognition and unbiased by the primary task criterion and by rating criteria according to the independent model, slopes were calculated as a function of rating sensitivity d_m as well as primary task criterion θ (Fig. 7) and spread of rating criteria τ (Fig. 8). Again, these calculations were performed using Gaussian and logistic distributions of evidence and using logit, probit, and half-logit link functions.

3.3.1 Are GLM slopes sensitive to metacognition?

In Fig. 7 and 8, separate lines indicate different degrees of rating sensitivity d_m. Greater rating sensitivity was associated with increasing regression slopes at each level of primary task criterion (Fig. 7), and at each level of rating criteria spread (Fig. 8). This pattern holds across the different distributions of evidence and link functions (separate panels of Fig. 7 and 8). When rating sensitivity is 0, i.e. when the sensory evidence available to metacognition does not differentiate between the two stimuli, regression slopes become 0 when observers are unbiased towards the two response options. However, when rating sensitivity is 0 and when there is a bias, the slope will become negative. Overall, these results indicate GLM analysis is sensitive to metacognition in the independent model, but the sign of the slope should not be interpreted without consideration of the primary task criterion.

3.3.2 Do GLM slopes depend on the primary task criterion?

As can be seen from each panel of Fig. 7, slopes are influenced by the primary task criterion θ according to the independent model as well. While there is always an effect of primary task criterion, direction and magnitude of the effect depends on a complex interaction between the amount of metacognition as indexed by the rating sensitivity d_m, the shape of the distributions of evidence, as well as the link functions. When the distributions of evidence are Gaussian and when a probit or logit link function is used, the relationship between primary task criterion θ and slopes appears to be u-shaped and similar across different levels of metacognition: Slopes reach a minimum at medium primary task criterion θ, and increase when θ is either 0 or maximal (see Fig. 7 upper and central panel to the left). When the distributions of evidence are logistic and when a probit or logit link functions is applied, GLM slopes decrease with increasing θ (see Fig. 7 upper and central panel to the right). When the half-logit link function is used, the slopes increase exponentially with θ for medium-to-large rating sensitivities. However, when rating sensitivity is low, slopes decrease with increasing θ. Overall, these observations indicate that slopes as measures of metacognition in the independent model can be biased by the primary task criterion; the kind of bias however depends on the other model parameters as well as on the choice of the link function.

FIG.7 ABOUT HERE

3.3.3 Do GLM slops depend on rating criteria?

Fig. 8 shows that GLM slopes are independent from the spread of rating criteria τ in one case: When the evidence is assumed to be logistically distributed, logistic regression slopes are independent from rating criteria (see Fig. 8 upper right panel). Probit regression slopes decrease when more conservative rating criteria are used. In contrast, when the evidence is assumed to be Gaussian, both logistic and probit regression slopes increase when more conservative rating criteria are applied (see Fig. 8 upper and central panel to the left). These effects are moderated by the amount of evidence available to ratings, i.e. rating sensitivity d_m. The larger d_m is, the more pronounced is the effect of rating criteria on GLM slopes. For the half-logit link function, slopes increase massively when the spread of rating criteria is very small (see Fig. 8 lower row). In summary, logistic regression slopes are unbiased by the spread of the rating criteria in the independent model when the distributions of evidence are logistic. When other link functions and Gaussian distributions are assumed, slopes can be heavily influenced by rating criteria.

FIG. 8 ABOUT HERE

3.3.4 Can rating criteria be assessed by regression intercepts?

To investigate if regression intercepts are sensitive to rating criteria and independent from metacognition and primary task criterion in the independent model, we calculated intercepts as a function of rating sensitivity d_m, primary task criterion θ, rating criteria spread τ, shape of the distributions, and link functions. Consistent across distributions and link functions, intercepts were negatively associated with the primary task criterion θ and positively associated with the rating sensitivity d_m (see Supplementary Fig. 3). The intercepts were only sensitive to the spread of rating criteria τ when d_m was above zero (see Supplementary Fig. 4). Overall, this means that according to the independent model – just as the hierarchical model - intercept effects could be due to rating criteria, degree of metacognition, and primary task criterion. In addition, true effects on rating criteria will remain undetected when metacognition is low.

4 Model fits of the independent and the hierarchical model in a low-contrast orientation discrimination task

The present analysis implies that logistic regression slopes are unbiased by rating criteria according to only one specific model of metacognition, namely the logistic independent model. In all models, the slopes are dependent on primary task criteria. Consequently, if researchers intend to use logistic regression as measure of metacognition, it would be useful to identify the cognitive model underlying the rating data. While measures of primary task criteria are readily available from SDT (Green & Swets, 1966; Macmillan & Creelman, 2005; Wickens, 2002), it might be a challenge to differentiate the independent model and logistic distributions from the hierarchical model and Gaussians. We reanalyze confidence ratings obtained in a recent low-contrast orientation discrimination experiment (Rausch & Zehetleitner, 2016) to investigate if this is possible using cognitive modeling and the maximum likelihood procedure.

4.1 Reanalysis

4.1.1 Experimental task

20 participants, all of which provided written informed consent, performed one training block and nine experimental blocks of 42 trials each of a low contrast orientation discrimination task. First, participants were presented with a fixation cross for 1 s. Then, the target stimulus, a binary grating oriented either horizontally or vertically, was presented for 200 ms with varying contrast levels of 0, 2.2, 3.9, 5.0, 5.5, and 6.9%. The screen remained blank afterwards until participants made a non-speeded discrimination response by key press whether the target had been horizontal or vertical. After each discrimination response, participants made two subjective reports, one regarding their visual experience of the stimulus, and one regarding their confidence in being correct in the discrimination task. For that, each question was displayed on the screen, which was: ‘‘How clearly did you see the grating?’’ or ‘‘How confident are you that your response was correct?’’ The sequence of questions was balanced across participants. Participants delivered subjective reports on a visual analog scale using a joystick, which means that participants selected a position along a continuous line between two end points by moving a cursor. The end points were labeled as ‘‘unclear’’ and ‘‘clear’’ for the experience scale and ‘‘unconfident’’ and ‘‘confident’’ for the confidence scale, i.e. observers indicated their experience or confidence by the selected cursor position on the continuous scale. If the discrimination response was erroneous, the trial ended by displaying the word ‘‘error’’ for 1 s on the monitor. There was no feedback with respect to the subjective report. Please refer to Rausch and Zehetleitner (2016) for a more detailed description of the experiment.

4.1.2 Models

We fitted eight different models to the data, which were characterized by all possible combinations of the following three features:

(i) The model could be either a hierarchical or an independent model as outlined above.
(ii) The distributions of evidence and noise could be either Gaussian or logistic.
(iii) The primary task criterion θ was either treated as a free parameter or fixed at 0.

In all eight models, we assumed that the discrimination sensitivity d varied across contrast levels, while the other parameters were assumed to be constant across contrast levels. Thus, each model involved six different sensitivity parameters d_1 - d_6, one for each contrast level. Moreover, each model involved a series of 11 rating criteria spread parameters τ_1 - τ_{11}. These parameters described how close the rating criteria were located to the primary task criterion θ: τ_1 denotes the location of the closest pair of criteria at both sides of θ; τ_2 referred to the second pair, and so on. In hierarchical models, the degree of metacognition was denoted by the internal noise parameter σ. For independent models, the rating sensitivity d_m was assumed to be a constant fraction of the discrimination sensitivity d, denoted by the parameter a. Overall, the models had 18 or 19 free parameters depending on if the primary task criterion θ was fixed at zero.

4.1.3 Model fitting

Model fitting was performed separately for each single participant. The fitting procedure involved the following computational steps. First, the continuous confidence ratings were discretized by dividing the continuous scale into equal 12 partitions. Analyses using four or eight bins gave similar results when used on the empirical data; however, 12 bins improved the recovery of model-generated data (see 4.1.6), which is why results based on 12 bins are reported. Second, we computed the frequency of each rating bin given orientation of the stimulus and the orientation response. Third, for each of the 8 models, the set of parameters was determined that maximized the likelihood of the data. To compute the likelihood, we made two widespread assumptions in SDT modeling (Dorfman & Alf, 1969; Maniscalco & Lau, 2016): (i) responses in each trial were assumed to be independent from each other and (ii) the joint probability of a task response and a subjective report given the stimulus was constant across trials. Formally, the likelihood of a set of parameters given primary task responses and subjective reports $\mathcal{L}(\mathrm{p}|R,C)$ is given by

$$\mathcal{L}(\mathrm{p}|R,C) \propto \prod_{i,j,k} P(R_i, C_j|S_k, p)^{n(R_i, C_j|S_k)} \tag{5}$$

where $P(R_i, C_j|S_k, p)$ denotes the probability of a primary task response in conjunction with a specific subjective report given the stimulus and the set of parameters, and $n(R_i, C_j|S_k)$ indicates the frequency how often the participants gave a specific response in conjunction with a specific subjective report given the stimulus. The set of parameters with the maximum likelihood was determined by minimizing the negative log likelihood as the latter is computationally more stable:

$$\log\big(\mathcal{L}(\mathrm{p}|R,C)\big) \propto \sum_{i,j,k} \log\Big(P\big(R_i, C_j\big|S_k, p\big)\Big) \times n\big(R_i, C_j\big|S_k\big) \tag{6}$$

To minimize the negative log likelihood, we used a general SIMPLEX minimization routine implemented in the R function optim (Nelder & Mead, 1965). The formulae for $\mathrm{P}(\mathrm{R_i}, \mathrm{C_j}|\mathrm{S_k}, \mathrm{p})$ are found in the Appendix as formulae (A6) – (A13) and (B5) – (B12). Internal noise was parametrized as the log of the standard deviation of the noise to allow negative values of the parameter during the fitting process. To maintain a fixed sequence of rating criteria, we did not directly fit $\tau_1 - \tau_{11,}$ instead, optimization was performed on the log distance to the nearest rating criterion closer to the primary task criterion.

4.1.4 Model selection

Following the fitting procedure, we assessed the relative quality of the eight candidate models using the Bayes information criterion (BIC, Schwarz, 1978) and the Akaike information criterion (AIC, Akaike, 1974). Conceptually, the BIC measures the degree of belief that a certain model is the true data-generating model relative to the other models under comparison, assuming that the true generative model is among the set of candidate models. In contrast, the AIC measures the loss of information when the true generative model is approximated by the candidate model. We used AIC_c, a variant of AIC that corrects for finite sample sizes (Burnham & Anderson, 2002). BIC and AIC_c take into account descriptive accuracy (i.e. goodness of fit) and parsimony (i.e. smallest number of parameters), but the BIC favors parsimony more heavily than the AIC_c does. BIC and AIC_c are given by the following formulae:

$$BIC = -2 \ \log\left(\mathcal{L}(\mathrm{p}|R, C)\right) + k \log(n) \tag{7}$$

$$AIC_c = -2 \ \log\left(\mathcal{L}(\mathrm{p}|R, C)\right) + 2k + \left(\frac{2k(k+1)}{(n-k-1)}\right) \tag{8}$$

where k indicates the number of parameters and n the number of observations.

4.1.5 Statistical testing

In experiments with a standard number of trials, even if the independent logistic model with fixed task criterion was true, it cannot be expected that models can be correctly identified for each single participant. However, in this case, it would still be expected that independent models obtained the best fit more frequently than hierarchical models, that models based on the logistic distribution would achieve the best fit more often than Gaussian models, and likewise that models with the free primary task criterion fixed at 0 would obtain better fits than models with a free primary task criterion.

