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

Frédéric Gosselin, Laurent Caplette, Valérie Daigneault and Jean-Maxime Larouche
Département de psychologie, Université de Montréal

Abstract

Researchers studying the mind often rely on behavioral tasks and differences, either in stimuli or in brain activity, between correct and incorrect trials. However, subjects often guess when they don't know the answer, leading to correct responses that result from the same causes as the incorrect responses: this is a source of noise that remains no matter the number of trials performed by the subjects. This paper presents a response reclassification procedure to reduce the noise caused by "false" correct responses using an independent source of reclassification evidence. We illustrate the procedure on data from Faghel-Soubeyrand et al. (2019) with response time as reclassification evidence. The reclassification procedure increased signal-to-noise ratio by about 13.5% with little bias. 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. Differences in stimuli or in brain states between correct and incorrect responses are then analyzed in various ways. When subjects don't know for sure which response alternative to choose from, however, they guess. Since guesses are sometimes correct only by chance, this leads to a proportion of the correct responses that result from the same causes as the incorrect responses, not those of the other correct responses. These false correct responses, as we call them, constitute an important source of noise that remains no matter the number of trials in an experiment. When there are two response alternatives and the correct response rate is 75%, for example, close to 25% of all responses are false correct responses. If all these false correct responses and only these false correct responses were reclassified as incorrect responses, the signal-to-noise ratio (SNR) would be increased by about 40%. Here, we present a new method that uses independent evidences to reclassify some of these false correct responses into incorrect responses to increase SNR in psychological experiments. We tested the method with response times as reclassification evidence on the data set from Faghel-Soubeyrand et al. (2019) comprosing 140 participants, and observed an increase in SNR of about 13.5%.

False correct responses, guesses and other useful definitions

We will suppose that a participant accumulates clues consciously or not in favor of the different response alternatives. Eventually, this participant has to select a response among the *A* response alternatives. We will use the terms *correct* and *incorrect* responses to refer to responses that a fallible being, such as 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. The original correct and incorrect responses of these original classifications are what psychologists refer to by correct and incorrect responses, respectively. In general, the correct responses consist of *true* and *false* correct responses – responses that an omniscient being would put in the "correct" and "incorrect" categories, respectively; and the incorrect responses consist in *true* and *false* incorrect responses – responses that an omniscient being would put in the "incorrect" and "correct" categories, respectively.

If an insufficient amount of information has been gathered by the participant to choose one response with a sufficiently high degree of certainty, the participant will *guess*. We will assume that, with two response alternatives, the focus of this article, guessing leads to a false correct response and to an incorrect response with the same probability of 0.5. For example, selecting one of the response alternatives randomly satisfies this assumption. We will assume for now that only *guesses* can lead an original response classification to contain incorrect and false correct responses at this stage. In subsequent sections, we will discuss other possible causes as well as their impact on the reclassification procedure introduced in this article.

Response classification efficiency

The proportion of all responses that are true correct and true incorrect responses might seem like a reasonable index of the efficiency of a 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 or a false

incorrect 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 between correct and incorrect responses would be slightly improved because it would prevent the cancellation so to speak of the effect of one true incorrect response by this false correct response. In fact, this improvement should be the same as adding exactly one response to the experiment, assuming that all trials are equivalent or interchangeable. Now, if you put this false correct response in the "incorrect" category, you would gain one true 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 classification efficiency, the difference between the proportion of all responses that are true correct or true incorrect responses, and the proportion of all responses that are false correct or false incorrect 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 responses. In practice, however, it should never go below 0, which is the expected response classification efficiency of a random response classification. For example, if the proportion of original correct responses is 0.75, the response classification efficiency is equal to 0.5 that is, a proportion of 0.25 of true incorrect responses plus a proportion of 0.50 of true correct responses minus a proportion of 0 of false incorrect responses and, finally, minus a proportion of 0.25 of false correct responses.

The remainder of this article will pertain to *response reclassification* — altering the original response classification with the goal of increasing response classification 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. 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*. We can thus express the above measure of response classification efficiency for reclassification as follows:

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

with E, the efficiency of the response reclassification; I_O, the number of original incorrect responses; H, the number of reclassification's hits; C, the number of reclassification's correct rejections; M, the number of reclassification's misses; F, reclassification's false alarms; and N, the total number of responses. The goal of the reclassification procedure can be framed as maximizing Equation 1 given *some reclassification evidence* that original correct responses might in fact be false correct responses.

