Reclassifying guesses to increase signal-to-noise ratio in psychological experiments

Frédéric Gosselin[1], Jean-Maxime Larouche[1], Valérie Daigneault[1], and Laurent Caplette[1,2]
[1]Département de psychologie, Université de Montréal
[2]Department of psychology, Yale University

Abstract

This paper introduces a novel procedure that can increase signal-to-noise ratio in psychological experiments that use accuracy as a *selection variable* for another dependent variable. This procedure relies on the fact that some correct responses result from guesses and reclassifies them as incorrect responses using a trial-by-trial reclassification evidence such as response time. It selects the optimal reclassification evidence criterion beyond which correct responses should be reclassified as incorrect responses. We show that the more difficult the task and the fewer the response alternatives, the more to be gained from this reclassification procedure. We illustrate the procedure on behavioral and ERP data from two different datasets (Caplette et al., 2020; Faghel-Soubeyrand et al., 2019) using response time as reclassification evidence. The reclassification procedure increased signal-to-noise ratio by 13-20%. Matlab and Python implementations of the reclassification procedure are freely available (https://github.com/GroupeLaboGosselin/Reclassification).

Since the seminal work of Gustav Fechner and Hermann von Helmholtz in the first half of the 19th century, researchers in psychology and related fields have tried to understand the mind by studying behavior. Typically, subjects are stimulated with an image, a sound or another physical stimulus, and respond to a specific question about this stimulus to the best of their ability on each trial. Often differences in stimuli or in brain states between correct and incorrect responses are then analyzed in various ways. In the widely used "subsequent memory paradigm", for example, the brain activity elicited by to-be-encoded stimuli is contrasted for correctly and incorrectly retrieved stimuli on a subsequent memory test (for a review, see Wagner, Koutstaal and Schacter, 1999). Likewise, in studies that employ the Bubbles technique, participants are asked to complete a task on partially revealed stimuli on every trial, and the information that leads to incorrect responses is subtracted from the information that leads to correct responses (Chauvin et al., 2005). Here, we present a novel procedure that can increase signal-to-noise ratio (SNR) in such psychological experiments that use accuracy as a *selection variable* for another dependent variable. This procedure reclassifies some correct responses as incorrect responses using a trial-by-trial reclassification evidence such as response time. When there are two response alternatives and the correct response rate is 75%, the SNR can be increased by as much as 41%. We show that the more difficult the task and the fewer the response alternatives, the more to be gained from this reclassification procedure. We apply the procedure on behavioral and EEG datasets (Caplette et al., 2020; Faghel-Soubeyrand & Gosselin, 2019) with response time as reclassification evidence. In both cases, the reclassification procedure increased signal-to-noise ratio by more than 13%.

False correct responses, guesses and other useful definitions

We will suppose that on any trial a participant accumulates clues consciously or not in favor of the different response alternatives. Eventually, this participant has to select a response among *A* response alternatives. If an insufficient amount of neural evidence has been gathered by the participant to choose one response with a sufficiently high degree of certainty, the participant *guesses*. We will assume that, with two response alternatives — the focus of this article — guessing leads to correct and incorrect responses with the same probability of 0.5. For example, randomly selecting one of the response alternatives satisfies this assumption. The probability of guessing as a function of neural evidence does not have to be a step function. In the simulation reported later and described in the Appendix, for example, it's a logistic function. We will also assume that only *guesses* can lead to incorrect responses. This particular observer model was introduced by Gustav Fechner and is called the *threshold model* (Green & Swets, 1966; Hautus, Macmillan & Creelman, 2021). It is a two-process model: a main recognition process and a guessing process. We will discuss the impact of other causes of errors made by participants on the reclassification procedure introduced in this article in subsequent sections.

In the remainder of this article, we will use the terms *correct* and *incorrect* responses to refer to responses that a researcher put in the "correct" and "incorrect" categories, respectively. We will refer to a complete collection of correct and incorrect responses as a *response classification*. We will call the special response classification provided by the actual responses of a participant during an experiment, the *original*

response classification. We will divide correct responses in *true* and *false* correct responses: true correct responses are responses that an omniscient being would put in the "correct" category because they resulted from the success of the observer's recognition process, whereas false correct responses are responses that an omniscient being would put in the "incorrect" category because they resulted from the observer's guessing process.

From response classification efficiency to response reclassification efficiency

The proportion of all responses that are true correct and incorrect responses might seem like a reasonable index of the efficiency of the original response classification. However, this index does not reflect adequately the information available when contrasting correct and incorrect responses, which, as we wrote in the introductory section, is our goal here. In this case, a false correct response should penalize classification efficiency twice. Think about it this way: If you could identify one false correct response and set it aside, your comparison of stimuli or brain states associated with correct vs incorrect responses would be slightly improved because it would prevent the nulling of the signal of one incorrect response by that of this false correct response. In fact, this improvement should be the same as adding exactly one response to the experiment, assuming that all trials contain equivalent information and are interchangeable. Now, if you put this false correct response in the "incorrect" category, you would gain one incorrect response, and your comparison would be slightly improved, once more. It would be equivalent to adding another trial to the experiment. This suggests, as a measure of the original response classification's efficiency, the difference between the proportion of all responses that are true correct or incorrect responses, and the proportion of all responses that are false correct responses. This measure varies between 1, when all responses are true correct responses or true incorrect responses, and -1, when all responses are false correct responses or false incorrect. In practice, however, it should never go below 0, which is the expected response classification efficiency of a random response classification (e.g., with two response alternatives, half of responses would be correct, by chance). For example, if the proportion of original correct responses is 0.75 in a task with two response alternatives, the original response classification efficiency is equal to 0.5. Indeed, we know that the proportion of incorrect responses is 0.25 (1 - 0.75); we know that this is also the proportion of false correct responses, since both types of responses result from guesses and guesses are assumed to result in both response alternatives with equal probability; furthermore, if 0.25 correct responses are false correct responses, that leaves 0.50 true correct responses (0.75 - 0.25). Thus, 0.25 incorrect responses plus 0.50 true correct responses minus 0.25 false correct responses results in an original response classification efficiency of 0.5. Note that the sum of the original correct and incorrect responses (I_O) is equal to the total number of responses (N). Thus, the efficiency of the original classification (E_O) in tasks with two response alternatives, as in the example we just gave, can be simplified to $E_O = 1 - \frac{2I_O}{N}$ (e.g., when the original correct response rate is 0.75: E_O = 1 – 2 * 0.25 = 0.5).

The remainder of this article will pertain to *response reclassification* — altering the original response classification with the goal of increasing efficiency as much as possible.

Response reclassification is limited to the original correct responses because, as we mentioned above, we assume that the original incorrect responses are all true incorrect responses. Thus, there is nothing to be gained from reclassifying original incorrect responses. Each reclassified response belongs to one of four categories. We will call the true correct responses reclassified as correct responses, the reclassification's *correct rejections* and the true correct responses reclassified as incorrect responses, the reclassification's *false alarms*; similarly, we will call the false correct responses reclassified as correct responses the reclassification's *misses* and the false correct responses reclassified as incorrect trials, the reclassification's *hits*. It is straightforward to adapt the above measure of response classification efficiency for this reclassification context:

$$E = (I_O + H + C - M - F)/N, \qquad \text{(Equation 1)}$$

with *E*, the efficiency of the response reclassification; *H*, the number of reclassification's hits; *C*, the number of reclassification's correct rejections; *M*, the number of reclassification's misses; and *F*, reclassification's false alarms. The goal of the reclassification procedure can be framed as maximizing Equation 1 given *some reclassification evidence* available on a trial-by-trial basis. The best possible outcome is a reclassification efficiency of 1. By taking the reciprocal of E_O, we get the maximum response reclassification efficiency gain. With two response alternatives and with a proportion of original correct responses equal to 0.75, as in the above example, $E_O^{-1} = 2$ (i.e., 1/0.5). This is equivalent to doubling the number of responses, or, put another way, it represents a SNR increase of about 41% (i.e., $\sqrt{2}$), assuming that all trials contain the same information. For tasks with two response alternatives, E_O is inversely proportional to I_O and, thus, E_O^{-1} is directly proportional to I_O. In other words, the more difficult the task, the smaller the original response classification efficiency and, more importantly, the more to be gained from the reclassification procedure. We generalize this result to tasks with more than two response alternatives in a subsequent section.

The reclassification procedure

What makes a reclassification evidence useful for reclassification is that it tends to be different for true correct than for false correct (or incorrect) responses. Importantly, our implementation of the reclassification procedure adapts to the polarity of the relationship between the true or false correct responses, and reclassification evidence. Thus, if guesses associated with smaller reclassification evidence values (e.g. shorter response times) dominated false correct incorrect responses, the correct trials associated with the smallest reclassification evidence values would be reclassified as incorrect trials. Instead, we will illustrate the reclassification procedure with the data from a hypothetical experiment with false correct responses associated with larger reclassification evidence (e.g. longer response times) than true correct responses, and with an original correct response rate of 75% with two response alternatives (see appendix for simulation details). The dashed gray curve on Figure 1 represents the resulting hypothetical frequency distribution of original correct reclassification evidence and the solid curves its decomposition in true (black line) and false (gray line) correct frequency distributions.