Therefore, we determined for each participant which of the eight candidate models achieved the minimal BIC and minimal AIC_c. Then, we performed three tests if those model features that imply that logistic regression slopes are independent from criteria are more likely to result in the best BIC or AIC_c: First, we tested if independent models were more likely to achieve the best BIC or AIC_c than hierarchical models. Second, we assessed if models with the primary task criterion fixed at 0 achieved the best BIC/AIC_c with a greater probability than models with the primary task criterion as free parameter. Finally, we examined if models based on logistic distributions attained minimal BIC and AIC_c more frequently than models based on Gaussians.

Statistical testing was based on Bayes factors for proportions implemented in the R library BayesFactor (Morey & Rouder, 2015). Bayes factors provide continuous measures of how the evidence supports the alternative hypothesis over the null hypothesis and vice versa (Dienes, 2011; Rouder, Speckman, Sun, Morey, & Iverson, 2009). Specifically, the Bayes factor indicates how the prior odds about alternative hypothesis and null hypothesis need to be multiplied to obtain the posterior odds of the two hypotheses. As null hypothesis, we assumed that both variants of the model were equally likely to achieve the best fit. As one-sided alternative hypothesis, we assumed a logistic distribution of the logits of the probability around 0 over the interval 0.5 and 1 with a scale parameter of 0.5.

4.1.6 Recovery of model generated data

To investigate the reliability of the model selection process, we generated random data sets based on each participant’s parameter set obtained during the fitting process (analogous procedure to Maniscalco & Lau, 2016). We used only the parameter sets of two models: The independent logistic model with primary task criterion fixed at 0 was used because it implies that slopes are independent of confounds by criteria and thus logistic regression slopes can be used as measure of metacognition without concern. The alternative model was the hierarchical Gaussian model with free primary task criterion because it is maximally different from the independent logistic model with primary task criterion fixed at 0. This procedure yielded a mock replica of each participant's behavioral data. We then fitted all eight models to these simulated data and performed model selection using AICc and BIC. If the model selection methodology is reliable, then independent, logistic and fixed θ models should be selected when the data is generated according to the independent logistic model with primary task criterion fixed at 0. Likewise, when the hierarchical Gaussian model with free primary task criterion is used to create the data, hierarchical, Gaussian models with a free primary task parameter should be preferred during model selection. We replicated the analysis using 100, 200, 378 (the same number of trials as in the real experiment), 600, 1200, 2500, 5000, 10000, 20000 and 50000 trials to estimate the number of trials required to perform a reliable model selection.

4.2 Results

The fitting procedure converged for all participants except for one participant, where the fitting of the hierarchical Gaussian model did not converge. On average, the best model fit was obtained by the hierarchical logistic model with fixed primary task criterion both in terms of BIC (M = 1432.8) as well as AICc (M =1363.7). According to the BIC, the strength of evidence in favor the hierarchical logistic model with fixed primary task criterion as indexed by the difference in BIC was always substantial (all other models: M's ≥ 1437.7). In contrast, according to the AICc, there was only a small benefit of the hierarchical logistic model with fixed primary task criterion compared to the independent logistic model with free task criterion parameter (M = 1364.5) and the hierarchical logistic model including a free task criterion parameter (M = 1364.9). As can be seen from Fig. 9, both the hierarchical logistic model with task criterion fixed at zero and the hierarchical logistic model including a free task criterion parameter provided qualitatively good accounts of the distributions of discrimination accuracy and confidence ratings.

4.2.1 Is there support for the independent model?

Independent models of metacognition only provided the best model fit in 30% of the participants according to the BIC, and in 40% of the participants according to the AIC_c. The Bayes factor analysis indicated evidence against the hypothesis that independent models are more likely to attain the best fit both in terms of BIC, $BF_{10} = 0.19$, as well as in terms of AIC_c, $BF_{10} = 0.29$. This result implies that prior beliefs that a flavor of the independent models is the generative model of the present data should be attenuated. Accordingly, estimates of metacognition in the present data set based on GLM slopes are likely to be affected by rating criteria.

4.2.2 Can the primary task criterion be fixed at 0?

Models with the primary task criterion fixed at zero attained the best model fit in 60% of the participants according to the BIC, and in 40% according to the AIC_c. The Bayes factor analysis does not provide any support in favor of the hypothesis that models with the primary task criterion fixed at zero are associated with better BICs, $BF_{10} = 1.02$. However, there was evidence against the hypothesis that models with the primary task criterion fixed at zero are associated with better AIC_cs, BF_{10} = 0.29. Consequently, there is some indication that GLM slopes in the present data would also influenced by the primary task criterion, but the evidence is not consistent.

4.2.3 Is the evidence distributed logistically?

Models based on the assumption of logistic distributions achieved the minimal BIC in 85% of the participants according to BIC and even in 90% according to $AIC_{c.}$. The Bayes factors indicated strong evidence that the logistic models were more likely to produce better fits than Gaussian models, both in terms of BIC, BF_{10} = 56.1, as well with regard to AIC_c, BF_{10} = 231.6. This means that only one out of the three conditions for logistic regression slopes to be independent of criteria was met in the present data set.

4.2.4 Can the model underlying simulated data be recovered?

Fig. 10 shows the results of the model recovery analysis of data simulated according to the independent logistic model with fixed task criterion (squares) and data generated according to the hierarchical Gaussian model with a free task criterion (circles).

As can be seen from the panels on the left, when the data was simulated according to the independent logistic model with fixed task criterion, the model was nearly always correctly identified as being one of the independent models. However, when the true model was the hierarchical Gaussian model with free task criterion, the model was relatively often misclassified as independent: When the trial number was small ($N \leq 200$), model recovery was even below chance. 5000 trials were required to obtain a tolerable recovery rate of approximately 70%. Increasing the trial number even more did not substantially improve model recovery.

The central panels of Fig. 10 show model recovery with respect to the free primary task criterion vs. the primary task criterion fixed at 0. For the BIC as model selection criterion, 5000 trials were necessary to detect the free task criterion parameter with a tolerable accuracy of 70%. In contrast, for the AICc, 1200 trials were sufficient to reach 70%. Notably, the AICc does not favor models with smaller number of parameter as heavily as the BIC does. It can also be seen that model recovery accuracy decreased with trial number for the independent logistic model with fixed task criterion (dotted lines). However, at the same time, model recovery accuracy increased with sample size for the hierarchical Gaussian model with free task criterion (straight lines). An explanation for this pattern may be that AICc and BIC both favor more parsimonious models. For small sample sizes, both AICc and BIC prefer models with a smaller number of parameters, which is why models are selected where the primary task criterion is fixed at 0. The bias towards smaller results will result in a correct classification with respect to the primary task criterion when the true model is the independent logistic model with fixed task criterion, but an error will occur when the true model is the hierarchical Gaussian model with free task criterion.

Finally, model recovery with respect to logistic and Gaussian distributions is depicted in the right panels of Fig. 10. For large sample sizes, logistic distributions were more often correctly identified than Gaussians. However, 1200 trials were sufficient to obtain a tolerable recovery rate of approximately 70% for both distributions.

5 Discussion

The analysis presented here investigated whether GLM slopes as measures of metacognition are biased by the spread of rating criteria and the primary task criterion. We showed analytically that logistic regression slopes are independent from rating criteria only according to one specific model of metacognition: the independent model based on logistic distributions. When other distributions were assumed, when other link functions were used, or when a hierarchical model was adopted, regression slopes always depended on the spread of rating criteria. The direction and magnitude of these effects depended on the other model parameters. The primary task criterion was related to regression slopes in all considered models. Depending on the model parameters, the relationship between slopes and task criterion were increasing, decreasing, or even u-shaped. An analysis of regression intercepts revealed that intercepts were insensitive to rating criteria when the amount of metacognition was too low. In addition, we examined if these models can be identified empirically on an existing data set, observing that a massive number of trials is required to distinguish between hierarchical and independent models with tolerable accuracy.

5.1 Is logistic regression a biased measure of metacognitive sensitivity?

When the aim of a study is to estimate the degree of metacognitive sensitivity, it is generally accepted that a suitable measure should be independent from the primary task criterion and rating criteria (Barrett et al., 2013). However, the present study revealed that logistic regression slopes depend on the primary task criterion independent of the underlying model of metacognition. Logistic regression slopes also depend on rating criteria except for one specific model of metacognition: When the sensory evidence considered for primary task responses is stochastically independent from the sensory evidence used in rating decisions, and when these two types of evidence both form logistic distributions, then logistic regression slopes are independent from rating criteria. This means that when researchers encounter an empirical effect on logistic regression slopes, without knowledge about the underlying model of metacognition, there will be at least three possible explanations for the effect: (i) the effect can be mediated by participants' degree of metacognition of the processes engaged in performing the task, (ii) the effect might also be due to participants' bias towards one of the task alternatives, and (iii) the effect may also depend entirely on differences how liberal or conservative participants' rating criteria are. Likewise, researchers might be unable to observe real effects on participants' degree of metacognition when there is also a difference between participants' bias or between rating criteria because the effects of metacognition and criteria could balance out.

5.2 Does the present analysis generalize to other models of decision-making and metacognition?

The present analysis was based on specific assumptions about the decision process as well as the cognitive model underlying metacognition. Concerning the decision process, we assumed the standard SDT model. Concerning metacognition, only two models, the hierarchical and the independent model were considered. Is it reasonable to assume that the characteristics of GLM slopes outlined here generalize to other models of decision-making and metacognition?

As the literature provides a multitude of competing models of decision making, confidence and / or metacognition, it is not feasible to investigate the eligibility of logistic regression for each model proposed in the literature. Examples for different models include the bounded accumulation model (Kiani, Corthell, & Shadlen, 2014), the collapsing confidence boundary model (Moran, Teodorescu, & Usher, 2015), the consensus model (Paz, Insabato, Zylberberg, Deco, & Sigman, 2016), the reaction time account (Ratcliff, 1978), the self-evaluation model (Fleming & Daw, 2016), and two-stage signal detection theory (Pleskac & Busemeyer, 2010). The attempt seems even futile because the number of models in the literature is continuously increasing. However, the two models tested here represent two complementary prototypes of models of metacognition: The hierarchical model is the simplest possible variant of models where the evidence used for the decision between the primary task alternatives also informs the decision between different rating criteria. The view that the evidence used in the decision process is also involved in metacognition is a standard tenet of theories about metacognition. In contrast, the independent model assumes that evidence used for the rating decision is sampled entirely independently from the evidence used for the task response. It is an open question whether datasets exist that can be conveniently described by the independent model.

The majority of existing models in the literature are closely related to one of the two models or the models constitute a combination of the two models. For example, post-decisional accumulation models can be seen as a combination of the hierarchical and the independent model, where a second, independent sample of evidence is acquired later in time than the sensory evidence used for performing the task (Moran et al., 2015; Pleskac & Busemeyer, 2010). When a model assumes that rating decisions are informed by both evidence considered in the primary task as well as evidence sampled in parallel, it is reasonable to expect that biases apparent in the hierarchical and the independent model persist when the two sources of evidence are combined.

Of course, it is still possible that there will be new theories which imply that logistic regression is not affected by primary task criteria and rating criteria. However, the implications of the present study do not require that such a model does not exist. What the study implies is that according to plausible models of metacognition, logistic regression is affected from primary task criterion and rating criteria. As a consequence, researchers who wish to use logistic regression to measure metacognitive sensitivity need to show that their effects cannot be alternatively explained by rating criteria and primary task criteria.