The reclassification procedure

We will illustrate the reclassification procedure using response time (RT) as reclassification evidence. We believe that RT will likely be the most popular reclassification evidence because RT and the responses themselves can be — and typically are — measured concomitantly without additional effort. That being said, RT is by no means the only possible reclassification evidence, and almost certainly not the best one. We will discuss other promising reclassification evidences toward the end of this article. What makes RT potentially useful for reclassification is that it tends to be faster on average for correct than for incorrect responses during object recognition (e.g. Luce, 1986). This is shown on Figure 1 for a hypothetical experiment with 75% of original correct responses. The dashed gray line represents a hypothetical distribution of original correct RT and

the solid gray line, the companion slower hypothetical distribution of incorrect RT (identified as *false correct responses* in the legend for a reason that should soon become clear).

The reclassification procedure provides a reclassification evidence criterion — above or below which original correct responses are reclassified as incorrect responses — or, in the particular case we are concerned with, an RT criterion — above which original correct responses are reclassified as incorrect responses — that maximizes Equation 1. To compute *H*, *C*, *M* and *F*, the unknowns of Equation 1, we need the RT frequency distributions of true and false correct responses. Given our assumption that all false correct responses result from guesses, false correct RT and the original incorrect RT must be drawn from the same population and are equiprobable with two response alternatives. In other words, the false correct RT frequency distribution is identical to the incorrect RT frequency distribution (Figure 1, solid gray line). Furthermore, the original correct RT frequency distribution is the sum of the true and false correct RT frequency distributions. Therefore, the true correct RT frequency distribution (Figure 1, solid black line) is equal to the difference between the correct RT frequency distribution (Figure 1, dashed black line) and the false correct RT frequency distribution.

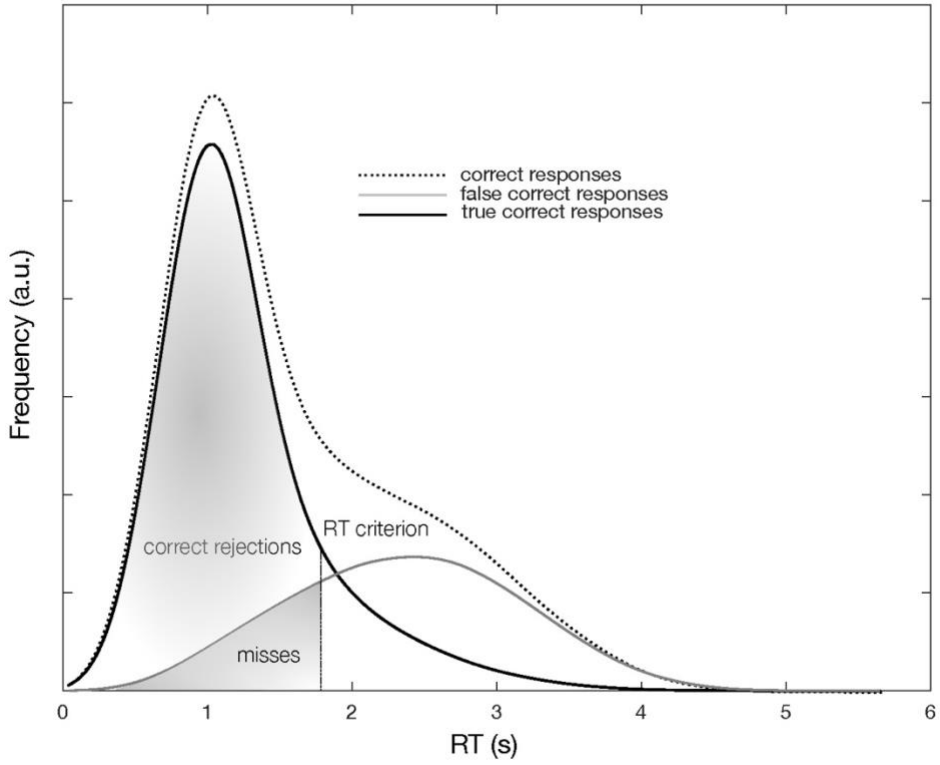

Figure 1. Hypothetical frequency RT distributions of original correct, true correct and false correct responses. Here, the frequency RT distribution of false correct responses is identical to the frequency RT distribution of original incorrect responses. The correct rejection and miss areas are shown with respect to an arbitrary RT criterion.