In this case, the reclassification procedure provides a reclassification evidence criterion above or below which original correct responses are reclassified as incorrect

responses that maximizes Equation 1 (see reclassification evidence criterion on Figure 1). To compute *H*, *C*, *M* and *F*, the unknowns of Equation 1, we need the reclassification evidence frequency distributions of true and false correct responses. Given our assumption that all false correct responses result from guesses, false correct reclassification evidence and the original incorrect reclassification evidence must be drawn from the same population and are equiprobable with two response alternatives. In other words, in this particular case, the false correct reclassification evidence frequency distribution is identical to the incorrect reclassification evidence frequency distribution (Figure 1, solid gray line). With more than two response alternatives, the incorrect and the false correct response reclassification evidence frequency distributions are not the same but are proportional to one another (see section *Multiple response alternatives*). Furthermore, as we stated from the get-go, the original correct reclassification evidence frequency distribution is the sum of the true and false correct reclassification evidence frequency distributions. Therefore, the true correct reclassification evidence frequency distribution (Figure 1, solid black line) is equal to the difference between the correct reclassification evidence frequency distribution (Figure 1, dashed black line) and the false correct reclassification evidence frequency distribution. This portion of the reclassification procedure might be improved, in some instances, by fitting a specific density function to the correct and false correct reclassification evidence frequency distributions prior to subtracting the latter from the former. However, the reclassification procedure was meant to be used with any reclassification evidence for which we do not necessarily have knowledge about the appropriate distribution to be used. For this reason, we did not include a curve-fitting stage to our implementation of the reclassification procedure.

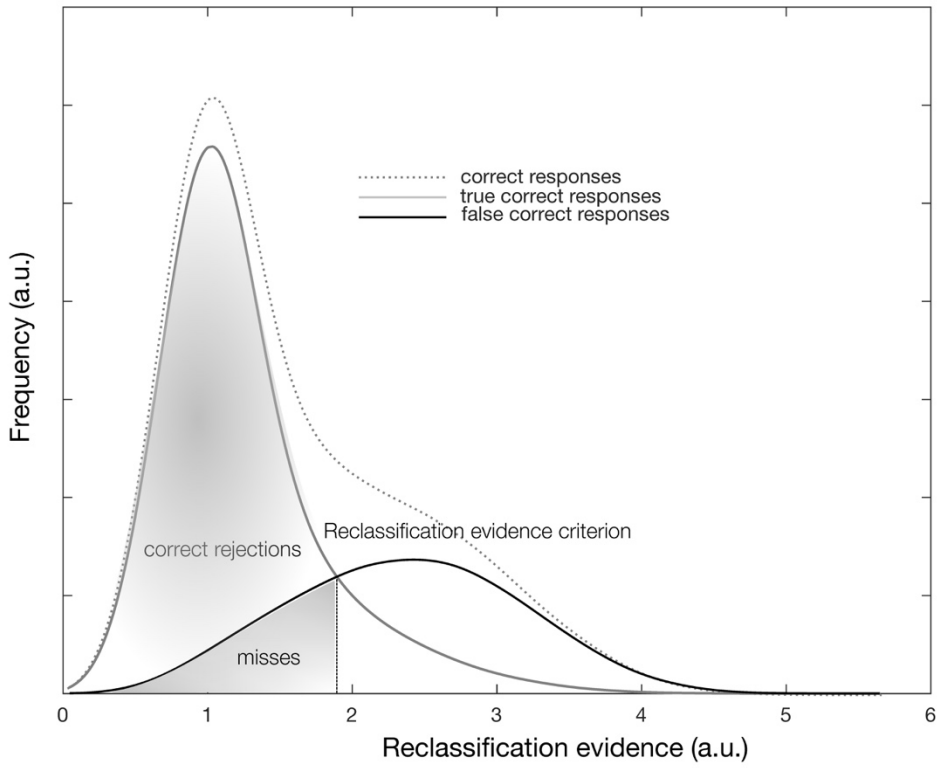


Figure 1. Reclassification of simulated data. Response reclassification is restricted to the original correct responses. This figure shows a hypothetical reclassification evidence frequency distribution of original correct responses, as well as its decomposition in reclassification evidence frequency distributions of true correct and false correct responses. Here, with two response alternatives, the frequency reclassification evidence frequency distribution of false correct responses is identical to the frequency reclassification evidence frequency distribution of original incorrect responses. The correct rejection and miss areas are shown with respect to the optimal reclassification evidence criterion. Note that the true correct responses above criterion are false alarms, and that the false correct responses above criterion are hits.

With these true and false correct reclassification evidence frequency distributions under our belt, we can calculate *H*, *C*, *M* and *F* as a function of reclassification evidence criterion. In fact, with two response alternatives, $H + M = I_O$ (the number of false correct responses — $H + M$ — is equal to the number of original incorrect responses) and $C + F = N - 2\ I_O$ (the number of true correct responses — $C + F$ — is equal to the total number of responses minus twice the number of original incorrect responses — the first time because the number of original correct responses is equal to the number of responses minus the number original incorrect responses, and the second time because the number of true correct responses is equal to the number of correct responses minus the false correct responses, which, as we already wrote, is equal to the number of original incorrect responses). Thus, Equation 1 can be simplified as follows, without *F* and *H*:

$$E = \big(4I_O - N + 2(C - M)\big)/N. \qquad \text{(Equation 2)}$$

The number of true correct responses below a given reclassification evidence criterion, like the one shown on Figure 1, is equal to *C* (the number of reclassification's correct rejections); similarly, the number of false correct responses below the same given reclassification evidence criterion is equal to *M* (the number of reclassification's misses). The curve in Figure 2 represents reclassification efficiency as a function of reclassification evidence criterion for the true and false correct reclassification evidence frequency distributions shown on Figure 1. The curve quickly rises from 0 (when all original correct responses are reclassified as incorrect responses), peaks at an efficiency of 0.736 with an reclassification evidence criterion of 1.885, and then, slowly goes down to E_O, the efficiency of the original response classification, which, in this case, is 0.5 (when no original correct responses are reclassified). The best reclassification efficiency is thus equivalent to an increase of 47% of responses (0.736/0.5) or a SNR increase of about 21% (√1.470), assuming that all false correct responses carry the same information. The reclassification efficiency curve displayed on Figure 2 was computed from the simulated original correct responses, incorrect responses, and reclassification evidences, which is all the information that we have following an actual experiment. However, since this is a simulation, we have access to the ground truth, i.e., the omniscient classification, and thus to the true efficiency. This true efficiency curve overlaps almost perfectly the one calculated from Equation 2. Note that Equation 2 peaks where (*C* - *M*) peaks. This maximum is attained where the derivative of (*C* - *M*) is equal to 0 that is, precisely where the true and false correct frequency distributions meet. This will become important for our analysis of the effect of lapses on the reclassification procedure below.

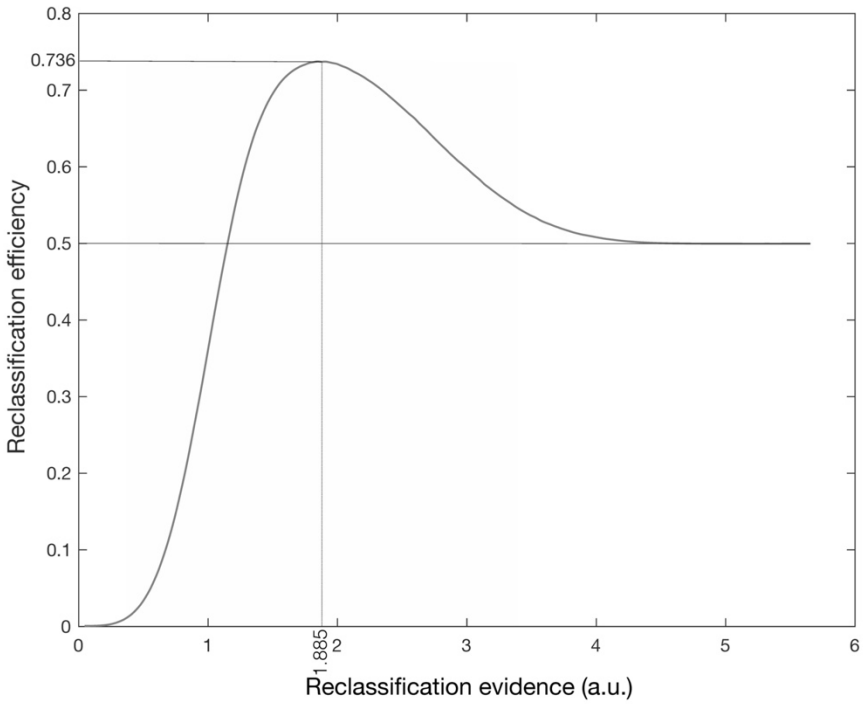


Figure 2. Reclassification efficiency as a function of reclassification evidence criterion based on the hypothetical true correct and false correct reclassification evidence frequency distributions shown in Figure 1. The maximum efficiency of about 0.736 is attained with a reclassification evidence criterion

of about 1.885. The height of the tail of the reclassification efficiency curve, 0.5, corresponds to the efficiency of the original classification.

The hypothetical situation illustrated in Figures 1 and 2 was designed to illustrate the reclassification procedure with a relatively clear-cut case. Nonetheless, it represents an intermediate situation in which the true and false correct reclassification evidence frequency distributions are only partially overlapping and, thus, the application of the reclassification procedure leads to the reclassification of *some* of the original correct responses — richer in false correct responses than in true correct responses — as incorrect responses (i.e., a reclassification efficiency between the original classification efficiency and 1). If the reclassification evidence was uncorrelated with neural evidence, unlike in the above simulation, the true and false correct reclassification evidence frequency distributions would coincide and no original correct response would be reclassified as incorrect — the reclassification would be null — and thus, the reclassification efficiency would be equal to the efficiency of the original response classification. At the other extreme, a reclassification evidence perfectly correlated with neural evidence, results in true and false correct reclassification evidence frequency distributions with no or little overlap, and a high reclassification efficiency. A residual overlap between the frequency distributions, here, would originate from the probabilistic relationship between neural evidence and original response classification (it is the case in the above simulation; see Appendix).