5.3 Can the independent model be empirically identified?

To exclude the possibility that effects on regression slopes are not caused or masked by rating criteria, it would be useful to identify the underlying model of metacognition. Unfortunately, the present model recovery analysis revealed that the amount of trials required to correctly classify a true underlying hierarchical Gaussian model is massive. Moreover, the hierarchical Gaussian model was still occasionally misclassified as independent or logistic even with extreme trial numbers. In a similar comparison between different models of metacognition, two thirds of the participants were excluded to reduce noise in the data and improve model selection, implying that these models are also not trivial to distinguish based on other data sets (Maniscalco & Lau, 2016). Likewise, Gaussian and logistic distributions are known to produce similar results in many applications (DeCarlo, 1998; Wickens, 2002). Consequently, when researchers intend to model the cognitive architecture of metacognition, they will need to ensure that both the sample size and the trial number are sufficiently large to ensure that classification errors are outnumbered by correct classifications. However, for those researchers who are only interested in measuring metacognition, the standard application of logistic regression as measure of metacognitive sensitivity, cognitive modeling will usually not be a feasible option because the number of trials in standard experiments is typically too small. Future studies might be able to provide more efficient methods to distinguish between the hierarchical Gaussian and the independent logistic model.

5.4 What other methods can be used to estimate metacognitive sensitivity?

What are the options to avoid the confound of metacognitive sensitivity with task criteria and rating criteria? There are two options: (i) researchers can resort to measures of metacognitive sensitivity other than GLM slopes, or (ii) they can control for primary task and rating criteria statistically.

What are the alternatives to logistic regression slopes (cf. Fleming & Lau, 2014)? A method that recently received a considerable amount of attention is meta-d'. Meta-d' quantifies the degree of metacognition in terms of a standard SDT model where task responses and subjective reports are made based on identical evidence (Maniscalco & Lau, 2012). There are several reasons to use meta-d': Meta-d' is reasonably robust to changes of primary task criteria and rating criteria in an optimal observer SDT model, a decreasing signal SDT model and an increasing signal SDT model (Barrett et al., 2013). The decreasing signal SDT model is closely related to the hierarchical model in the present study, while the increasing signal SDT model can be seen as combination of the hierarchical model and the independent model in the present study. In addition, meta-d' provides control over performance in the primary task. Meta-d' has even the unique advantage of allowing comparisons with primary task performance as metacognitive sensitivity and primary task performance are measured on the same scale (Fleming & Lau, 2014; Maniscalco & Lau, 2012). Thus, meta-d' is able to assess imperfect metacognition as well as metacognition better than expected from task performance.

An argument against the use of meta-d' is that meta-d' requires assumptions about the distributions of evidence during the decision process. The most common choice are equal Gaussian distributions (Barrett et al., 2013; Maniscalco & Lau, 2012), but other distributions have been implemented as well (Rausch et al., 2015). However, to our knowledge, no study so far has investigated how often the distributional assumptions of meta-d' are in fact violated, or how sensitive meta-d' is to violations of these distributional assumptions. Moreover, meta-d' has also never been assessed in an independent model of metacognition.

SDT approaches that rely on distributional assumptions are often criticized because it is often hard to test of these assumptions are justified. Indeed, in a classical paper, Nelson (1984) recommended gamma correlations to avoid the distributional assumptions made by parametric SDT methods. Although many researchers have since used gamma correlations, simulations suggested that rating criteria strongly impact on gamma correlations, making results obtained by gamma ambiguous as they could be due to metacognition or due to criterion setting (Masson & Rotello, 2009).

An non-parametric SDT approach to estimate metacognitive sensitivity is by means of type 2 ROC-curves (Fleming et al., 2010). This method provides a measure of metacognitive sensitivity free of bias by rating criteria and distributional assumptions. However, the number of trials required to estimate type 2 ROC-curves can be massive (Nelson, 1984), and type 2 ROC curves do not provide any control over performance in the primary task and the associated criteria (Fleming & Lau, 2014).

Overall, it appears to us that meta-d' is the most useful measure of metacognitive sensitivity, although more research on the distributional assumptions required by meta-d' would be desirable. Meanwhile, when the underlying distributions of evidence cannot be ascertained, it may be useful to check if meta-d' and type 2 ROC-curve converge to the same results.

5.5 How can criteria be controlled?

When researchers do not wish to resort to other methods than logistic regression slopes, they should at least control primary task criteria and rating criteria statistically. This approach requires of course the assessment of criteria independent of metacognitive sensitivity. As logistic regression will often be used when researchers do not wish to make explicit assumptions about the underlying model of metacognition, the measure of criteria should also be model-free.

Logistic regression intercepts, which are occasionally used as measure of rating criteria (Wierzchoń et al., 2012), do not fulfill these requirements. According to the present analysis, intercepts also depend on model parameters assumed to reflect the degree of metacognition. Moreover, when metacognition is low, intercepts are no longer sensitive to rating criteria. Overall, rating criteria need to be controlled by other measures than regression intercepts.

A model based approach to assess rating criteria is based on the standard SDT model used to estimate meta-d'. When the standard SDT model is assumed, rating criteria and the primary task criterion can be directly estimated from the model (Rausch & Zehetleitner, 2016). The advantage of this approach is that it allows to control for primary task criterion and rating criteria at the same time. Unfortunately, this approach again requires assumptions about the distributions of evidence. In addition, it has never been investigated if these estimates are unbiased when the data-generating model is not the standard SDT model or a hierarchical model of metacognition.

SDT theory provides numerous other indices for response bias (Green & Swets, 1966; Macmillan & Creelman, 2005; Wickens, 2002). A non-parametric measure of primary task criterion is β_K, which is calculated from the empirical receiver operating characteristic (Kornbrot, 2006). When researchers construct type 2 ROC-curves, they can compute B_{roc}, the analog of β_K for type 2 ROC-curves, as a measure of conservative and liberal rating criteria (Fleming et al., 2010). A disadvantage of these measures is again that the number of trials required for ROC-curves is large. Nevertheless, when distributional assumptions need to be avoided, β_K and B_{roc} appear to be the most promising approach.

It should be noted that in order to control the impact of criteria on regression slopes, it is not sufficient to demonstrate that the mean primary task criterion and the mean rating criteria are the same between two conditions. The reason is that the effects of criteria on GLM slopes often follow non-linear trends. Thus, when the variance of criteria is greater in one condition, the effect of the maximal and minimal criteria will not necessarily balance out. Consequently, slopes can be different between two conditions solely due to different variances of primary task criteria or rating criteria. Researchers who would like to rule out than an effect on slopes is not due to criteria should assess if the distributions of primary task criteria and rating criteria are the same between conditions.

6 Conclusion

Logistic regression slopes as measures of metacognitive sensitivity always depend on the primary task criterion and are independent from rating criteria according to only one specific model of metacognition, namely the independent logistic model. It is argued that researchers who want to quantify metacognitive sensitivity using logistic regression should provide evidence that the underlying model is the independent logistic model where the primary task criterion is fixed at zero. However, this will often not be feasible as number of trials required to allow accurate model selection is massive. Alternatively, they can also control the primary task criteron and rating criteria statistically or use alternative methods to measure metacognitive sensitivity.

7 Acknowledgements

This research was supported by grant ZE 887/3-1 of the Deutsche Forschungsgesellschaft (DFG) (to MZ). The funders had no role in study design, data collection, analysis, decision to publish, or preparation of the manuscript. We are grateful to Sebastian Hellmann for helpful comments.

8 References

Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Cintrol*, *19*(6), 716–723.

Baranski, J. V, & Petrusic, W. M. (1994). The calibration and resolution of confidence in perceptual judgments. *Perception & Psychophysics*, *55*(4), 412–28.

Barrett, A. B., Dienes, Z., & Seth, A. K. (2013). Measures of metacognition on signal-detection theoretic models. *Psychological Methods*, *18*(4), 535–52. http://doi.org/10.1037/a0033268

Brockhoff, P. B., & Christensen, R. H. B. (2010). Thurstonian models for sensory discrimination tests as generalized linear models. *Food Quality and Preference*, *21*(3), 330–338. http://doi.org/10.1016/j.foodqual.2009.04.003

Brockhoff, P. M., & Müller, H. G. (1997). Random effect threshold models for dose-response relationships with repeated measurements. *Journal of the Royal Statistical Society Series B-Methodological*, *59*(2), 431–446.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: a practical information-theoretic approach* (2nd ed.). New York: Springer.

Cul, A. Del, Dehaene, S., Reyes, P., Bravo, E., & Slachevsky, A. (2009). Causal role of prefrontal cortex in the threshold for access to consciousness. *Brain*, *132*, 2531–2540. http://doi.org/10.1093/brain/awp111

DeCarlo, L. T. (1998). Signal detection theory and generalized linear models. *Psychological Methods*, *3*(2), 186–205. http://doi.org/10.1037//1082-989X.3.2.186

Dienes, Z. (2004). Assumptions of Subjective Measures of Unconscious Mental States. *Journal of Consciousness Studies*, *11*(9), 25–45.

Dienes, Z. (2011). Bayesian Versus Orthodox Statistics: Which Side Are You On? *Perspectives on Psychological Science*, *6*(3), 274–290. http://doi.org/10.1177/1745691611406920

Dorfman, D. D., & Alf, E. (1969). Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals—Rating-method data. *Journal of Mathematical Psychology*, *6*, 487–496. http://doi.org/10.1016/0022-2496(69)90019-4

Evans, S., & Azzopardi, P. (2007). Evaluation of a “bias-free” measure of awareness. *Spatial Vision*, *20*(1), 61–77. http://doi.org/10.1163/156856807779369742

Fleming, S. M., & Daw, N. D. (2016). Self-evaluation of decision performance: A general Bayesian framework for metacognitive computation. *Psychological Review*, 1–59. http://doi.org/10.1002/smll.))

Fleming, S. M., & Lau, H. (2014). How to measure metacognition. *Frontiers in Human*

Neuroscience, *8*, 1–9. http://doi.org/10.3389/fnhum.2014.00443

Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., & Rees, G. (2010). Relating introspective accuracy to individual differences in brain structure. *Science*, *329*(5998), 1541–1543. http://doi.org/10.1126/science.1191883

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), 843–76.

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

Ince, D. C., Hatton, L., & Graham-Cumming, J. (2012). The case for open computer programs. *Nature*, *482*, 485–488. http://doi.org/10.1038/nature10836

Kiani, R., Corthell, L., & Shadlen, M. N. (2014). Choice certainty is informed by both evidence and decision time. *Neuron*, *84*(6), 1329–1342. http://doi.org/10.1016/j.neuron.2014.12.015

Kornbrot, D. E. (2006). Signal detection theory, the approach of choice: model-based and distribution-free measures and evaluation. *Perception & Psychophysics*, *68*(3), 393–414.

Kunimoto, C., Miller, J., & Pashler, H. (2001). Confidence and accuracy of near-threshold discrimination responses. *Consciousness and Cognition*, *10*(3), 294–340. http://doi.org/10.1006/ccog.2000.0494

Lipovetsky, S. (2015). Analytical closed-form solution for binary logit regression by categorical predictors. *Journal of Applied Statistics*, *42*(1), 37–49. http://doi.org/10.1080/02664763.2014.932760

Luce, R. D., & Suppes, P. (1965). Preference, utility, and subjective probability. In & E. G. R. D. Luce, R. Bush (Ed.), *Handbook of Mathematical Psychology*. New York: Wiley.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection Theory. A user's guide*. Mahwah, NY: Lawrence Erlbaum Associates.