With these true and false correct RT frequency distributions under our belt, we can calculate *H*, *C*, *M* and *F* as a function of RT criterion. In fact, $H + M = I_O$ and $C + F = N - 2\ I_O$ with two response alternatives. Thus Equation 1 can be rewritten as follows:

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

The number of true correct responses below a given RT criterion, like the one shown on Figure 1, is equal to *C* (i.e. the number of reclassification's correct rejections); similarly, the number of false correct responses below the same given RT criterion is equal to *M* (i.e. the number of reclassification's misses). The curve in Figure 2 represents reclassification efficiency as a function of RT criterion for the true and false correct RT 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 RT criterion of 1.885 s, and, then, slowly goes down to the efficiency of the original classification which, in this case, is 0.5 (when no original correct responses are reclassified). The reclassification efficiency is thus equivalent to an increase of 47% of responses (0.736/0.5) or a SNR increase of about 21% (√0.470), assuming that slower and faster false correct responses carry the same information. Note that Equation 2 peaks where (*C* - *M*) peaks. What the remainder of the equation does is modify linearly this simplified equation to provide a reclassification efficiency that varies between 0 and 1 from random to perfect reclassification. This maximum is attained where the derivative of Equation 2 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 and false beliefs on the reclassification procedure later on.

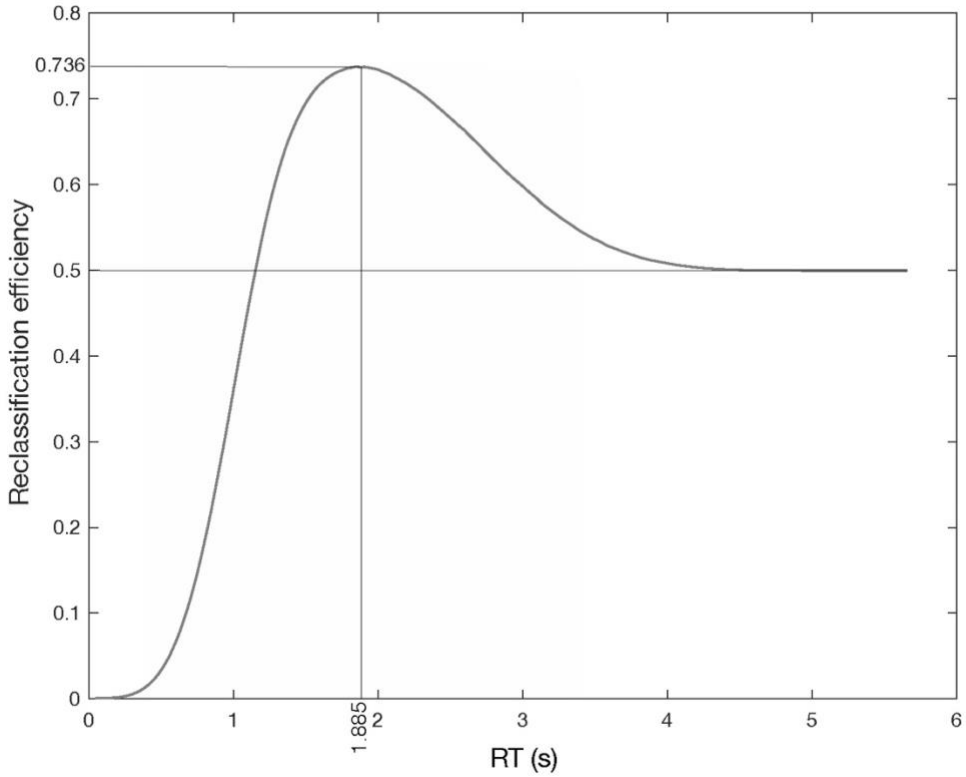

Figure 2. Reclassification efficiency as a function of RT based on the hypothetical true correct and false correct RT frequency distributions shown in Figure 1. The maximum efficiency of about 0.736 is attained with a RT criterion of about 1.885 s. The height of the tail of the reclassification efficiency curve, 0.5, corresponds to the efficiency of the original classification.