Comparison with other procedures

In sum, our reclassification procedure boils down to subtracting the reclassification evidence frequency distribution of false correct responses (which is equal to the frequency distributions of incorrect responses with two response alternatives assuming the threshold observer model) from the frequency distribution of original correct responses to isolate the frequency distribution of true correct responses. The reclassification criterion, beyond which the original correct responses are reclassified as incorrect response, is where these true and false correct frequency distributions meet. For comparison purposes, it is useful to divide this procedure in three stages: 1) splitting the correct reclassification evidence frequency distribution in true correct and false correct reclassification evidence distributions; 2) deriving a reclassification evidence criterion from these true and false correct reclassification evidence frequency distributions; and 3) reclassifying the correct responses beyond this criterion as incorrect responses.

The first stage of our procedure — removing false correct responses from the correct response reclassification evidence frequency distribution, assuming that all errors are due to guesses — is the least original stage. It is similar to the kill-the-twin procedure (Eriksen, 1988; Miller & Lopes, 1991; Gondan & Heckel, 2008) which obtains the true correct responses by eliminating, for each original incorrect response, one original correct response with a similar RT (presumably a false correct response). In this literature, the true correct RTs are then used to test race models reducing the bias caused by fast guesses in correct RTs. In other words, the kill-the-twin procedure doesn't go beyond the first stage of our reclassification procedure. More recently, Glickman, Gray and

Morales (2005) developed a latent-variable approach to combine speed and accuracy information that bears many similarities to the first stage of our reclassification procedure. They assume two underlying processes: an error-free process, which is equivalent to the process responsible for the true correct responses, and a guessing process, which includes both rapid guesses and slow guesses and is responsible for incorrect responses. The guessing process in our reclassification procedure is also assumed to be responsible for all incorrect responses but it is unimodal unlike theirs. Another difference between the two procedures is that they fit Weibull density functions to the true correct and the true incorrect RT frequency distributions. We didn't implement any curve fitting in the companion Matlab and Python functions, as we have already explained, because we wanted the reclassification procedure to be applicable to any reclassification evidence, including reclassification evidences with unknown density functions. The ultimate goal of Glickman, Gray and Morales is to characterize the two competing latent stochastic processes. They propose, for example, to include these parameters in statistical models to better compare performance across experimental conditions. Thus, once more, these authors did not go beyond the first stage of our reclassification procedure.

As far as we know, our reclassification procedure is entirely novel with respect to the last two stages. Some authors have proposed to reject correct responses beyond a certain *response criterion*. However, our particular criterion, computed from the true and false correct reclassification evidence frequency distributions — any reclassification evidence, not just RT—, is novel to the best of our knowledge. For example, proposals related to our reclassification procedure have been made in the context of psychometric testing with the goal of identifying and, ultimately, rejecting items associated with fast guessing (Wise, 2019). Fast guessing tend to occur because of disengagement due, for example, to low-stake situations (for a recent survey, see Kang & Ratcliff, 2021). These procedures aim at finding a RT criterion under which it is likely that a student was guessing. Superficially, this is very close to the criterion-finding stage in our reclassification procedure using RT as the reclassification evidence. However, as we will see, the characteristics of these criteria all differ from those of our criterion. Schnipke (1995) proposed to take the local minimum between the two modes of the overall response time frequency distribution. This criterion, however, can only be calculated when the response time frequency distribution is bimodal and if there are enough trials to unambiguously identify the minimum. Interestingly, Schnipke's criterion is equivalent to our reclassification evidence criterion when there is no overlap between the two underlying distributions. Lee and Jia (2014; see also Guo et al., 2016) proposed a criterion that exploits the idea that fast guesses are expected to result in chance performance. They proposed to use the first RT associated with accuracies that diverge in a statistically significant manner from chance performance. This is difficult to apply in several real-life situations because a lot of repeated trials are required. For this reason, Wise and Ma (2012) considered various percentages of an item's average response time as thresholds and used the point at which response accuracy in a large database of test items began to increase beyond chance level as a criterion for setting a threshold. They recommended the 10% point for their so-called *normative threshold method*. These psychometric procedures are quite similar to artifact exclusion procedures often used to "clean" the

data from psychological experiments (for a recent review, see Berger & Kiefer, 2021). These artifact exclusion procedures can be applied to any type of continuous measure, and they typically produce two thresholds — one under which trials are rejected and one above which trials are rejected. Several relative criteria have been proposed over the years. The two most popular ones are probably the mean of the measure plus or minus two standard deviations of the measure, and Tukey's fences — smaller than the 25th percentile of the measure minus 1.5 times the interquartile range of the measure, and greater than the 75th percentile of the measure plus 1.5 times the interquartile range of the measure (Tukey, 1977).

Finally, we believe that we are the first ones to ever propose the third and final stage of our reclassification procedure, that is, reclassifying some correct responses as incorrect responses to increase the SNR in psychological experiments that use accuracy as a selection variable for another dependent variable.

Empirical tests of the reclassification procedure

We tested the reclassification procedure with the Faghel-Soubeyrand et al. (2019) and Caplette et al. (2020) datasets using, in both cases, RT as reclassification evidence. RT is likely to be the most used reclassification evidence since it is almost always recorded concomitantly with behavioral responses. We computed the SNR gain of the reclassified responses relative to the original responses following the application, for each participant, of the reclassification procedure. We then compared this to the SNR gain achieved with several other methods. First, we calculated the mean percentage of reclassified trials across participants, and then, for each participant, reclassified that percentage of the slowest correct responses as incorrect. Second, we computed the SNR gain associated with other procedures discussed in the previous section: Tukey's, Wise and Ma's and the mean RT + 2 RT standard deviation. We used only greater-than criteria, just like in the reclassification procedure (but bear in mind that our implementation of the reclassification procedure adapts to the polarity of the relationship between the true or false correct responses, and reclassification evidence). Finally, we tried both rejecting and reclassifying the responses above the criteria. In all cases, reclassification outperformed rejection. Therefore, we only report the reclassification SNR gains below.

Faghel-Soubeyrand et al. (2019). These researchers examined the use of facial features in 140 individuals during a sex discrimination Bubbles task (available at https://osf.io/jfz68). Three hundred color face images (150 human females) from Dupuis-Roy et al. (2009) were used to generate the stimuli. These face images were scaled, rotated, and translated so that the position of the eyes, the nose, and the mouth coincided as much as possible while preserving relative distances between them. Inter-pupil distances were 40 pixels on average. Face images were randomly flipped on the vertical axis on every trial to control for possible information asymmetries (e.g., illumination differences). Stimuli were created by superimposing an opaque grey mask punctured by randomly located Gaussian windows of 3 pixels of standard deviation, or *bubbles*, on randomly selected face images (see Figure 3a). The number of bubbles per bubble mask was adjusted on a trial-by-trial basis using the QUEST algorithm (Watson & Pelli, 1983) to maintain performance at a rate of 75% of correct responses throughout

the entire experiment. Stimuli subtended 3.08 x 3.08 degrees of visual angle (128 x 128 pixels). Participants had to identify the sex of the face by pressing one of two keyboard keys.

To reveal the use of information of a particular participant, a classification image (CI) was computed. It consists, essentially, of averaging the bubble masks associated with correct responses, on the one hand, and those associated with incorrect responses, on the other hand, and in subtracting element-by-element the latter from the former (Chauvin et al., 2005). This is the ideal experiment to test the reclassification procedure. First, accuracy is a selection variable for the bubble masks in the CI computation. The success of the response reclassification can be assessed as SNR gain in these CI. As a measure of SNR, we used the standard deviation of the pixels of the group CIs — the sums of all 140 individual CIs. The standard definition of SNR is the ratio between the mean of the signal and the standard deviation of the noise. Given that the goal of classification image experiment is usually to discover what the nature of the signal is, it doesn't make sense, here, to use this definition of SNR. However, we know that we should observe a greater standard deviation over the entire classification image than that of noise, which, by design, is equal to 1 in our z-scored classification images. As a measure of SNR, we used the standard deviation of the pixels of the group CIs — the sums of all 140 individual CIs. This SNR measure is more conservative than the standard measure because noise regions are included in the computations. As a measure of SNR gain, we divided the standard deviation of the group CI obtained from the accuracies transformed, for example, with our reclassification procedure, by the standard deviation of the group CI obtained from the original accuracies. Second, each participant of Faghel-Soubeyrand et al. (2019) completed 300 trials, which is relatively few trials for a Bubbles experiment. This dataset thus provides an opportunity to test the SNR gain following response reclassification near the limit of the Bubbles method's sensitivity.

The mean RT per trial in the Faghel-Soubeyrand et al. (2019) experiment was 1.61 s (SD = 0.91 s). The mean original correct and incorrect RT were 1.50 s (SD = 0.83) and 1.93 s (SD = 1.18; $t(139) = 11.445$; $p<.001$), respectively. This confirms that RT is a promising response reclassification evidence. A reclassification RT criterion was computed for each subject. The average RT criterion selected by the reclassification procedure was 2.78 s (SD = 2.55 s). The minimum RT criterion was 0.49 s and the maximum 24.99 s. The maximum proportion of reclassified responses was 0.33 and the minimum, 0 (for 5.71% of all participants the reclassification was null, that is, no correct response was reclassified as an incorrect response). An average of 9.97% of all responses were reclassified (SD = 7.67%).