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, *21*(1), 422–30. http://doi.org/10.1016/j.concog.2011.09.021

Maniscalco, B., & Lau, H. (2016). The signal processing architecture underlying subjective reports of sensory awareness. *Neuroscience of Consciousness*, (November 2015), 1–17. http://doi.org/10.1093/nc/niw002

Masson, M. E. J., & Rotello, C. M. (2009). Sources of bias in the Goodman-Kruskal gamma coefficient measure of association: implications for studies of metacognitive processes. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *35*(2), 509–27. http://doi.org/10.1037/a0014876

Moran, R., Teodorescu, A. R., & Usher, M. (2015). Post choice information integration as a causal determinant of confidence: Novel data and a computational account. *Cognitive Psychology*, *78*, 99–147. http://doi.org/10.1016/j.cogpsych.2015.01.002

Morey, R. D., & Rouder, J. N. (2015). BayesFactor: Computation of Bayes factors for common designs. R package version 0.9.10-1.

Morin, A., Urban, J., Adams, P. D., Foster, I., Sali, A., Baker, D., & Sliz, P. (2012). Shining Light into Black Boxes. *Science*, *336*(6078), 159–160. http://doi.org/10.1126/science.1218263

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, *7*, 308--313.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, *95*(1), 109–133. http://doi.org/10.1037//0033-2909.95.1.109

Paz, L., Insabato, A., Zylberberg, A., Deco, G., & Sigman, M. (2016). Confidence through consensus: a neural mechanism for uncertainty monitoring. *Scientific Reports*, *6*, 21830. http://doi.org/10.1038/srep21830

Peters, M. A. K., & Lau, H. (2015). Human observers have optimal introspective access to perceptual processes even for visually masked stimuli. *eLife*, *4*(OCTOBER2015), 1–30. http://doi.org/10.7554/eLife.09651

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. http://doi.org/10.1037/a0019737

R Core Team. (2014). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.r-project.org/

Rahnev, D. a, Maniscalco, B., Luber, B., Lau, H., & Lisanby, S. H. (2012). Direct injection of noise to the visual cortex decreases accuracy but increases decision confidence. *Journal of Neurophysiology*, *107*(6), 1556–63. http://doi.org/10.1152/jn.00985.2011

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108.

Rausch, M., Müller, H. J., & Zehetleitner, M. (2015). Metacognitive sensitivity of subjective reports of decisional confidence and visual experience. *Consciousness and Cognition*, *35*, 192–205. http://doi.org/10.1016/j.concog.2015.02.011

Rausch, M., & Zehetleitner, M. (2014). A comparison between a visual analogue scale and a four point scale as measures of conscious experience of motion. *Consciousness and Cognition*, *28*, 126–140. http://doi.org/10.1016/j.concog.2014.06.012

Rausch, M., & Zehetleitner, M. (2016). Visibility is not equivalent to confidence in a low contrast orientation discrimination task. *Frontiers in Psychology*, *7*, 591. http://doi.org/10.3389/fpsyg.2016.00591

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–37. http://doi.org/10.3758/PBR.16.2.225

Sandberg, K., Bibby, B. M., & Overgaard, M. (2013). Measuring and testing awareness of emotional face expressions. *Consciousness and Cognition*, *22*(3), 806–9. http://doi.org/10.1016/j.concog.2013.04.015

Sandberg, K., Timmermans, B., Overgaard, M., & Cleeremans, A. (2010). Measuring consciousness: Is one measure better than the other? *Consciousness and Cognition*, *19*(4), 1069–1078. http://doi.org/10.1016/j.concog.2009.12.013

Schwarz, G. (1978). Estimating the dimensions of a model. *The Annals of Statistics*, *6*(2), 461–464. http://doi.org/10.1214/aos/1176348654

Scott, R. B., Dienes, Z., Barrett, A. B., Bor, D., & Seth, A. K. (2014). Blind insight: Metacognitve discrimination depite chance task performance. *Psychological Science*, *25*, 1–20. http://doi.org/10.1177/0956797614553944

Siedlecka, M., Paulewicz, B., & Wierzchoń, M. (2016). But I Was So Sure! Metacognitive Judgments Are Less Accurate Given Prospectively than Retrospectively. *Frontiers in Psychology*, *7*(February). http://doi.org/10.3389/fpsyg.2016.00218

Wickens, T. D. (2002). *Elementary signal detection theory*. New York: Oxford University Press.

Wierzchoń, M., Asanowicz, D., Paulewicz, B., & Cleeremans, A. (2012). Subjective measures of consciousness in artificial grammar learning task. *Consciousness and Cognition*, *21*(3), 1141–53. http://doi.org/10.1016/j.concog.2012.05.012

Wierzchoń, M., Paulewicz, B., Asanowicz, D., Timmermans, B., & Cleeremans, A. (2014). Different subjective awareness measures demonstrate the influence of visual identification on perceptual awareness ratings. *Consciousness and Cognition*, *27*, 109–120. http://doi.org/10.1016/j.concog.2014.04.009

Williams, J., Ramaswamy, D., & Oulhaj, A. (2006). 10 Hz flicker improves recognition memory in older people. *BMC Neuroscience*, *7*, 21. http://doi.org/10.1186/1471-2202-7-21

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: confidence and error monitoring. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *367*(1594), 1310–21. http://doi.org/10.1098/rstb.2011.0416

Table 1. Parameters of the hierarchical and the independent model of metacognition

Name	Symbol	Conceptual interpretation	Part of which models?
Sensitivity	d	Objective discrimination ability of the observer between the two stimulus alternatives	hierarchical and independent
Primary task criterion	θ	Bias of the observer towards one of the two stimulus alternatives	hierarchical and independent
Rating criteria spread	τ	Degree of conservativeness of subjective reports	hierarchical and independent
Internal noise	σ	Amount of distortion during metacognitive read-out of sensory evidence	hierarchical
Rating sensitivity	d_m	Ability of the second, metacognitive channel to discern between the two stimulus alternatives	independent

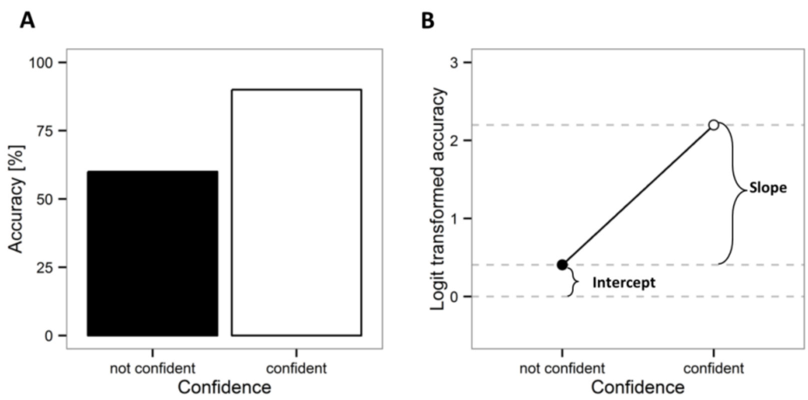

Fig. 1. Quantifying the relationship between trial accuracy and subjective reports by logistic regression. (A): Data of a hypothetical experiment. Task accuracy in % correct is plotted separately for two categories of subjective reports, “not confident” and “confident”. (B): Same data, but accuracy transformed into logits. Logistic regression is based on fitting a linear function on such transformed data. The slope of the regression line is interpreted as metacognitive sensitivity, the intercept as criterion.

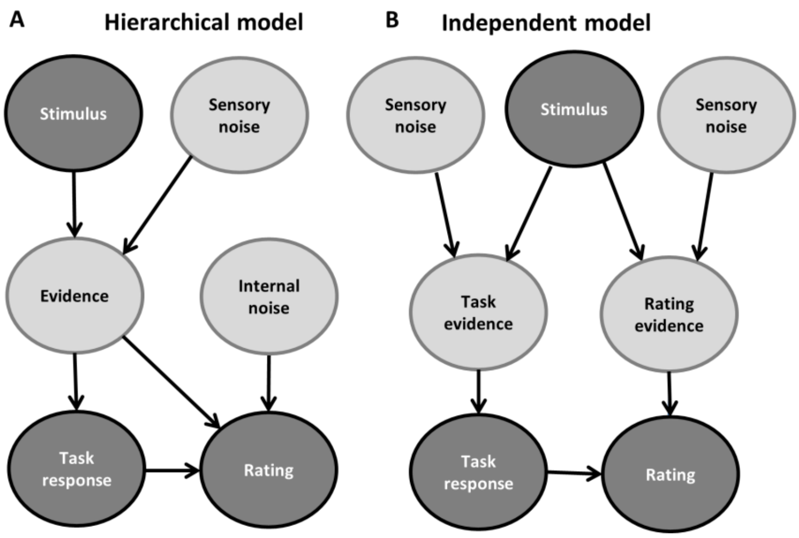

Fig. 2. The hierarchical and the independent model of metacognition. According to the hierarchical model, the rating is generated by the same evidence as the response to the primary task, but the evidence is distorted by internal noise. The task response determines which criteria are applied to select a subjective report. According to the independent model, evidence is created in parallel by two channels. The task response is selected based on the evidence selected in one of the channels. The rating response depends on whether the evidence sampled independently in the second channel confirms the response.

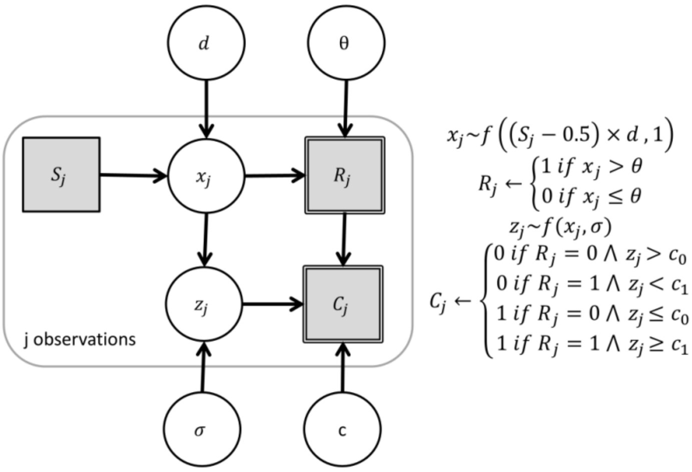

Fig. 3. Graphical model of the hierarchical model of metacognition. In each trial, the observer is faced with sensory evidence x, which depends on the stimulus S as well as the observers' sensitivity to discriminate between the two stimuli d. The response R is selected based on a comparison between x and the task bias θ. In addition, the decision variable z for selecting one out of several rating options depends on x as well as on unsystematic noise σ. The rating C depends on the decision variable z, the response R, and the rating criterion c.