A special reclassification is the null reclassification that is, the original response classification. In the original reclassification, $C = N - 2I_O$ and $M = I_O$; replacing these terms in Equation 2, we obtain that the efficiency of the original classification is $E_O = 1 - \frac{2I_O}{N}$. For a proportion of original correct

responses of 0.75, for instance, as in our above example, this gives us $E_O = 0.5$. And the maximum reclassification efficiency gain is equal to $E_O^{-1} = 2$ (i.e. 1/0.5). This is equivalent to doubling the number of responses. Or, put otherwise, it represents a SNR increase of $\sqrt{2}$.

Empirical test of the reclassification procedure

We tested the reclassification procedure with the Faghel-Soubeyrand et al. (2019) data set (available at https://github.com/GroupeLaboGosselin/Reclassification). These researchers examined the use of facial features in 140 individuals during a sex discrimination Bubbles task. Three hundred color face images (150 men) 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. The number of bubbles was adjusted on a trial-by-trial basis to maintain accuracy as close as possible from a proportion of 75% correct throughout the entire experiment. Stimuli subtended 3.08 x 3.08 degrees of visual angle (128 x 128 pixels). Each participant completed 300 trials, which is relatively few trials for a Bubbles experiment. This is an opportunity to test the SNR gain following reclassification near the limit of the Bubbles method's sensitivity.

The mean accuracy was 75.09% (standard deviation across subjects (SD) = 1.68%), which is quite close to target. To achieve this accuracy rate, a mean number of bubbles of 39.57 (SD = 13.40) was required. The mean RT was 1.61 s (SD = 0.91 s). The mean original incorrect and correct RT were 1.50 s (SD = 0.83) and 1.93 s (SD = 1.18), 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). An average of 9.97% of all responses were reclassified (SD = 7.67%). The average predicted reclassification efficiency was 57.16% (SD = 6.06%). The mean predicted reclassification gain that is, the ratio between the predicted reclassification efficiency and the null reclassification efficiency, was 114.14% (SD = 11.80%). This is equivalent to a predicted SNR gain of 106.84%, assuming that all responses carry the same amount of information.

To test if the response reclassifications translated to an increase in the SNR in the classification images, we did the following: We summed the centers of all bubbles in each trial, smoothed the resulting images with a Gaussian filter having a standard deviation of 5 pixels, and z-scored them. The individual classification images (CI) were computed by summing the z-scored bubble masks weighted by accuracies (or reclassified accuracies, RT and RT ranks), transformed in z-scores, and by dividing the resulting image by the square root of the number of trials (i.e. $\sqrt{300}$). The group CIs were obtained by summing the 140 individual classification images and dividing the sum by square of the number of participants (i.e. $\sqrt{140}$). We compared the group CI derived from accuracies with the one derived from reclassified accuracies but also with the ones derived from RT and from RT ranks (see Figure 3). As a measure of SNR, we used the standard deviation of the pixels of the group CIs. This makes sense as a global measure of SNR because the CI are transformed in z-scores. So the greater the departure from a standard deviation of 1, the more signal in a CI. As a measure of SNR gain, we divided the standard deviation of the group CI obtained from reclassified accuracy by the standard deviation of the group CI obtained from the

standard accuracy. Reclassified accuracy led to a SNR gain of 113.46%. This is higher than corresponding SNR gains from the group CIs derived from RT (89.40%) and from RT ranks (104.46%) also shown on Figure 3. It is also higher than what was predicted from Equation 2. This suggests that the slower false correct responses contain more signal than faster ones or that slower true correct response contain less signal than faster ones. If we only include the individual CI for which at least 1 response was reclassified (132/140 or a proportion of 94.29%), the SNR gain goes up to 114.52%. If only the 65 subjects that had at least 10% of their responses above RT criterion are included, the average SNR gain becomes 116.86% (see Figure 5; this criterion probably seems arbitrary at this stage but it should become less so once you'll have read the section on lapses).