Reclassified accuracy led to a SNR gain of 113.46%. Accuracies resulting from reclassifying the average proportion of individual responses reclassified by our reclassification procedure — the slowest 9.97% correct responses — led to a smaller but nonetheless important SNR gain of 110.22%. Both these SNR gains are greater than the SNR gains associated with Tukey's fences (107.55%), Wise and Ma's criterion (109.20%) and the mean + 2 st. dev. criterion (106.04%) (see Figure 3b).

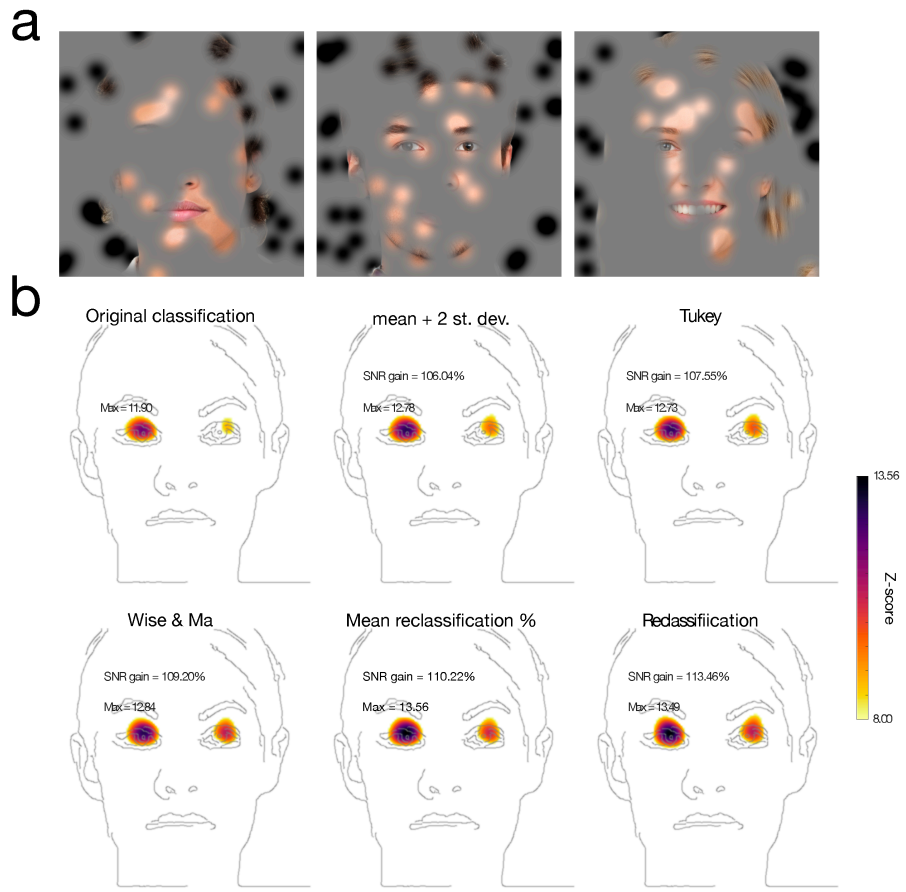


Figure 3. Test of the reclassification method on the Faghel-Soubeyrand et al. (2019) dataset. a) Stimuli similar to those of Faghel-Soubeyrand et al. (2019). The three faces partially revealed were created by a generative adversarial network. In the actual experiment, real face photographs were used. b) High values in the group classification images (CI) computed from different weights for the Faghel-Soubeyrand et al. (2019) data set as well as various statistics. The inferno color map that we used is perceptually uniform. The face outline is shown to help with interpretation.

Caplette et al. (2020). We also tested the reclassification procedure on part of a second dataset from Caplette et al. (2020) available at https://osf.io/3r782. In the experiment that was analyzed here, participants performed a sex discrimination task on face stimuli while their electroencephalic (EEG) activity was recorded. Face stimuli were altered in a similar way to the study of Faghel-Soubeyrand et al. (2019) but we will not consider this aspect of the experiment here. Instead, we will assess whether EEG activity differs between correct and incorrect responses, and whether reclassifying these responses using RTs can increase our ability to detect differences (i.e., increase the SNR). Mean accuracy was 82.9%. The mean RT was 0.686 s (SD = 0.079 s). The mean original correct and incorrect RT were 0.677 s (SD = 0.072 s) and 0.720 s (SD = 0.102 s; $t(11) = 3.000$, $p=0.012$), respectively, indicating that response times could be a good reclassification evidence. A reclassification RT criterion was computed for each session. The average RT criterion selected by the reclassification procedure was 0.890 s (SD = 0.113 s). The minimum RT criterion was 0.720 s and the maximum 1.105 s. The maximum proportion of reclassified responses was 25.75% and the minimum, 0.1%. An average

of 7.66% of all responses were reclassified (SD = 11.80%). The average predicted reclassification efficiency was 55.68 % (SD = 9.07%). The mean predicted reclassification gain, that is, the ratio between the predicted reclassification efficiency and the null reclassification efficiency, was 111.36% (SD = 18.14%).

To test if reclassifying the responses indeed led to an increase in SNR, we compared the EEG activity for (original or reclassified) correct and incorrect trials for each subject and computed *t* statistics: this resulted in electrode x time maps of *t* statistics. We compared the map derived from original accuracies with the one derived from reclassified accuracies (see Figure 4). To compute the SNR of each map, we used the same standard deviation measure as above; we then divided the SNR of the map obtained with reclassified accuracy by the one of the map obtained with the original classification. Reclassified accuracy led to a SNR gain of 120.32%, which is higher than results derived using the mean + 2 st. dev. criterion (107.36%), Tukey's criterion (105.36%), Wise & Ma's criterion (100.80%) or the mean percentage of reclassified responses (111.25%).

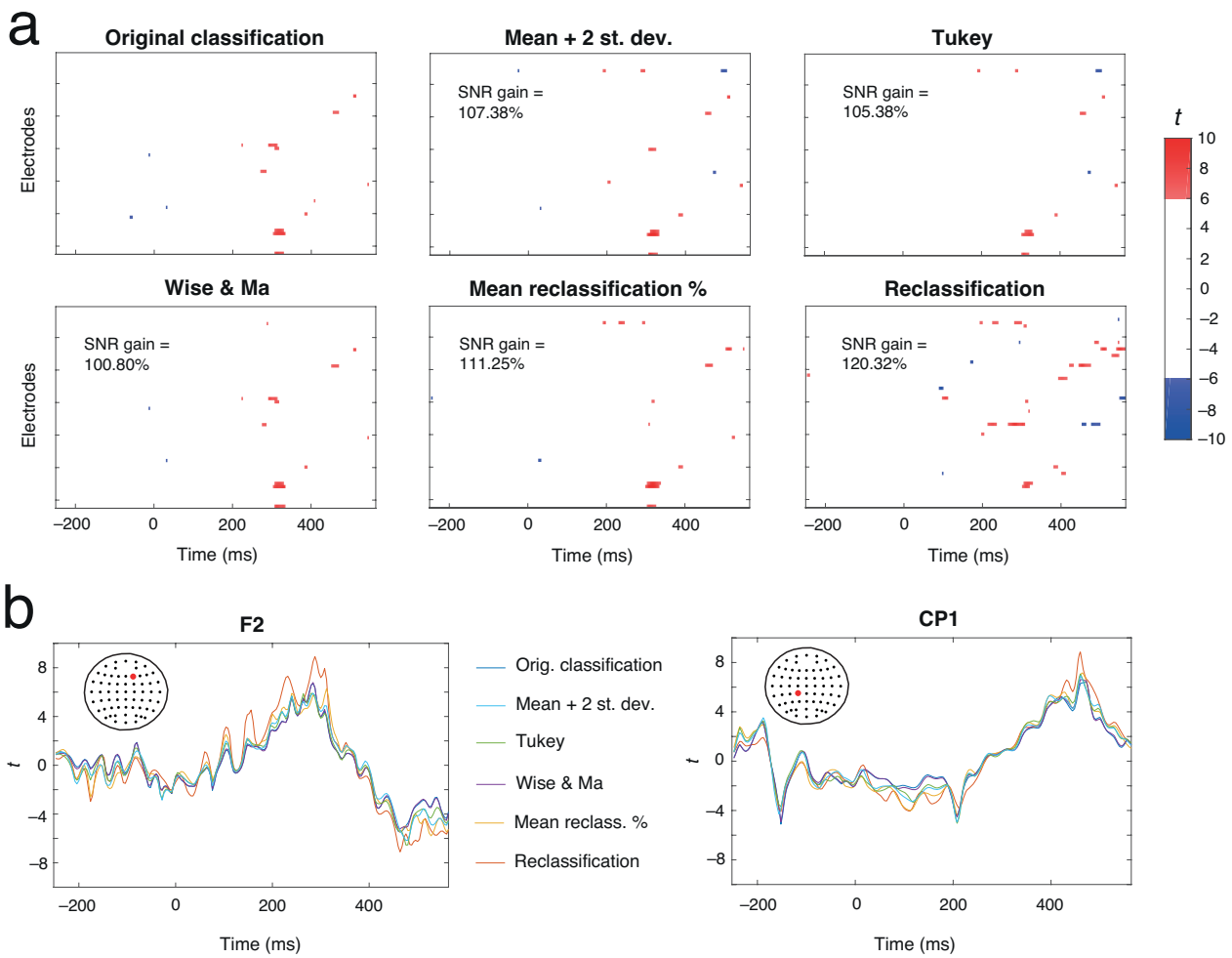


Figure 4. Test of the reclassification procedure on the Caplette et al. (2020) dataset. a) Results derived using different analysis methods. *t* statistics of high magnitude for all electrodes and time points are

shown for each method. b) *All* t statistics for two specific electrodes (left and right). Each line corresponds to a different analysis method (legend in the middle).