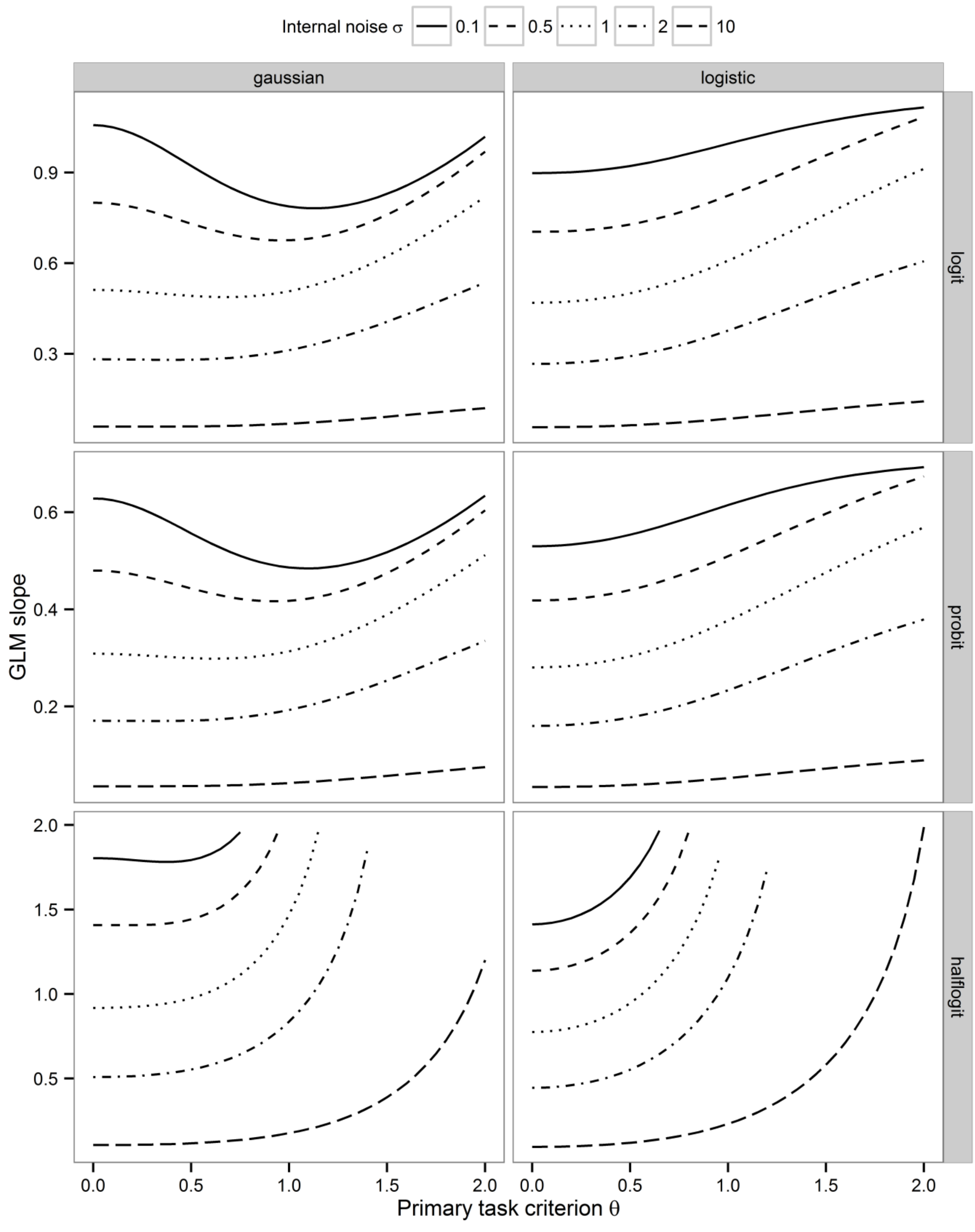

Fig. 4. Generalized linear regression slopes according to the hierarchical model of metacognition as a function of primary task criterion θ (X-Axis), internal noise σ (different lines), distributions of evidence (different columns) and link function (different rows). The other parameters were fixed to d = 1 and τ = 1.

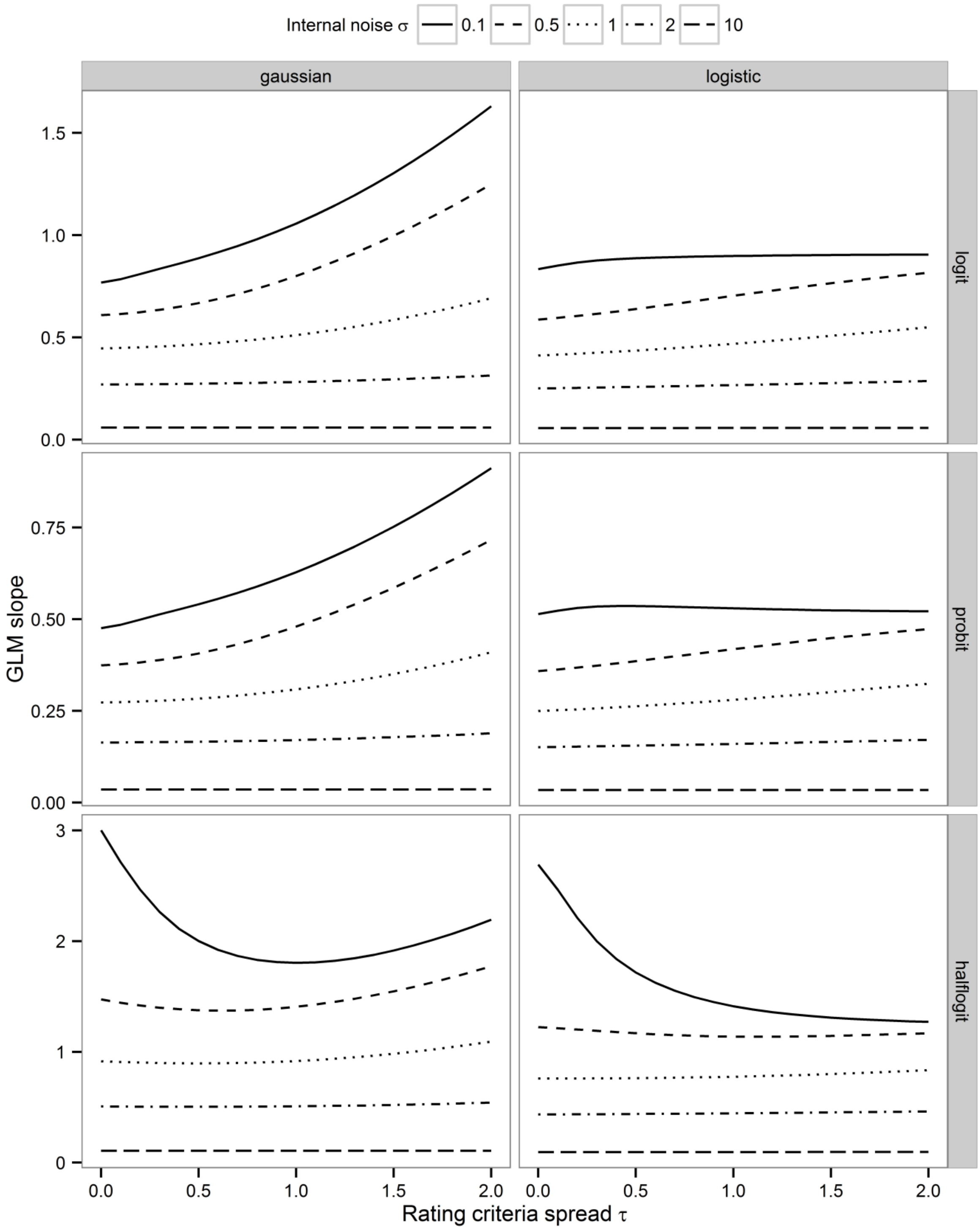

Fig. 5. Generalized linear regression slopes according to the hierarchical model of metacognition as a function of rating criteria spread τ (X-Axis), internal noise σ (different lines), distributions of evidence (different columns) and link function (different rows). The other parameters were fixed: $d = 1$, $\theta = 0$.

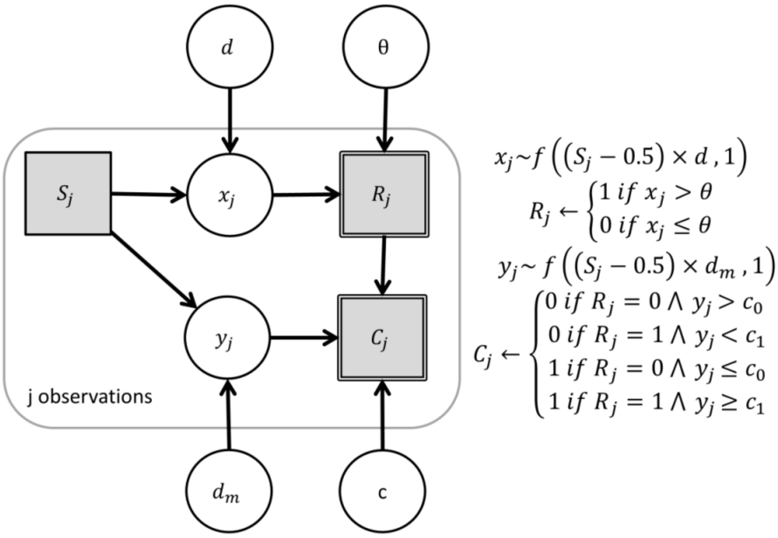

Fig. 6. Graphical model representing the dual evidence model of metacognition. Each trial, the observer is faced with sensory evidence x, which depends on the stimulus S as well as the observers' sensitivity to discriminate between the two stimuli d. The response R is selected based on a comparison between x and the task bias θ. In contrast to the single evidence model, the decision variable y for selecting one out of the several rating options depends on the stimulus S as well as on the metacognitive access parameter d_m. The rating C depends on the decision variable y, the response R, and the rating criterion c.

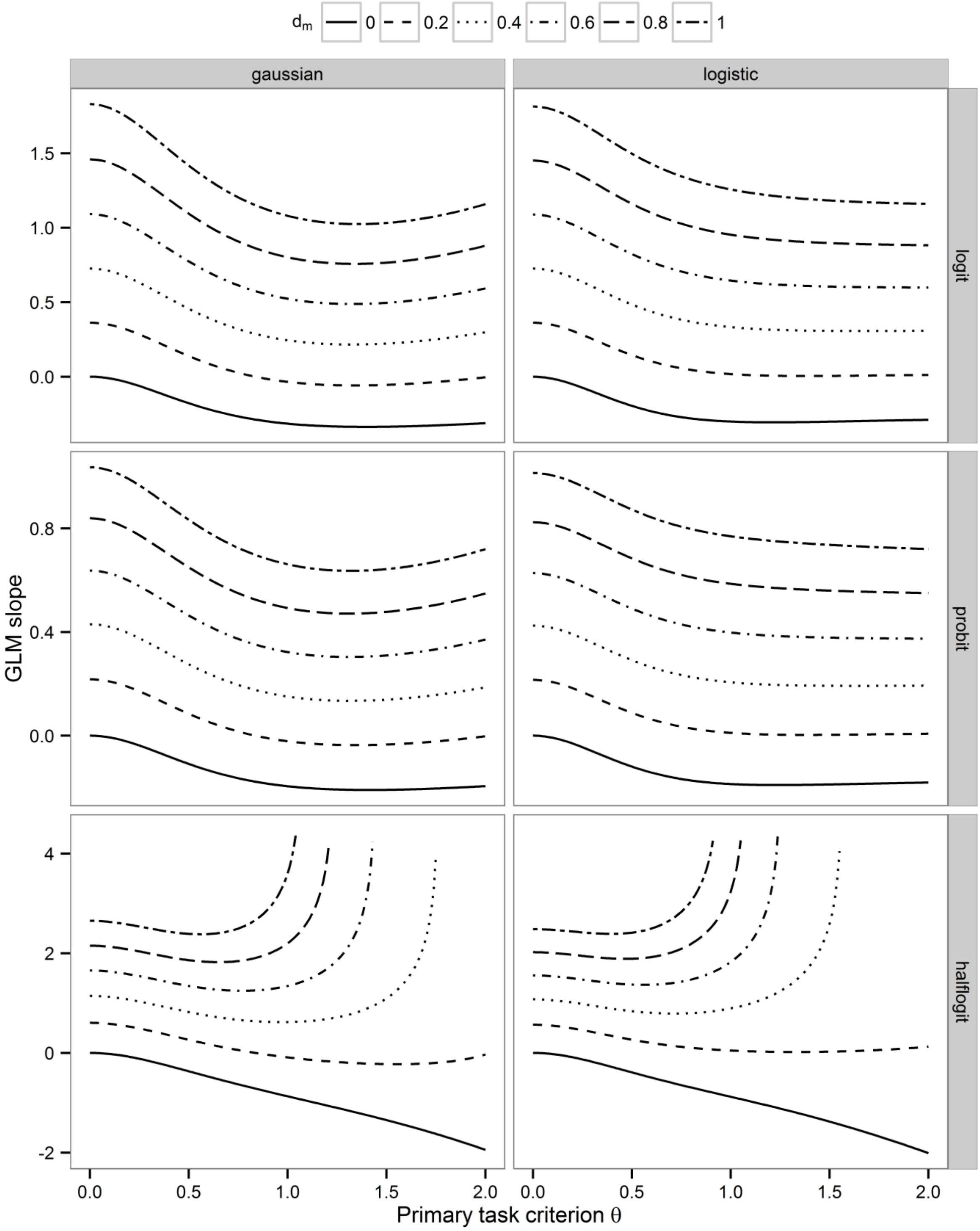

Fig. 7. Generalized linear regression slopes according to the independent model of metacognition as a function of primary task criterion θ (X-Axis), rating sensitivity d_m (different lines), distributions of evidence (different columns) and link function (different rows). The other parameters were fixed: d = 1, τ = 1.