It is possible that slower false correct responses — the ones that were reclassified — entail different mechanisms than faster ones — the ones that weren't reclassified. This is all the more pertinent that, as we have mentioned earlier, reclassified responses seem to contain more signal than the average response. Is it only quantitatively different or is it also qualitatively different? As a global measure of bias, we used the complement of the Pearson correlation between the standard group CI and the reclassified group CI. This measure is equal to 0.038 here. If we only include the individual CIs for which at least 1 response was reclassified, this bias goes up a little to 0.043, and if we include only the individual CIs for which at least 6% of responses are above criterion this bias attains 0.079. In any case, these compare favorably to the biases of 0.205 for the RT derived group CI and of 0.197 the RT rank derived group CI. Note also that the maximum z-score are located in the eye region on the left of the image for the accuracy and reclassified group CIs whereas they are located in the opposite eye region for the RT and RT rank group CIs. In sum, the reclassification procedure seems to work: It increased SNR by about 13.5% with little bias for the Faghel-Soubeyrand et al. (2019) data set.

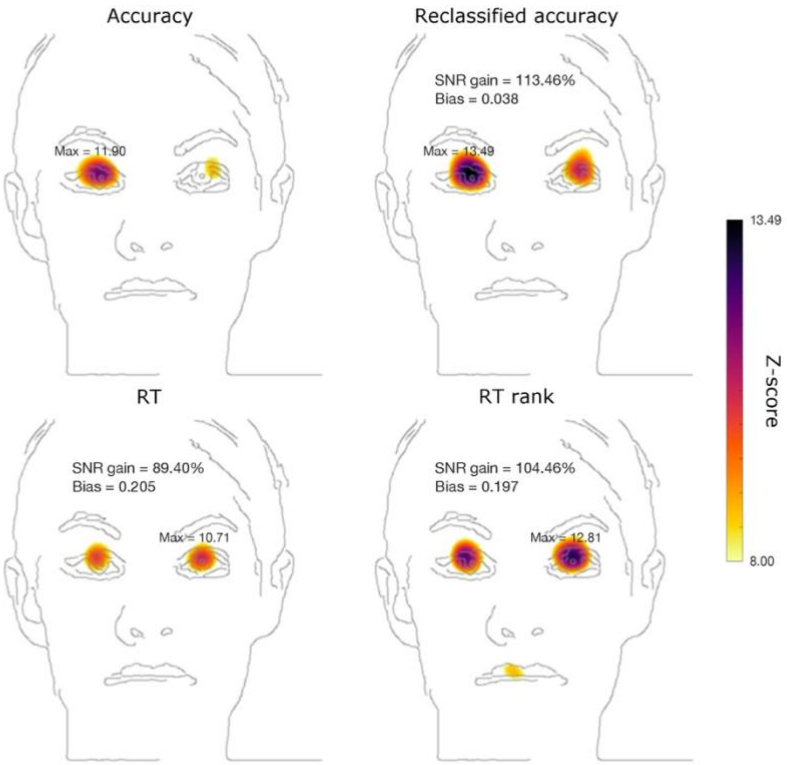

Figure 3. High values in classification images (CI) computed from different regressors for the Faghel-Soubeyrand et al. (2019) data set. The used color map is perceptually uniform.

Lapses

When the participants' vigilance drops, when they are distracted, when they look away, when they blink or, more generally, when they respond irrespective of what was asked of them, participants lapse. These lapses 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 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 of true incorrect. 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, lapses result is an underestimation of the true reclassification evidence criterion. The extent of this underestimation depends on the

characteristics of 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. A little bit like the tide that runs quickly toward the shore when the slope is gentle.

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 given by $\frac{L}{2\sigma\sqrt{2\pi}}$. This is how much the false correct reclassification evidence frequency distribution would be pushed up in the worse-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 4, we'll use a liberal lapse rate of 0.05. Figures 4c and 4d illustrate, respectively, a narrow —small σ — and an extended — large σ — lapses reclassification evidence frequency distributions centered on the optimal reclassification criterion — the worse-case scenario mentioned above. Note that the effect of lapses is very small in Figure 4d. And for the more probable scenarios — lapses reclassification evidence frequency distributions not centered on the optimal criterion — the effect of lapses is negligible (see Figure 4a and 4b).