Others sources of errors

The reclassification procedure assumes that all errors during a psychological experiment come from blind guessing. Three other potential sources of errors are discussed in the literature: lapses, internal noise and inefficiency. We examine their impact on the reclassification procedure in the two following sections.

Lapses. When the participants' vigilance drops, when they are distracted, when they look away, when they blink or, more generally, when they respond irrespectively of what was asked of them, participants lapse. These lapses can be present in both original correct and orginal incorrect responses. However, they should be considered neither true correct, nor true incorrect responses because they are not related to the stimuli and do not inform us about the strategies used by the participants to resolve the task. In other words, they come from a third underlying cognitive process (in addition to the recognition and guessing processes). In this section, we'll examine the effect of lapses on the reclassification procedure. We'll suppose that lapses represent *L* responses and that they come from a unique reclassification evidence distribution, which may or may not be the same as the reclassification evidence distributions of true correct or true incorrect responses. With two response alternatives, $\frac{L}{2}$ lapses are expected to be hiding both in the original correct responses and in the original incorrect responses. This implies that false correct reclassification evidence frequency distribution which is estimated by the original incorrect reclassification evidence frequency distribution also contain $\frac{L}{2}$ lapses. It also implies that lapses cancel out in the true correct reclassification evidence frequency distribution which is obtained by subtracting the false correct reclassification evidence frequency distribution from the original correct reclassification evidence frequency distribution. Finally, this means that the false correct reclassification evidence frequency distribution is pushed up and, therefore, this curve meets the true correct reclassification evidence frequency distribution at a higher point than it would without lapses. In other words, in the presence of lapses, the reclassification procedure proposes a reclassification evidence criterion which is liberal or, if you prefer, it reclassifies too many correct responses as incorrect ones. The extent of this overreclassification depends on the characteristics of the lapses' reclassification evidence frequency distribution. Where on the x-axis the false correct and true correct reclassification evidence frequency distributions meet is also determined by the steepness of the downward slope of the latter. The gentler, the more the impact.

For the sake of the argument, we will suppose that the lapses reclassification evidence frequency distribution is a Gaussian distribution with an area of $\frac{L}{2}$, a standard deviation of σ and a mean of μ. The height of this distribution is thus given by $\frac{L}{2\sigma\sqrt{2\pi}}$. This is by how much the false correct reclassification evidence frequency distribution would be pushed up in the worst-case scenario — if the lapses reclassification evidence frequency distribution is centered on the optimal reclassification criterion. Note that this

height is proportional to L and inversely proportional to σ. Lapse rate — $\frac{L}{N}$ — can be estimated using easy catch trials during experiments. For example, Manning and colleagues (2014) used this procedure to measure lapse rate during an experiment examining the development of global motion processing. They observed lapse rates of 0.04 in 5 year-olds, of 0.02 in 7 year-olds, of 0.01 in 9 and 11 year-olds, and less than 0.01 in adults. In the hypothetical cases illustrated in Figure 5 , we'll use a generous lapse rate of 0.05. In all other respects, this hypothetical case is identical to the one shown in Figure 1 and described in the Appendix. In particular, the optimal reclassification criterion calculated from ground truth accuracy is equal to 1.885. Figures 5c and 5d illustrate, respectively, a narrow (small σ) and an extended (large σ) lapses reclassification evidence frequency distribution centered on the optimal reclassification criterion — the worst-case scenario mentioned above. Note that the effect of lapses is very small in Figure 5d (the prescribed reclassification criterion is 1.830 whereas the optimal one is 1.885). And for the more probable scenarios — lapses reclassification evidence frequency distributions not centered on the optimal criterion — the effect of lapses is negligible (the prescribed and optimal reclassification criteria are equal to the thousandths; see Figure 5a and 5b).

Lapses affect the results of the reclassification in one other way. In the original response classification, they occur equally often in the correct and in the incorrect trials. After reclassification, however, they are necessarily more numerous in the incorrect than in the correct trials. How much so will depend on the characteristics of the reclassification evidence frequency distribution of lapses. In Figure 5b, for example, almost all lapses originally classified as correct responses are reclassified as incorrect, whereas, in Figure 5a, few, if any, are.

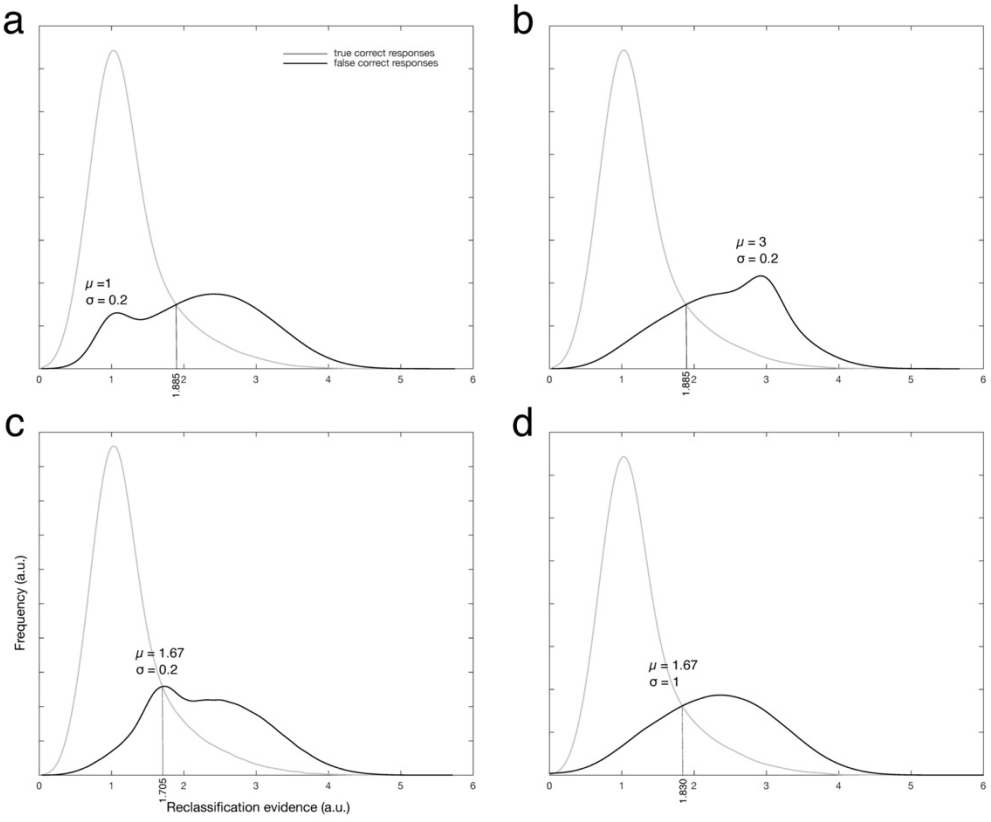


Figure 5. The effect of lapses on the reclassification method. These are the same hypothetical reclassification evidence frequency distributions than those presented in Figure 1, except for the inclusion of lapses. The optimal reclassification criterion calculated from ground truth accuracy is thus equal to 1.885. We modeled the lapses' reclassification evidence frequency distributions as Gaussian frequency distributions with an area corresponding to a liberal lapse rate of 0.05. Lapses with a narrow reclassification evidence frequency distribution (σ = 0.2 a.u. in these examples) and a) clearly below (μ = 1 a.u. in this example) or b) clearly above criterion (μ = 3 a.u. in this example) do not interfere with the meeting point between the hypothetical false correct and the true correct reclassification evidence frequency distributions — in both cases, the prescribed and optimal reclassification evidence criteria are equal to the thousandths. Lapses near reclassification evidence criterion (μ = 1.67 a.u. in these examples) lead to an underestimation of this criterion more pronounced for c) narrow (σ = 0.2 a.u. in this example) than d) broad reclassification evidence frequency distributions (σ = 1 a.u. in this example).

Inefficiency and noise. The reclassification procedure assumes that all original incorrect responses come from guesses. In addition to lapses, discussed in the previous section, two other sources of errors have been considered in the literature: noise and inefficiency. One popular observer model, the signal detection theory (SDT) model, assumes that only noise and inefficiency are responsible for incorrect responses (e.g., Hautus, Creelman & Macmillan, 2021; Green & Swets, 1966). In the SDT model observer, the effect of noise added to the stimuli (i.e., external noise) and noise added to the internal template which is compared with the noisy stimuli (i.e., internal noise) is the same. In a detection task, for example, it increases the variance of the neural activation frequency distributions for signal-present and signal-absent trials without changing their means. (You can think of neural activation, here, as a signed version of neural evidence, with the sign indicating neural evidence in favor of one vs. the other response alternatives.) This, in turn, decreases the model observer's sensitivity. This can be easily understood when considering the most common measure of sensitivity: *d'*. The *d'* index is defined as the difference between the means of the neural activation distributions for signal-present and signal-absent trials (which is invariant to an increase in noise) divided by the standard deviation of either of these distributions (which increases when noise increases). An increase of inefficiency (i.e., the use by the observer of a template less similar to the one used by the best possible observer for the task at hand – the ideal observer), the other cause for errors in the SDT model observer, also decreases sensitivity but differently than noise. Indeed, an increase of inefficiency decreases the difference between the means of the neural activation distributions for signal-present and signal-absent trials without affecting the variances of the two distributions. In sum, adding external noise to the stimuli, adding internal noise to the SDT observer model's template, or making the model's template less efficient all decrease sensitivity. (Note that noise and inefficiency do not affect the other main characteristic of the SDT model observers: the bias.)