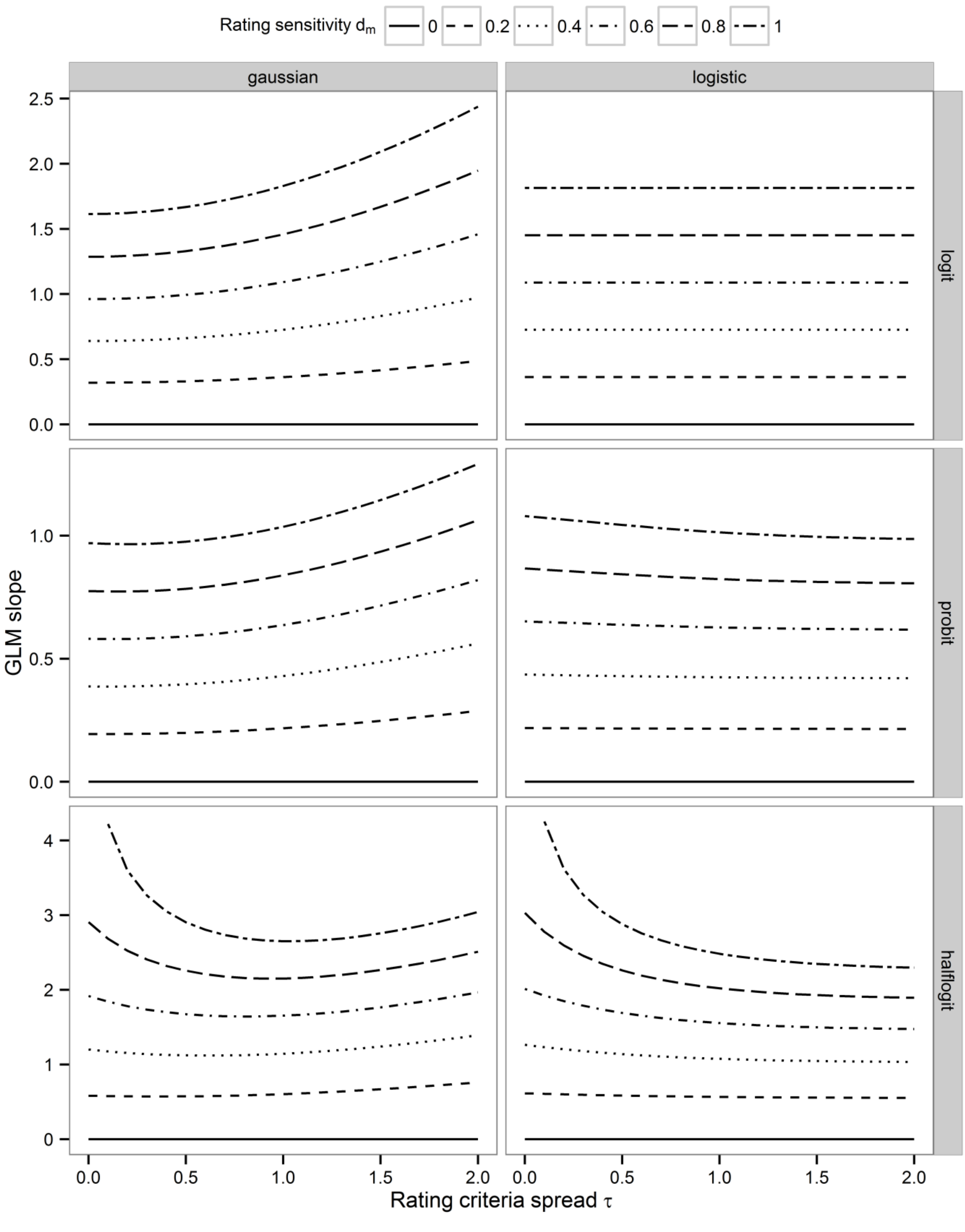

Fig. 8. Generalized linear regression according to the independent model of metacognition as a function of rating criteria spread τ (X-Axis), rating sensitivity d_m (different lines), distributions of evidence (different columns) and link function (different rows). The other parameters were fixed: d = 1, θ = 0.

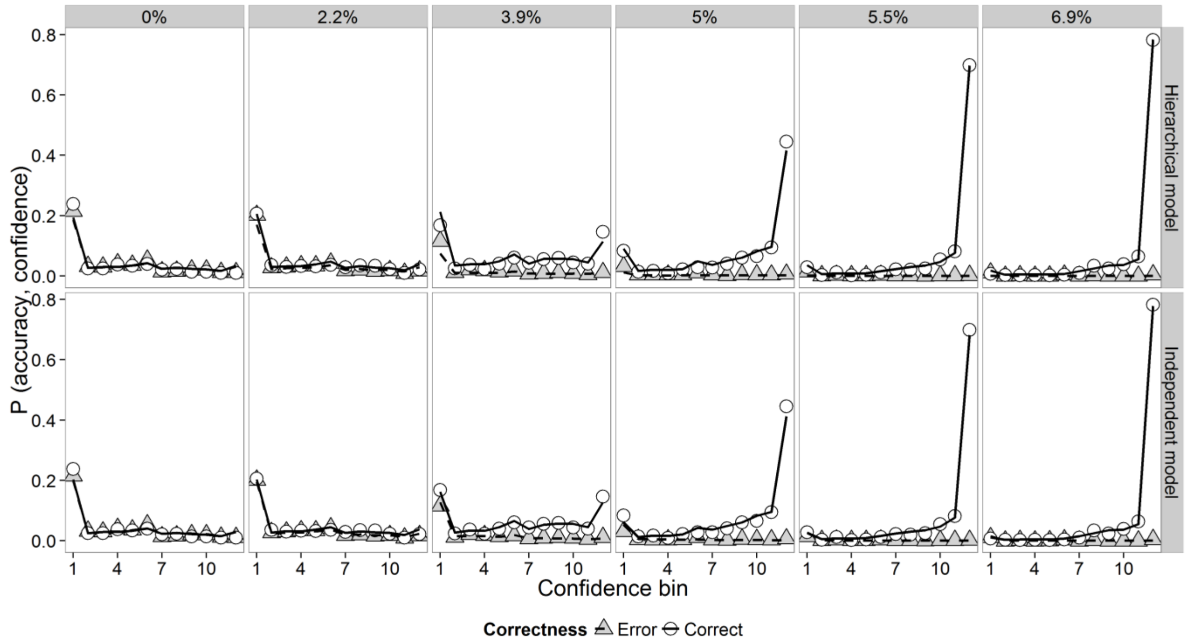

Fig. 9. Joint probability of primary task accuracy (symbols) and confidence bins (on the X-Axis) as a function of contrast levels (separate columns). Different rows indicate the prediction of the two best performing models, the hierarchical logistic model without task bias (upper row) and the independent logistic model with task bias (lower row).

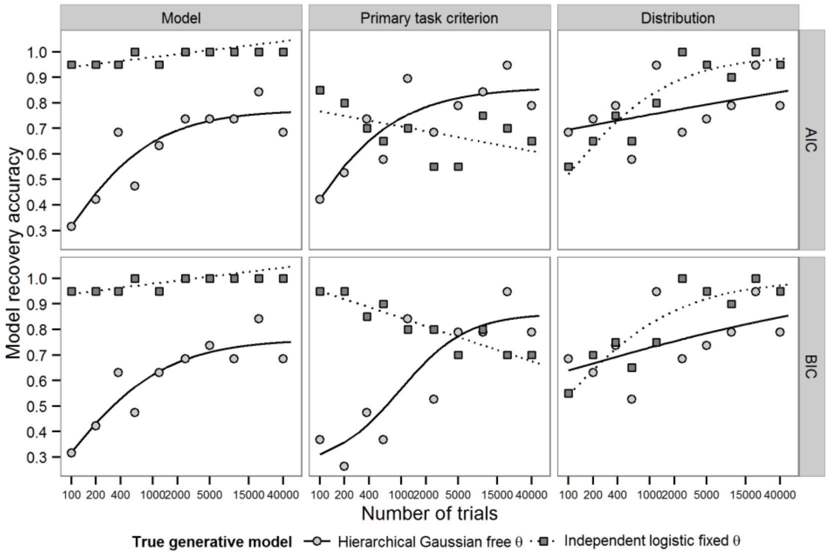

Fig. 10. Accuracy of model recovery as a function of model feature (model: hierarchical vs. independent, bias: θ free vs. fixed, distribution: logistic vs. Gaussian; in different columns), the true generative model (different symbols), number of simulated trials (X-Axis) as well as the goodness-of-fit measure (AIC_c vs BIC).

9 Appendix A

According to SDT theory, when participants are instructed to indicate which out of two possible alternative stimuli was presented, their perceptual systems provides sensory evidence, with is randomly sampled out of a distribution depending on the stimulus in the current trial. This distribution f and the corresponding cumulative density function F can be the normal distribution or the logistic distribution, which we both parametrize here by mean and standard deviation. When $S = 1$, then

$$x \sim f_{\frac{d}{2},1} \quad \text{(A1)}$$

and correspondingly, when $S = 0$, then

$$x \sim f_{-\frac{d}{2},1} \quad \text{(A2)}$$

In the hierarchical model, the decision variable z evaluated when selecting a subjective report depends on x overlaid by additive noise. Consequently, z will be distributed with a mean of x and a standard deviation of σ:

$$z \sim f_{x,\sigma} \quad \text{(A3)}$$

Participants are assumed to report confidence about a response $R = 1$ when $z > c_1$, and about a response $R = 0$ when $z < c_0$. The probability of giving a correct response and being correct given $S = 1$ can be computed as:

$$P(T = 1 \wedge C = 1 | S = 1) = P(x > \theta | S) \times P(\mathrm{z} > c_1 | x > \theta, S = 1) \quad \text{(A4)}$$

$$= \int_{\theta}^{\infty} P(x) \times P(\mathrm{z} > c_1 | x)\, dx \quad \text{(A5)}$$

$$= \int_{\theta}^{\infty} f_{\frac{d}{2},1}(x) \times \left(1 - F_{x,\sigma}(c_1)\right) dx \quad \text{(A6)}$$

$P(T = 0 \wedge C = 0 | S = 0)$, $P(T = 0 \wedge C = 0 | S = 1)$, $P(T = 0 \wedge C = 1 | S = 0)$, $P(T = 0 \wedge C = 1 | S = 1)$, $P(T = 1 \wedge C = 0 | S = 0)$, $P(T = 1 \wedge C = 0 | S = 1)$, and $P(T = 1 \wedge C = 1 | S = 0)$ are computed analogously:

$$P(T = 0 \wedge C = 0 | S = 0) = \int_{\theta}^{\infty} f_{-\frac{d}{2},1}(x) \times F_{x,\sigma}(c_1)\, dx \quad \text{(A7)}$$

$$P(T = 0 \wedge C = 0 | S = 1) = \int_{-\infty}^{\theta} f_{\frac{d}{2},1}(x) \times \left(1 - F_{x,\sigma}(c_0)\right) dx \quad \text{(A8)}$$

$$P(T = 0 \wedge C = 1 | S = 0) = \int_{\theta}^{\infty} f_{-\frac{d}{2},1}(x) \times \left(1 - F_{x,\sigma}(c_1)\right) dx \quad \text{(A9)}$$

$$P(T = 0 \wedge C = 1 | S = 1) = \int_{-\infty}^{\theta} f_{\frac{d}{2},1}(x) \times F_{x,\sigma}(c_0) dx \quad \text{(A10)}$$

$$P(T=1 \wedge C=0|S=0) = \int_{-\infty}^{\theta} f_{-\frac{d}{2},1}(x) \times \left(1 - F_{x,\sigma}(c_0)\right) dx \tag{A11}$$

$$P(T=1 \wedge C=0|S=1) = \int_{\theta}^{\infty} f_{\frac{d}{2},1}(x) \times F_{x,\sigma}(c_1)\ dx \tag{A12}$$

$$P(T=1 \wedge C=1|S=0) = \int_{-\infty}^{\theta} f_{-\frac{d}{2},1}(x) \times F_{x,\sigma}(c_0)\ dx \tag{A13}$$

For simplicity, we assume that $c_1 - \theta = \theta - c_0 = \tau$, leaving only one parameter of rating criteria τ, which represents the conservativeness of rating criteria.