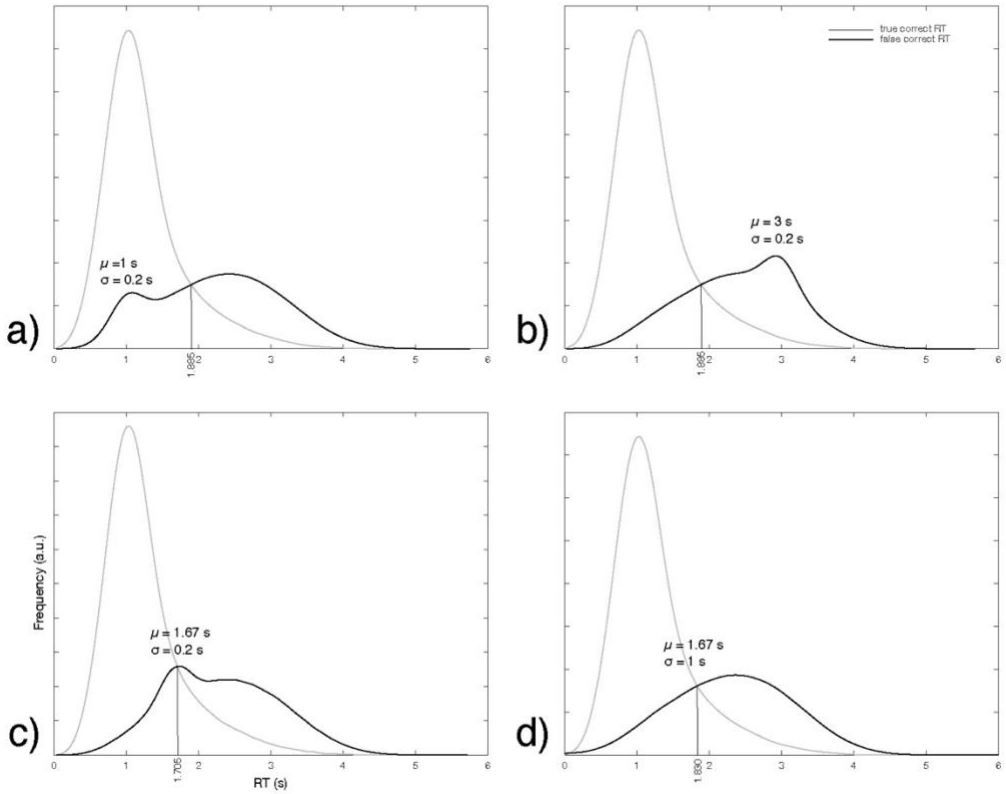

Figure 4. These are the same hypothetical RT frequency distributions than those presented in Figure 1, except for lapses. We modeled lapses RT frequency distributions as Gaussian frequency distributions with an area corresponding to a liberal lapse rate of 0.05. Lapses with a narrow RT frequency distribution σ = 0.2 s in these

examples) and a) clearly below (μ = 1 s in this example) or b) clearly above criterion (μ = 3 s in this example) do not interfere with the meeting point between the hypothetical false correct and the true correct RT frequency distributions—hence they do not impact RT criterion. Lapses near RT criterion (μ = 1.67 s in these examples) lead to an underestimation of this criterion more pronounced for c) narrow (σ = 0.2 s in this example) than d) broad RT frequency distributions (σ = 1 s in this example).

It thus seems unlikely that lapses make much difference on the reclassification evidence criterion. In the end, however, it's an empirical question. We looked for traces of lapses in the Faghel-Soubeyrand et al. (2019) data set. Specifically, we tested if criterion above the chosen criterion lead to greater SNR than this chosen criterion. Only the 65 subjects that had at least 10% of responses above criterion were included in this analysis. We tried criterions in the vicinity of the selected criterion by steps of 1% of all responses (or 3 responses per participant). Results are presented in Figure 5. The maximum SNR gain is attained just 1% of all responses prior to the criterion selected by the reclassification procedure. Note that the SNR gain is greater than for the entire subject sample, as we have already mentioned above. This was expected because the entire sample contains subjects for which the procedure lead to a null reclassification. Importantly, SNR gain drops after the selected criterion contrary to what would happen if lapses impacted this criterion. This suggests that the effect of lapses on the reclassification procedure is negligible in practice.