As we have just explained, the SDT model assumes a *single process*, a recognition process: All original correct and incorrect responses result from applying this somewhat noisy and inefficient recognition process to somewhat noisy stimuli. You will remember that the goal of the reclassification procedure is to better separate the responses according to the underlying mechanisms at play (i.e., guessing and recognition) in order to use these reclassified responses as a selection variable used to contrast another dependent variable (e.g., bubble masks). It's thus nonsensical to apply the

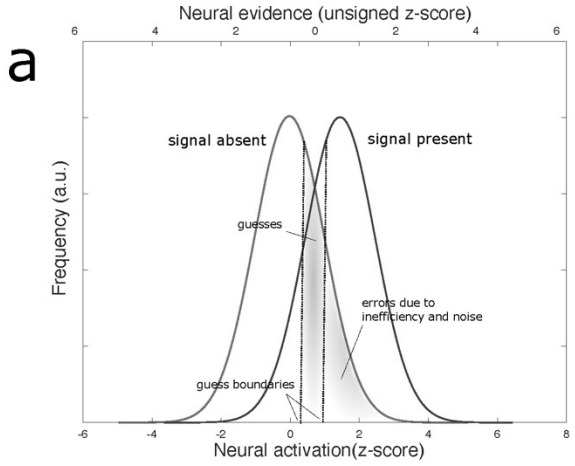
reclassification procedure to this standard SDT observer model. In the remainder of this section, we describe a hybrid observer model that can err because of inefficiency and noise, just like the standard SDT observer model, but also because of guessing. In fact, this hybrid observer model lives on a continuum between the SDT observer model (when all errors are caused by inefficiency and noise) and the threshold observer (when all errors are caused by guess), which is at the heart of the reclassification procedure. This hybrid observer model will allow us to discuss the effect of a gradual violation of the *all-errors-are-guesses* assumption of the reclassification procedure.

Consider the neural activation frequency distributions for signal-present and signal-absent trials shown in Figure 6a. In the hybrid observer model, neural evidence (represented as the top x-axis in the figure) is defined as the absolute value of the neural activation (represented as the bottom x-axis in the figure) minus the neural activation where the two neural activation frequency distributions meet. We posit that guesses are all located between two sharp guess boundaries centered on the zero neural evidence mark in both the neural activation frequency distributions. In other words, half the responses randomly chosen between these two boundaries (in both frequency distributions) are true incorrect responses while the other half are false correct responses. On Figure 6a, the area of guesses for the neural activation frequency distribution of signal absent trials is identified. An identical but mirror-reversed guess area exists in the neural activity frequency distribution of signal present trials but isn't represented on the figure to avoid crowding. Finally, all responses associated with a neural activation greater than the rightmost guess boundary in the neural activation frequency distribution for signal absent trials are inefficiency and noise errors (i.e. the false alarms in this hypothetical detection task). The area corresponding to these errors is identified on Figure 6a. Although not represented on the figure, all responses associated with a neural activation smaller than the leftmost guess boundary in the neural activation distribution for signal present trials are also inefficiency and noise errors (i.e. the misses). Importantly, all the responses between the two guess boundaries are true incorrect responses — responses resulting from the guessing process — whereas all the other responses are true correct responses — responses resulting from the somewhat noisy and inefficient recognition process, *including, somewhat counterintuitively, errors due to noise and inefficiency*. These true correct and true incorrect responses are the ground truth accuracy against which we will compare the original accuracy (the one computed by a researcher after the completion of the experiment by an observer), on the one hand, and the reclassified accuracy, on the other hand.

The isoquant maps on the left of Figure 6b show reclassification gain following the application of the reclassification procedure to this hybrid observer model as a function of the proportion of original correct responses varying between 0.7 and 0.8, and of the proportion of errors due to inefficiency and noise (vs. guesses) varying between 0.1 (almost the threshold observer model) and 0.9 (almost the SDT observer model). The reported reclassification gains are ratios of the reclassification efficiencies and the original classification efficiencies. Here, reclassification efficiency was calculated as the mean of the reclassified accuracies equal to the ground truth accuracies minus the mean of the reclassified accuracies different from the ground truth accuracies; and the original classification efficiency corresponds to the mean of the original accuracies equal to the

ground truth accuracies minus the mean of the original accuracies different from the ground truth accuracies. As mentioned in the section *From response classification efficiency to response reclassification efficiency*, the lower the proportion of original correct responses, the better the reclassification efficiency gain. More to the point, the fewer the errors due to inefficiency and noise relative to those due to guesses, the better the reclassification efficiency gain. You will recall that, for the true and false correct reclassification evidence frequency distributions to be distinguishable at all by our reclassification procedure, the reclassification evidence must be correlated with this neural evidence (see section *The reclassification procedure*). In the simulations presented on Figure 6b, the strength of this correlation decreases from the top isoquant maps to the bottom ones. As expected, the better the correlation between neural evidence and reclassification evidence the better the reclassification efficiency gain.

The isoquant maps on the right of Figure 6b show the difference between the best reclassification efficiency criterion calculated from ground truth accuracy and the one prescribed by the reclassification procedure expressed as z-scores as a function of the proportion of original correct responses and of the proportion of errors due to inefficiency and noise (vs. guesses). As you can appreciate, all these differences are either equal to zero, which means that the reclassification criterion by the reclassification procedure is optimal, or positive, which means that this prescribed reclassification criterion is liberal. Thus, the reclassification procedure tends to be liberal in the presence of errors due to noise and inefficiency. This happens partly because, under these circumstances, the reclassification procedure overestimates the number of false correct responses. Indeed, a fraction of the incorrect responses are caused by inefficiency and noise, and shouldn't be included in the false correct responses as we explained above. Furthermore, because these errors due to inefficiency and noise are not present in the original correct responses, the reclassification procedure also underestimates the number of true correct responses. These overestimation and underestimation, respectively, of the false and true correct reclassification evidence frequency distributions *both* contribute to an overestimation of the reclassification evidence criterion. Moreover, in the hybrid observer model, inefficiency and noise errors occur near the intersection of the true and false correct reclassification evidence frequency distributions, that is, where they hurt the criterion prescribed by the reclassification procedure the most. This situation is not unlike the lapses' worst-case scenario shown on Figure 5c. There is one notable difference though: Contrary to inefficiency and noise errors, lapses do not affect the true correct reclassification evidence frequency distribution. In other words, these inefficiency and noise errors are more detrimental to the reclassification procedure than lapses, everything else being equal.


a
Neural evidence (unsigned z-score)
signal absent
signal present
guesses
errors due to inefficiency and noise
guess boundaries
Frequency (a.u.)
Neural activation(z-score)