Eq. (A6) – (A12) can be used to compute the probability of being correct given a subjective report of $C = 0$ and $C = 1$, respectively.

$$P(T=1|C=0) = \frac{P(T=1 \wedge C=0)}{P(C=0)} \tag{A14}$$

$$P(T=1 \wedge C=0) = \frac{1}{2} \times \big(P\,(T=1 \wedge C=0 \mid S=0) + P\,(T=1 \wedge C=0 \mid S=1)\big) \tag{A15}$$

$$P(C=0) = \frac{1}{2} \times \big(P\,(T=1 \wedge C=0 \mid S=0) + P\,(T=1 \wedge C=0 \mid S=1) + P\,(T=0 \wedge C=0 \mid S=0) + P\,(T=0 \wedge C=0 \mid S=1)\big) \tag{A16}$$

$$P(T=1|C=1) = \frac{P(T=1 \wedge C=1)}{P(C=1)} \tag{A17}$$

$$P(T=1 \wedge C=1) = \frac{1}{2} \times \big(P\,(T=1 \wedge C=1 \mid S=0) + P\,(T=1 \wedge C=1 \mid S=1)\big) \tag{A18}$$

$$P(C=1) = \frac{1}{2} \times \big(P\,(T=1 \wedge C=1 \mid S=0) + P\,(T=1 \wedge C=1 \mid S=1) + P\,(T=0 \wedge C=1 \mid S=0) + P\,(T=0 \wedge C=1 \mid S=1)\big) \tag{A19}$$

10 Appendix B

According to the independent model, participants select the response R based on a comparison between the sample of sensory evidence x and the primary task criterion θ, just as in the standard SDT model. However, subjective reports are assumed to be based on a second independent sample of sensory evidence y. The distribution f and the corresponding cumulative distribution function F are assumed to be the same as those from which x is sampled, except for the mean, which is described by the metacognitive access parameter d_m. When $S = 1$, then

$$y \sim f_{\frac{1}{2}d_m,1} \quad \text{(B1)}$$

and correspondingly, when $S = 0$, then

$$y \sim f_{-\frac{1}{2}d_m,1} \quad \text{(B2)}$$

Participants are assumed to report confidence about a response $R = 1$ when $y > c_1$, and about a response $R = 0$ when $y < c_0$. The probability of giving a correct response and being correct given $S = 1$ in the dual evidence model is obtained by:

$$P(T = 1 \wedge C = 1 | S = 1) = P(x > \theta | S) \times P(\mathrm{y} > c_1 | S = 1) \quad \text{(B4)}$$

$$= \left(1 - F_{\frac{d}{2},1}(\theta)\right) \times \left(1 - F_{\frac{1}{2}d_m,1}(c_1)\right) \quad \text{(B5)}$$

$P(T = 0 \wedge C = 0|S = 0)$, $P(T = 0 \wedge C = 0|S = 1)$, $P(T = 0 \wedge C = 1|S = 0)$, $P(T = 0 \wedge C = 1|S = 1)$, $P(T = 1 \wedge C = 0|S = 0)$, $P(T = 1 \wedge C = 0|S = 1)$, and $P(T = 1 \wedge C = 1|S = 0)$ are computed analogously:

$$P(T = 0 \wedge C = 0 | S = 0) = \left(1 - F_{-\frac{d}{2},1}(\theta)\right) \times F_{-\frac{1}{2}d_m,1}(c_1) \quad \text{(B6)}$$

$$P(T = 0 \wedge C = 0 | S = 1) = F_{\frac{d}{2},1}(\theta) \times \left(1 - F_{\frac{1}{2}d_m,1}(c_0)\right) \quad \text{(B7)}$$

$$P(T = 0 \wedge C = 1 | S = 0) = \left(1 - F_{-\frac{d}{2},1}(\theta)\right) \times \left(1 - F_{-\frac{1}{2}d_m,1}(c_1)\right) \quad \text{(B8)}$$

$$P(T = 0 \wedge C = 1 | S = 1) = F_{\frac{d}{2},1}(\theta) \times F_{\frac{1}{2}d_m,1}(c_0) \quad \text{(B9)}$$

$$P(T = 1 \wedge C = 0 | S = 0) = F_{-\frac{1}{2}d_m,1}(\theta) \times \left(1 - F_{-\frac{1}{2}d_m,1}(c_0)\right) \quad \text{(B10)}$$

$$P(T = 1 \wedge C = 0 | S = 1) = \left(1 - F_{\frac{1}{2}d,1}(\theta)\right) \times F_{\frac{1}{2}d_m,1}(c_1) \quad \text{(B11)}$$

$$P(T = 1 \wedge C = 1 | S = 0) = F_{-\frac{1}{2}d_m,1}(\theta) \times F_{-\frac{1}{2}d_m,1}(c_0) \quad \text{(B12)}$$

Supplementary Material

to

Should metacognition be measured by logistic regression?

Manuel Rausch [1,2] and Michael Zehetleitner [1,2]

[1] Katholische Universität Eichstätt-Ingolstadt, Eichstätt, Germany
[2] Ludwig-Maximilians-Universität München, Munich, Germany

Correspondence should be addressed at:

Manuel Rausch

Katholische Univerität Eichstätt-Ingolstadt

Psychologie II

Ostenstraße 25, "Waisenhaus"

85072 Eichstätt

Germany

Phone: +49 8421 93 21639

Email: manuel.rausch@ku.de

http://www.ku.de/ppf/psychologie/psych2/mitarbeiter/m-rausch/

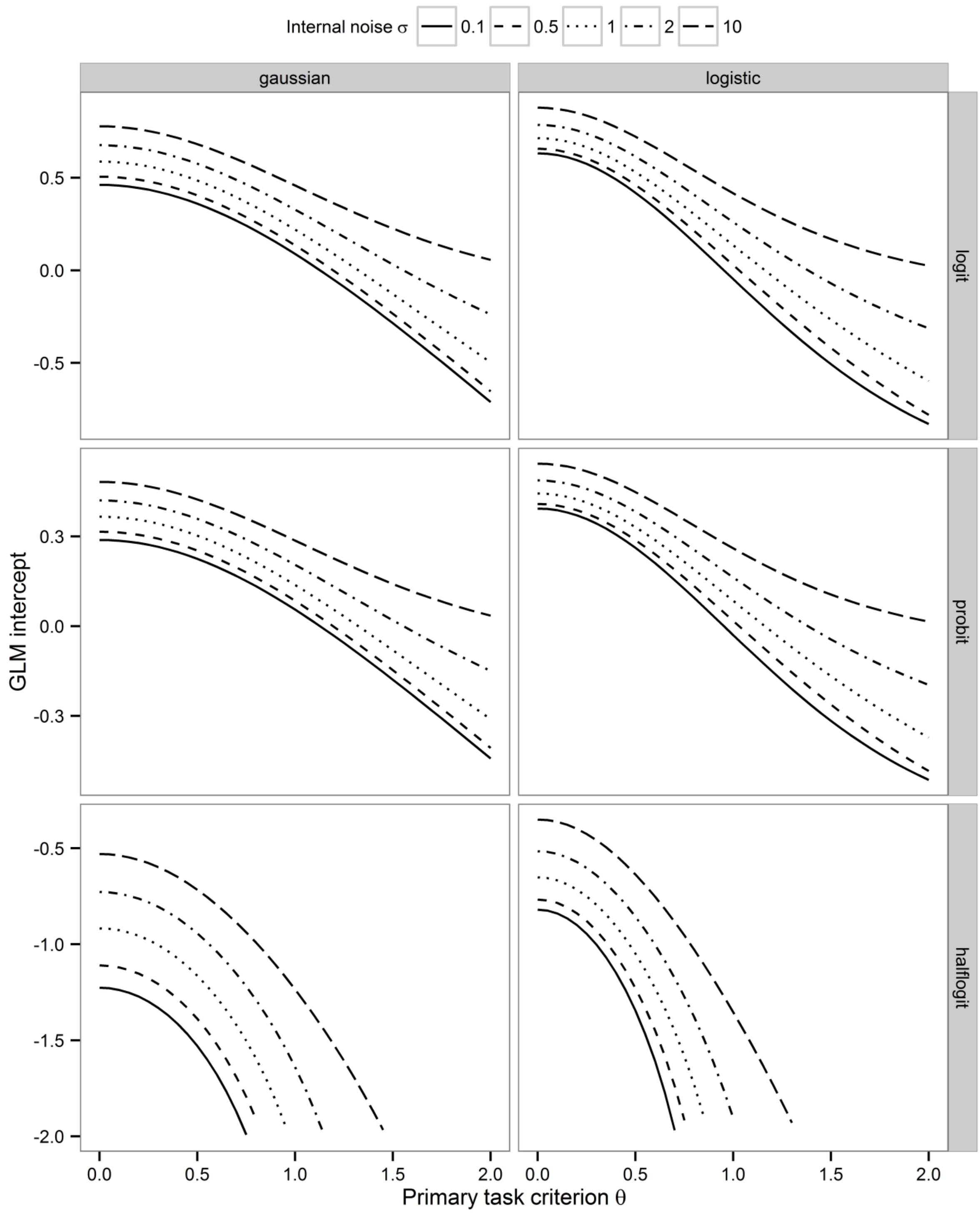

Figure S1. Generalized linear regression intercepts according to the hierarchical model of metacognition as a function of primary task criterion θ (X-Axis), internal noise σ (different lines), distributions of evidence (different columns) and link function (different rows). The other parameters were fixed: d = 1, τ = 1.