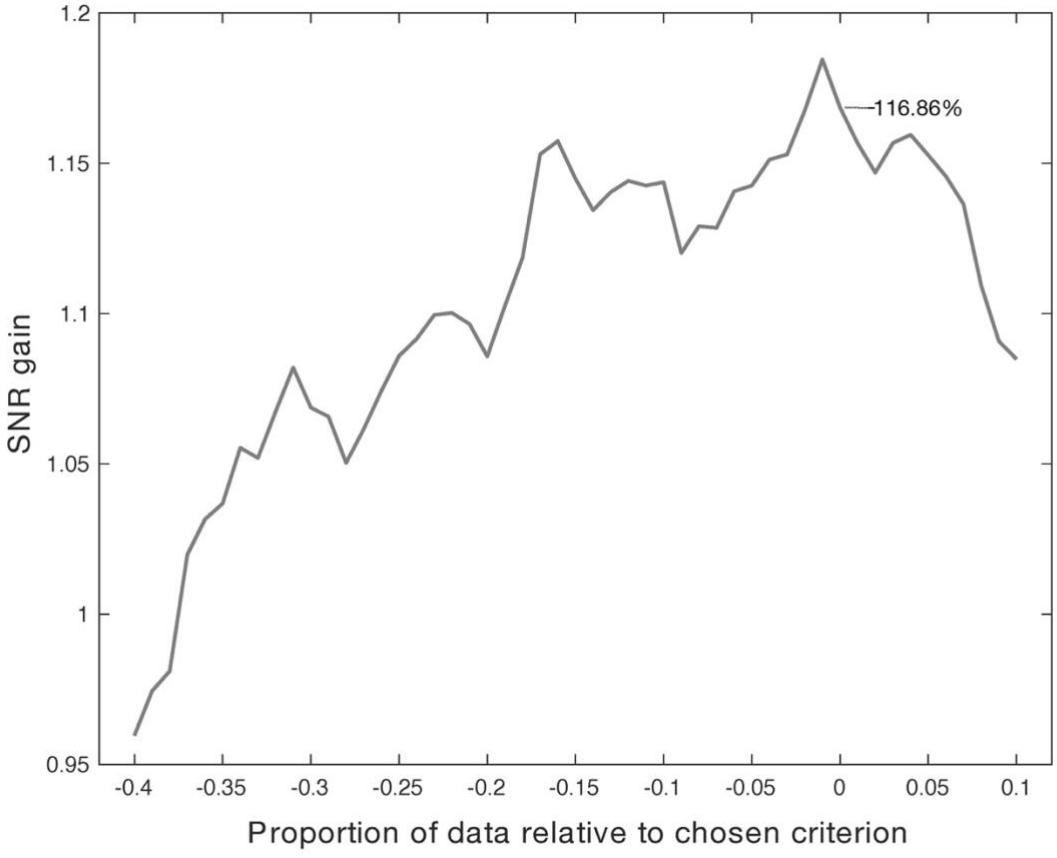

Figure 5. SNR gain as a function of the proportion of sorted RT included above or below criterion selected by the reclassification procedure by steps corresponding to 1% of all responses.

False beliefs

If the information of a stimuli is partly revealed, as in a Bubbles experiment, for example, false beliefs—when a participant overestimate or underestimate the weight of a particular clue for a given response alternative—can lead to incorrect responses but not to false correct responses. Indeed, false believes could dominate for some stimulus samples and, as a result, the participant could select the wrong response alternative. This is potentially problematic for the reclassification procedure presented in this article. Remember that our estimation of the distribution of false correct responses is the distribution of original incorrect responses. If a fraction of these original incorrect responses are the result of false believes and if none of the false correct responses are the result of false believes, we would overestimate the number of false correct responses and, hence, we would underestimate the number of true correct responses. Whether this would make a difference or not on the reclassification evidence criterion depends on the number of original incorrect responses due to false believes and on the characteristics of the reclassification evidence frequency distribution for original incorrect due to false believes. This situation is similar to that following lapses described above. There is one notable difference though: Contrary to false believes, lapses do not affect the true correct reclassification evidence frequency distribution. Thus the overestimation and the underestimation, respectively, of the false and the true correct reclassification evidence frequency distributions *both* contribute to underestimate the reclassification evidence criterion following false beliefs. In other words, false believes are more detrimental to the reclassification procedure than lapses, everything else being equal. As we wrote in the section on lapses, however, we found no empirical support for such underestimation of the reclassification evidence criterion in the data of Faghel-Soubeyrand et al. (2019).