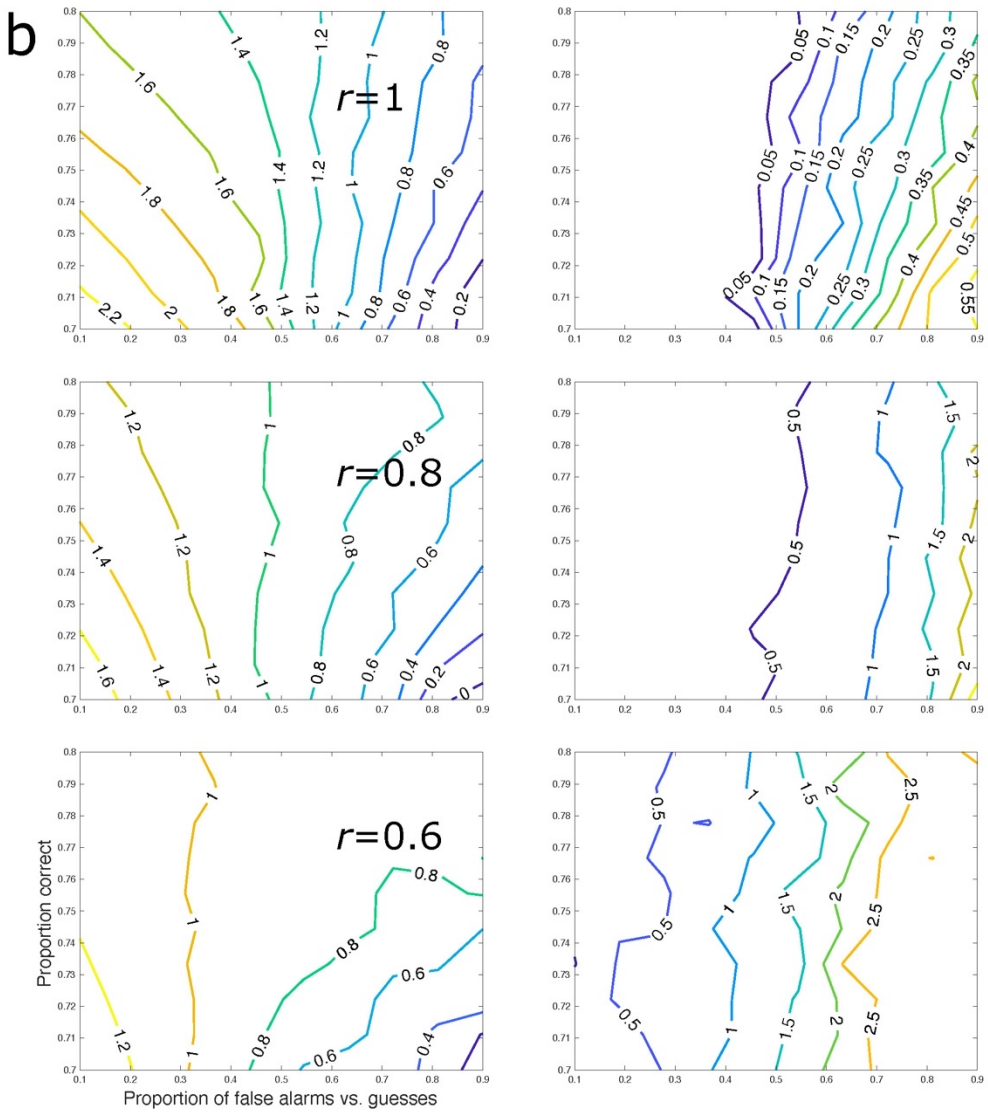

b
r=1
r=0.8
r=0.6
Proportion correct
Proportion of false alarms vs. guesses

Figure 6. Impact of applying the reclassification procedure to a hybrid observer model that makes mistakes because of both guesses and noise and inefficiency. a) Illustration of the main elements of the hybrid observer model with a proportion of correct responses equal to 0.75 and the proportion of errors due to inefficiency and noise vs. guesses equal to 0.5. Note that the areas of guesses and of errors due to inefficiency and noise are identified only on the neural activation frequency distribution of signal-absent trials to avoid crowding. However, identical but mirror-reversed areas also exist on the neural activation frequency distribution of signal present trials. b) Isoquant maps on the left show the reclassification efficiency gain following the application of the reclassification procedure on the original accuracy as a function of the proportion of original correct responses (*y* axes) and of the proportion of errors due to noise and inefficiency (vs guesses; *x* axes). Isoquant maps on the right show the difference between the best reclassification efficiency criterion obtained from ground truth accuracy and the one prescribed by the reclassification procedure expressed in z-scores as a function of the proportion of original correct responses and of the proportion of errors due to noise and inefficiency vs. guesses. The isoquant maps in successive rows, from the top to bottom, were obtained with a reclassification evidence less and less associated with the neural evidence.

Despite these theoretical limitations, in the two empirical tests of the reclassification procedure presented in this article, we did not observe evidence for a high proportion of errors due to inefficiency and noise relative to errors due to guesses. In other words, the human observers in these experiments did not behave like SDT model observers, they did not inhabit the rightmost portions of the isoquants maps of Figure 6b. This might be because these experiments differ from the psychophysical experiments that led to the formulation of the SDT model observer inasmuch as they used high-contrast stimuli with no external additive noise. It is possible that, in other situations, human observers would behave more like SDT model observers.

Multiple response alternatives

So far, we discussed the reclassification procedure exclusively in the context of tasks with two response alternatives. In this section, we explore the generalization of the procedure to tasks with multiple response alternatives. If an observer guessed among exactly x alternatives, these guesses would be (true) incorrect responses with a probability of $\frac{x-1}{x}$ and false correct responses with a probability of $\frac{1}{x}$. The incorrect and false correct responses would no longer be equiprobable as they are with two response alternatives but they would still be drawn from the same reclassification evidence distribution. This means that the false correct reclassification evidence frequency distribution would be equal to the incorrect reclassification evidence frequency distribution scaled by the factor $\frac{1}{x-1}$. And the true correct reclassification evidence frequency distribution would be equal to the correct reclassification evidence frequency distribution minus this false correct reclassification evidence frequency distribution.

The trouble is that with $A>2$ response alternatives, guesses consist of selecting one response among *a maximum of A* alternatives, not necessarily among exactly A alternatives. Indeed, it seems plausible that, after having accumulated some clues, the observer would have eliminated some response alternatives, on some trials at least, and would thus guess among the remaining alternatives. Given the proportion of responses (w_x) for which the observer would guess among $x = 2, \ldots,$ and A response alternatives, the adequate scaling factor to be applied to the original incorrect reclassification

evidence frequency distribution would be equal to $S = \sum_{x=2}^{A} \frac{w_x}{x-1}$. Unfortunately, these proportions are unknown. Unless one makes additional assumptions, the reclassification procedure thus cannot be fully generalized to more than two alternatives. For instance, we could also assume that w_x gives more weight to low numbers of alternatives (high probabilities for small x) when the task is easy, and more weight to high numbers of alternatives (high probabilities for high x) when the task is difficult. Specifically, we could posit that $w'_x = e^{kdx}$, where $d = \frac{2I_O}{N(A-1)} - 1$ is task simplicity (it varies linearly from -1 to 1 for most difficult to easiest), and k is a free slope parameter which could be determined using past studies or a cross-validation procedure. Finally, we would need to normalize the values of w_x so that they sum to 1, $w_x = \frac{w'_x}{\sum_{x=2}^{A} w'_x}$. With this scaling factor, S, we would be able to compute the false and true correct reclassification evidence frequency distributions.

Regardless of this possible definition of S, we have the following two equalities: $H + M = SI_O$ (the number of false correct responses — $H + M$ — is equal to the number of original incorrect responses multiplied by the scaling factor S) and $C + F = N - I_O(1 + S)$ (the number of true correct responses — $C + F$ — is equal to the total number of responses minus the original incorrect responses or, if you prefer, to the number of correct responses, minus the number of false correct responses $-SI_O$). Thus Equation 1 can be generalized to as follows:

$$E = \big(2I_O(1 + S) - N + 2(C - M)\big)/N. \quad \text{(Equation 3)}$$

The original classification efficiency is obtained when $C = N - I_O(1 + S)$ and $M = SI_O$. Replacing these terms in Equation 3, we obtain $E_O = (4I_O + 2SI_O + N)/N$. The maximum reclassification gain — E_O^{-1} — is thus inversely proportional to I_O. In other words, the more difficult the task, the more to be gained from the reclassification procedure. This generalizes the result already presented in the context of tasks with two response alternatives to tasks with more than two response alternatives. Furthermore, E_O^{-1} is inversely proportional to S, which, under the additional assumptions presented above, increases with A. In fact, S increases with A, under all additional assumptions for which the model observer sometimes guesses between more than two alternatives. Therefore, the more response alternatives in a task, the less to be gained, in all likelihood, from the reclassification procedure.

Other reclassification evidences

Good reclassification evidences must have two features: First, they must contain response reclassification information. The best reclassification evidences with respect to this feature will identify all true and false correct responses among the original correct responses (and lead to a response reclassification gain of E_O^{-1}). The worst reclassification evidences will lead to a null reclassification. Most reclassification evidences lead to an intermediate situation and allow to reclassify some original correct responses — richer in false correct responses than in true correct responses — as incorrect responses. Second, collecting good reclassification evidences must imply little additional efforts from the experimenters. Response times (RT) can be — and often are — measured concomitantly with the responses themselves with no additional effort whatsoever so

they are an obvious choice as reclassification evidence. But the procedure outlined in this article can be applied readily to other sources of trial-by-trial reclassification evidence.

Another piece of reclassification evidence that could be used in Bubbles experiments such as Faghel-Soubeyrand et al. (2019) is the similarity between the classification image computed from standard accuracies and the bubble masks presented on each trial such as a Pearson correlation. This is highly likely to contain reclassification information. Indeed, a participant is most likely to guess when the information used by this participant to do a task is masked. Although we can't compute new classification images from these reclassified accuracies and compare the SNR gain without "double dipping", we can compare the *predicted* SNR gains ($\sqrt{E_{max}} / \sqrt{0.5}$, with E_{max} the maximum E obtained from Equation 2) of the two reclassification evidences: 106.84% for RT (which, as you might recall, is smaller than the SNR gain of 113.5% that we actually measured) and an astonishing 141% for the classification-image similarities — this is very close to the maximum reclassification gain of $\sqrt{2}$ that can be observed in this experiment. These similarities could also be used for other purposes. For example, we have used a similar procedure to better separate electroencephalographic (EEG) data in true correct and true incorrect responses (in fact, this is similar to what Caplette et al., 2020, actually did). A cross-validation procedure could also be used in this case to reduce or eliminate double dipping (Kriegeskorte et al., 2009). For example, we could compute a classification image using half the data, calculate the similarities between this classification image and the bubble masks presented on the other half, reclassify the accuracies of this other half using these similarities, repeat the procedure swapping the data halves, compute two classification images with these reclassified accuracies and combine these classification images.

Gaze, pupil dilation and spontaneous blink rate, which can all be measured at the same time with a video eye-tracker, seem particularly promising as reclassification evidences. Gaze correlates with overt attention (for a review, see Hollingworth & Bahle, 2019) and with use of information (e.g., Blais et al., 2017), which are both likely associated with accumulation of evidence. Uggeldahl et al. (2016) showed that the frequency of gaze shifting in a discrete choice experiment is also associated with choice uncertainty, which is likely to occur before a guess. Gaze-tracking only requires a brief calibration and the occasional drift correction. Pupil dilation has been used as a measure of subjective task difficulty, which is what a participant struggling to accumulate enough cues to make an informed decision would experience, while spontaneous eyeblink rate correlates with processes underlying reinforcement learning (for a review, see Eckstein et al., 2017). In sum, all these video eye-tracker measures provide a lot of information about cognitive processes mostly at a moderate temporal frequency rate (3-4 Hz). If a remote video eye-tracker is used — a type of eye-tracker that is becoming very affordable — they can be measured without disrupting participants.