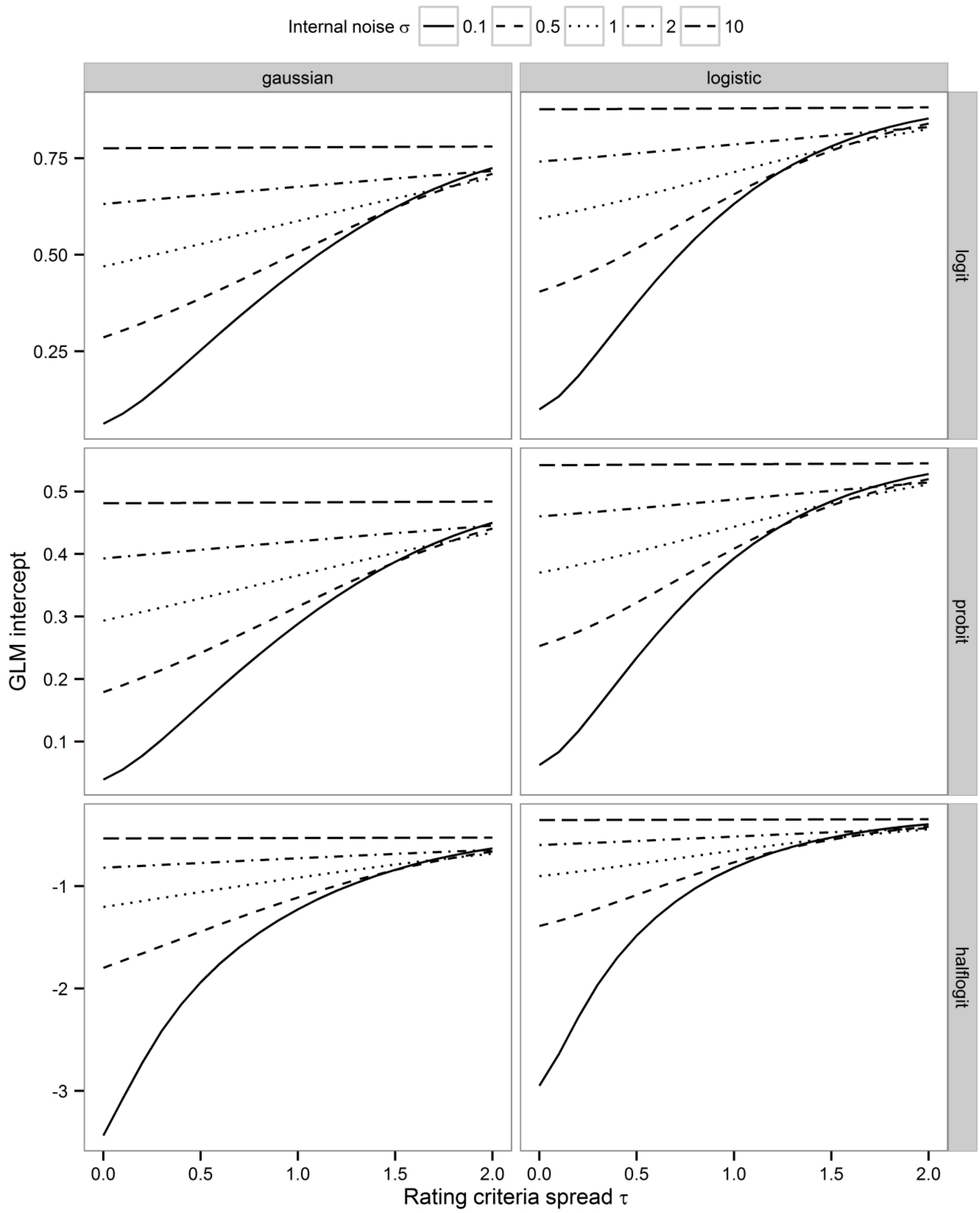

Figure S2. Generalized linear regression intercepts according to the hierarchical model of metacognition as a function of rating criteria spread τ (X-Axis), internal noise σ (different lines), distributions of evidence (different columns) and link function (different rows). The other parameters were fixed: d = 1, θ = 0.

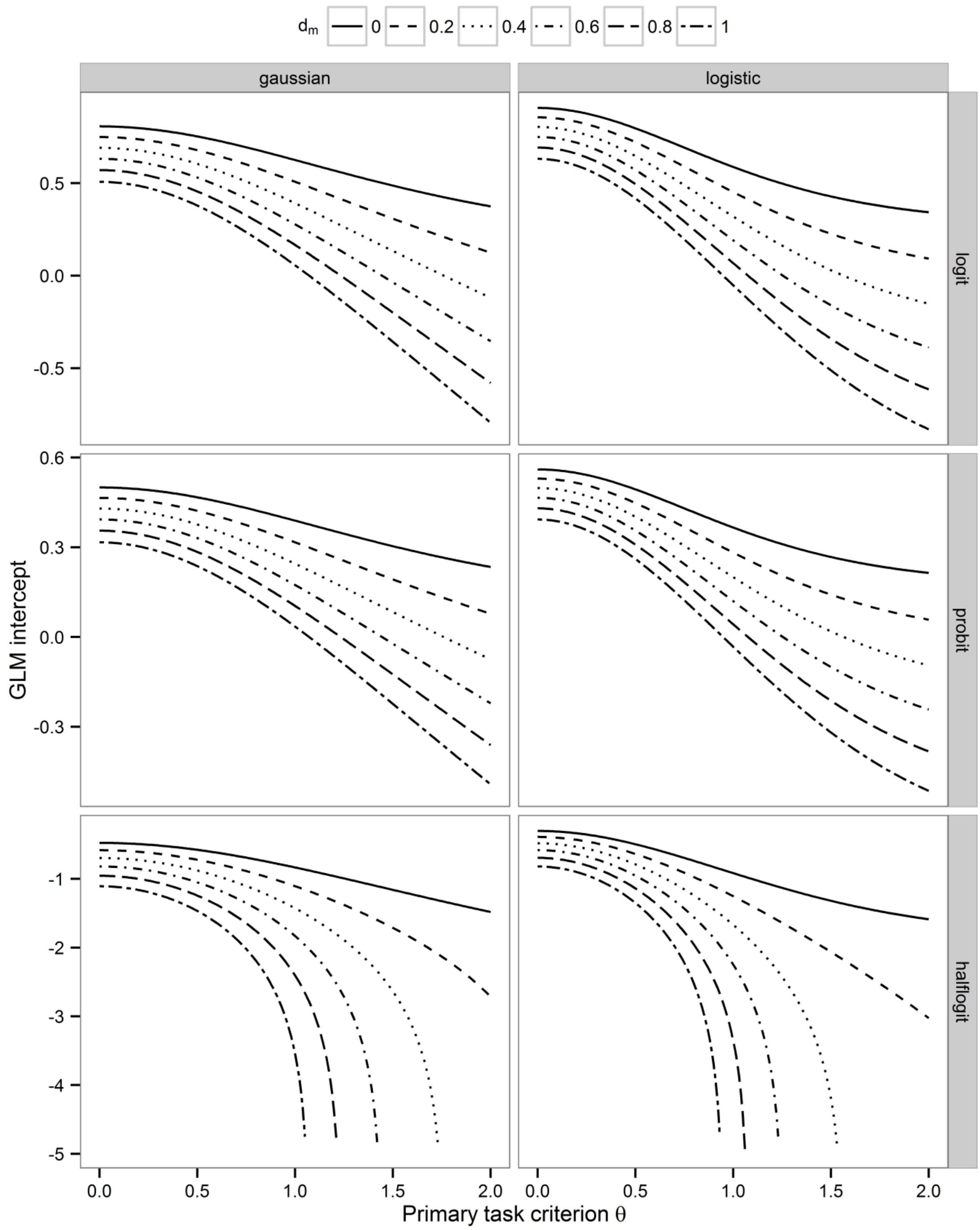

Figure S3. Generalized linear regression intercept according to the independent model of metacognition as a function of primary task criterion θ (X-Axis), rating sensitivity d_m (different lines), distributions of evidence (different columns) and link function (different rows). The other parameters were fixed: d = 1, τ = 1.

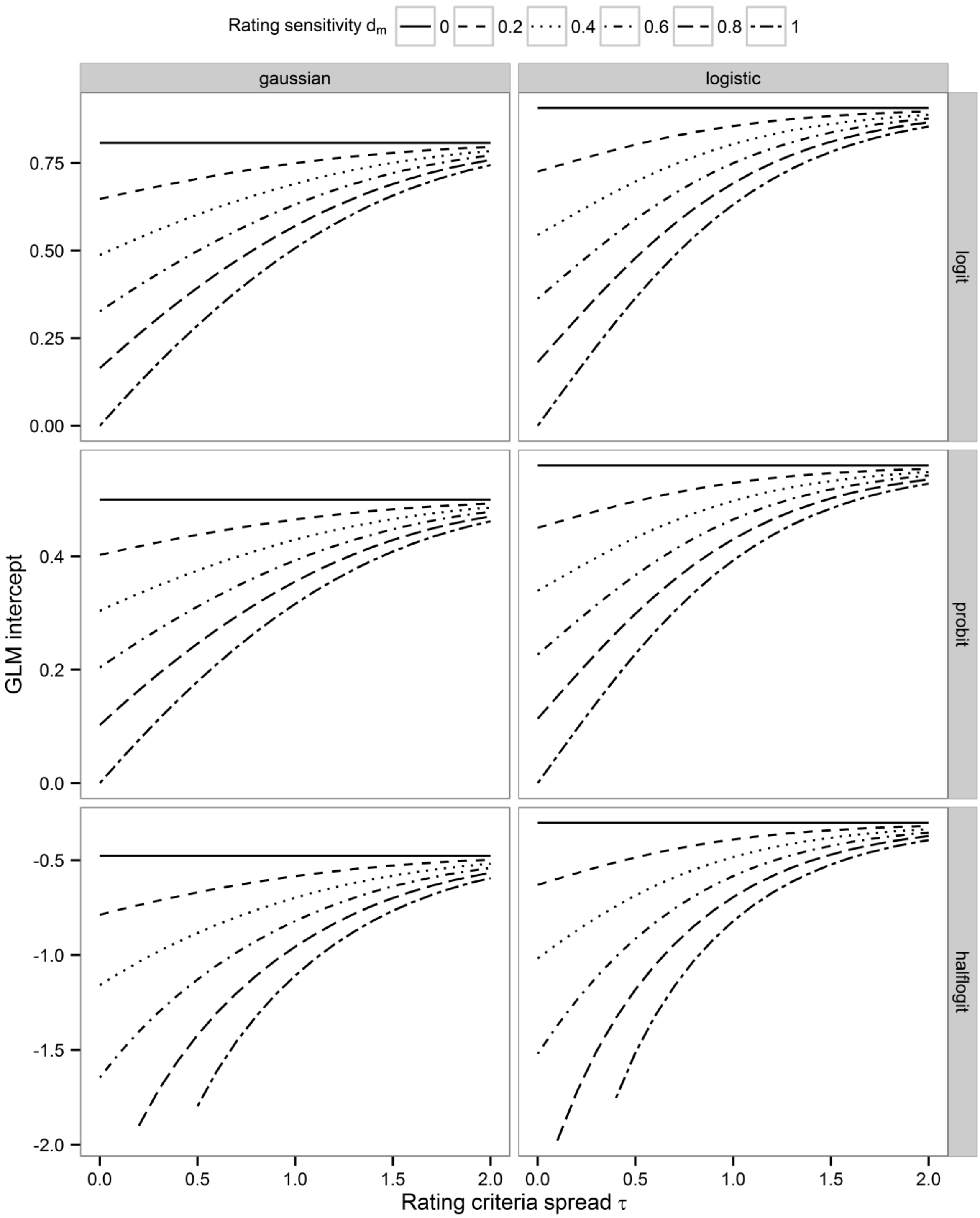

Figure S4. Generalized linear regression intercept according to the independent model of metacognition as a function of rating criteria spread τ (X-Axis),, rating sensitivity d_m (different lines), distributions of evidence (different columns) and link function (different rows). The other parameters were fixed: $d = 1$, $\theta = 0$