Multiple 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 $\frac{1}{x-1}$. The true correct reclassification evidence frequency distribution would remain equal to the correct reclassification evidence frequency distribution minus the false correct reclassification evidence frequency distribution.

The trouble is, 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 responses 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 would be equal to $S = \sum_{x=2}^{A} \frac{w_x}{x-1}$. Unfortunately, these proportions are unknown.

Here, we will 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 will 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 will 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 can compute the false and true correct reclassification evidence frequency distributions. We also have the following two equalities: $C + F = N - I_O(1 + S)$, and $H + M = SI_O$. Thus Equation 1 can be rewritten 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 $E_O = (4I_O + 2SI_O + N)/N$. E_O is thus proportional to S which, in turn, increases with A. In other words, the more response alternatives, the more efficient the original classification. But the maximum reclassification gain—E_O^{-1}—is inversely proportional to S. This makes the reclassification procedure less and less appealing as the number of response alternatives increases.

Other reclassification evidences

Response times (RT) can be — and often are — measured concomitantly with the responses themselves so they are an obvious choice as reclassification evidence. But the procedure outlined in this article can be applied directly to other sources of reclassification evidence. These must have two features: first, they must contain response reclassification information; and, second, they must imply little extra efforts. Another piece of reclassification evidence that could be used in Faghel-Soubeyrand et al. (2019) is the similarity between the classification image computed from standard accuracies and the bubble masks presented on each trial—a Pearson correlation, for example. 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 efficiency gain for each reclassification evidence: 114.14% for RT and an astonishing 198.86% for these similarities. 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 (Caplette et al., 2020). A cross-validation procedure could also be used in this case to reduce or eliminate double-dipping. 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-tracking and pupillometry seem particularly promising as reclassification evidences. Gaze correlates with overt attention, which correlates with use of information. Pupillometry correlates with increased processing in the brain, which is what would happen if a participant was struggling to accumulate enough cues to make an informed decision. Both techniques provide a lot of information about perception mostly at a moderate temporal frequency rate (3-4 Hz). They only require a brief calibration and the occasional drift correction. If a remote eye-tracker is used — a type of eye-tracker that is becoming very affordable — they are entirely non-invasive. Another promising source of reclassification evidence is electroencephalography (EEG). It measures cortical activity which most probably correlates in multiple ways with the failure to accumulate enough clues for a particular response alternative. It provides a large quantity of information about perception and cognition at a broad band or temporal frequencies (3-100 Hz). Some dry-electrode systems are inexpensive and fairly quick to setup. Both EEG and gaze-tracking/pupillometry

produce multivariate reclassification evidences. To combine the different components of each of these reclassification evidences, machine learning could be used. We could attempt to predict the original correct and incorrect responses with a continuous output from all these reclassification evidence variables using cross-validation. Then we could use this overall reclassification evidence to derive false correct and true correct frequency distributions just like we did in this paper with RT. The same scheme could be used to combine different sources of reclassification evidence.

Conclusion

We presented and tested a reclassification procedure that reclassify some original correct responses as incorrect responses using reclassification evidence. We observed a SNR gain of about 113.5% with little bias in the analyses of the data of Faghel-Soubeyrand et al. (2019) 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). Researchers should use it at least with response times because collecting response times requires no additional efforts. The reclassification procedure should be tested with additional reclassification evidences and with multiple response alternatives.

References

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

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.

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.

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

Luce, R. Duncan (1986). Response times: their role in inferring elementary mental organization. New York: Oxford.

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.