Another promising source of reclassification evidence is electroencephalography (EEG). Sheldon and Mathewson (2021) recently discovered that, in a visual orientation discrimination task, guesses were less frequent on trials with high EEG power in the 2- to 3-Hz bandwidths than low, and the difference started around 250-ms after stimulus onset. More generally, EEG provides rich broadband information (3-100 Hz) about

perception and cognition. Some dry-electrode systems are inexpensive and fairly quick to setup. Both EEG and eye-tracker produce multivariate information that could be used as reclassification evidence. To combine the different EEG channels, for example, various fusion methods could be used, including machine learning (e.g., Roy et al., 2019). We could attempt to maximize the predicted reclassification efficiency from all these reclassification evidence variables using cross-validation. Then we could use this global reclassification evidence to derive false correct and true correct frequency distributions exactly like we did in this paper with RT. The same scheme could be used to combine different sources of reclassification evidence such as EEG and eye-tracker (for a related proposal, see Qian et al., 2009).

Conclusion

We presented and tested a procedure that can increase signal-to-noise ratio (SNR) in psychological experiments that use accuracy as a *selection variable* for another dependant variable. This procedure reclassifies some original correct responses as incorrect responses using reclassification evidence. We show that the more difficult the task and the fewer the response alternatives, the more there is to be gained from reclassification. In fact, we observed SNR gains of about 113.5% and 120.3%, when reanalyzing the data of Faghel-Soubeyrand et al. (2019) and Caplette et al. (2020), respectively, using response time as reclassification evidence. Matlab and Python implementations of the reclassification procedure for two response alternatives presented in this article are freely available (https://github.com/GroupeLaboGosselin/Reclassification). We encourage researchers of trying it at least with response times since response times are usually collected in all behavioral tasks.

Appendix: Simulated Data

We simulated 5,000,000 trials, each comprising an accuracy and a reclassification evidence, to generate the data used to produce Figures 1 and 2. On each simulated trial, randomly selected a number between 0 and 1. This represented the quantity of neural evidence accumulated by the observer in favor of one of two response alternatives. A logistic function

$$f(x) = \frac{1}{1+e^{-k(x-x_0)}},$$

centered on x_0 = 0.5 and with a slope of k =11, was then applied to this accumulated neural evidence and determined the probability that the observer provided a true correct response. Note that the smaller the slope parameter *k*, the smaller the reclassification efficiency. Another number was randomly chosen between 0 and 1. If this number was smaller than the probability that the observer gave a true correct response, this trial's response was a true correct response; otherwise, it was an incorrect response, resulting from a guess. These true correct and incorrect responses are the omniscient being's classification — the ground truth. The actual response of the model observer, in this last case, was a false correct response with a 0.5 probability; otherwise, it was an incorrect response. The original response classification was all 5,000,000 incorrect and correct — true and false correct — responses of the model observer. The goal of the reclassification procedure is to recover the ground truth as much as possible from this original response classification using some reclassification evidences.

Next, we simulated the reclassification evidence for every trial. We drew a number randomly from a two-parameter Weibull density function by applying the Weibull inverse cumulative density function to a random number selected from a uniform distribution varying between 0 and 1:

$$f(x) = \alpha\big(-log(x)\big)^{\frac{1}{\beta}}$$

where β, the shape parameter, was equal to 4, and α, the scale parameter, was equal to the following inverted logistic function of neural evidence:

$$f(x) = 3 - \frac{2}{1 + e^{-k(x-x_0)}}$$

with x_0 = 0.5 and k = 10. The mean of this Weibull density function is equal to $\alpha\Gamma\left(1+\frac{1}{\beta}\right)$, where Γ is the gamma function. These stochastic relationships between neural evidence and the Weibull scale parameter — and thus the mean of the Weibull density function — , on the one hand, and between neural evidence and accuracy, on the other hand, is what made this reclassification evidence useful for reclassification.

References

Berger, A., & Kiefer, M. (2021). Comparison of different response time outlier exclusion methods: A simulation study. *Frontiers in Psychology*, *12*.

Blais, C., Fiset, D., Roy, C., Saumure Régimbald, C., & Gosselin, F. (2017). Eye fixation patterns for categorizing static and dynamic facial expressions. *Emotion*, *17*(7), 1107.

Caplette, L., Ince, R. A. A., Jerbi, K., & Gosselin, F. (2020). Disentangling presentation and processing times in the brain. *NeuroImage, 218*, 116994.

Chauvin, A., Worsley, K. J., Schyns, P. G., Arguin, M., & Gosselin, F. (2005). Accurate statistical tests for smooth classification images. *Journal of vision*, *5*(9), 1-1.

Dupuis-Roy, N., Fortin, I., Fiset, D., & Gosselin, F. (2009). Uncovering gender discrimination cues in a realistic setting. *Journal of Vision*, *9*(2):10, 1-8.

Eckstein, M. K., Guerra-Carrillo, B., Singley, A. T. M., & Bunge, S. A. (2017). Beyond eye gaze: What else can eyetracking reveal about cognition and cognitive development?. *Developmental cognitive neuroscience*, *25*, 69-91.

Eriksen, C. W. (1988). A source of error in attempts to distinguish coactivation from separate activation in the perception of redundant targets. Perception & Psychophysics, 44, 191_193.

Faghel-Soubeyrand, S., Dupuis-Roy, N. & Gosselin, F. (2019). Inducing the use of right eye enhances face-sex categorization performance. *Journal of Experimental Psychology: General, 148,* 1834-1841.

García-Pérez, M. A. (2010). Denoising forced-choice detection data. *British Journal of Mathematical and Statistical Psychology*, *63*(1), 75-100.

Glickman, M. E., Gray, J. R., & Morales, C. J. (2005). Combining speed and accuracy to assess error-free cognitive processes. *psychometrika*, *70*(3), 405-425.

Gondan, M., & Heckel, A. (2008). Testing the race inequality: A simple correction procedure for fast guesses. *Journal of Mathematical Psychology*, *52*(5), 322-325.

Granholm, E.& Steinhauer, S. R. (2004).Pupillometric measures of cognitive and emotional processes. *International Journal of Psychophysiology*, *52*, 1-6.

Guo, H., Rios, J. A., Haberman, S., Liu, O. L., Wang, J., & Paek, I. (2016). A new procedure for detection of students' rapid guessing responses using response time. Applied Measurement in Education, 29, 173–183. doi:10.1080/ 08957347.2016.1171766

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (Vol. 1, pp. 1969-12). New York: Wiley.

Hautus, M. J., Macmillan, N. A., & Creelman, C. D. (2021). *Detection theory: A user's guide*. Routledge.

Hollingworth, A., & Bahle, B. (2019). Eye tracking in visual search experiments. In *Spatial Learning and Attention Guidance* (pp. 23-35). Humana, New York, NY.

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., & Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nature neuroscience*, *12*(5), 535-540.

Lee, Y. H., & Jia, Y. (2014). Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. Large-scale Assessments in Education, 2, 8, 1–24. doi:10.1186/s40536-014-0008-1

Manning, C., Dakin, S. C., Tibber, M. S., & Pellicano, E. (2014). Averaging, not internal noise, limits the development of coherent motion processing. *Developmental Cognitive Neuroscience, 10*, 44-56.

Miller, J., & Lopes, A. (1991). Bias produced by fast guessing in distribution-based tests of race models. *Perception & Psychophysics*, *50*(6), 584-590.

Qian, M., Aguilar, M., Zachery, K. N., Privitera, C., Klein, S., Carney, T., & Nolte, L. W. (2009). Decision-level fusion of EEG and pupil features for single-trial visual detection analysis. *IEEE Transactions on biomedical Engineering*, *56*(7), 1929-1937.

Roy, Y., Banville, H., Albuquerque, I., Gramfort, A., Falk, T. H., & Faubert, J. (2019). Deep learning-based electroencephalography analysis: a systematic review. *Journal of neural engineering*, *16*(5), 051001.

Schnipke, D. L. (1995). Assessing speededness in computer-based tests using item response times (Unpublished doctoral dissertation). Baltimore, MD: Johns Hopkins University.

Sheldon, S. S., & Mathewson, K. E. (2021). To see, not to see or to see poorly: Perceptual quality and guess rate as a function of electroencephalography (EEG) brain activity in an orientation perception task. *European Journal of Neuroscience*.

John, T. (1977). Exploratory Data Analysis Addison-Wesley. *Reading, MA*.

Uggeldahl, K., Jacobsen, C., Lundhede, T. H., & Olsen, S. B. (2016). Choice certainty in discrete choice experiments: Will eye tracking provide useful measures?. *Journal of choice modelling*, *20*, 35-48.

Wagner, A. D., Koutstaal, W., and Schacter, D. L. (1999). When encoding yields remembering: insights from event-related neuroimaging. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 354, 1307–1324. doi: 10.1098/rstb.1999.0481

Watson, A. B., & Pelli, D. G. (1983). QUEST: A Bayesian adaptive psychometric method. Attention, Perception & Psychophysics, 33, 113–120. doi: 10.3758/BF03202828

Wise, S. L. (2019). An Information-Based Approach to Identifying Rapid-Guessing Thresholds, Applied Measurement in Education, 32:4, 325-336, DOI: 10.1080/08957347.2019.1660350

Wise, S. L., & Ma, L. (2012, April). Setting response time thresholds for a CAT item pool: The normative threshold method. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, Canada.