Errors of statistical inference in the Information Sampling Task

Daniel Bennett Murat Yucel Carsten Murawski

In a previous article (1), we detailed an error of statistical inference in $P(correct)$, one of two outcome metrics for the Information Sampling Task (IST) (2), and showed how this error was likely to lead to biased estimation of reflection impulsivity by standard analysis code. We also provided an alternative formulation of this measure that more accurately reflects the statistical structure of the IST.

In a recent comment, Axelsen, Jepsen and Bak (3) argue against the use of this revised $P(correct)$ measure. They suggest that, under certain assumptions about participants' prior beliefs, a true $P(correct)$ measure converges to the original measure published by Clark and colleagues (2) rather than to the revised measure suggested in our commentary (1). Although we agree that a subjective probability measure such as $P(correct)$ can only be defined meaningfully with respect to a prior belief, we are of the view that the recommendation of Axelsen and colleagues is untenable for several reasons.

First, Axelsen and colleagues' recommendation assumes that participants learn the structure of the IST quickly enough that they can be considered as having perfect knowledge of task parameters. However, in order for this to be the case, participants (i) would have to know that all trials are generated according to an identical underlying binomial rate, (ii) would have to assume that this held across different cost conditions, and (iii) would have to come to this knowledge after having seen only one or two trials. This is psychologically implausible and statistically untenable. In fact, since card colours differ across trials (magenta, cyan, etc.), participants are given an implicit cue that trials *differ* from one another. In one test dataset that we acquired, 7 different colours were present in 14 unique combinations, with no combination occurring more than 3 times. Given such cues, it would be entirely reasonable for participants to infer that the underlying rate was not stable, but was rather dependent on the particular colour in a given trial. Moreover, even if participants knew that a single binomial rate affected all trials, the statistical inference described by the authors could not take place instantaneously. The latter would be necessary, however, for the original $P(correct)$ measure to be a good description of participants' beliefs on all trials. Instead, beliefs could only converge upon this rate slowly, after a number of trials had been observed.

The authors' recommendation that participants be informed that trials are each generated according to the same underlying rate is unlikely to resolve these issues, since different clinical groups are likely to differ

markedly in their ability to use probabilistic information to update beliefs (e.g., (4)). As such, the crucial assumptions of Axelsen and colleagues are unlikely to hold in many participants completing the IST.

More importantly, even if we were to accept the strong assumption that participants perfectly and instantaneously update their beliefs in line with the generative probability embedded in the IST code, the authors' rationale for retaining the original $P(correct)$ measure nevertheless demonstrates an error of statistical inference that recapitulates that of the original publication (2). Axelsen and colleagues assert that under their assumptions, "the Bayesian formulation of $[P(correct)]$ can be shown to be exactly equal to the original equation" (3). This assertion is incorrect. The authors confuse the generative probability of different card colours as implemented in the software (0.5, in this case) with the relative proportion of card colours in any given trial (which may range from 0.2 to 0.8).

To illustrate this, consider that on each trial the number of cards of a given colour N_{colour} is a binomial random variable generated with probability p:

$$\binom{25}{N_{colour}} p^{N_{colour}} (1-p)^{25-N_{colour}}, 5 \leqslant N_{colour} \leqslant 20$$

, where $p = 0.5$ in case of the IST.

Within a trial, therefore, the probability of a given card being a particular colour is equal not to the generative probability p, but to the proportion of cards of that colour within the array:

$$Pr(colour) = \frac{N_{colour}}{25}$$

Crucially, and contrary to the assertion of Axelsen and colleagues, $Pr(colour) \neq p$. For instance, a particular trial may be randomly initialised with 20 blue cards and 5 yellow cards. On such a trial, the probability that any given card will be blue would be equal not to $p = 0.5$, as asserted by Axelsen and colleagues, but to $Pr(colour) = \frac{20}{25} = 0.8$. Furthermore, it is not only true that the measure suggested by Axelsen and colleagues is incorrect in certain cases: in fact, since there is an odd number of cards in each trial, it is in fact the case that $Pr(colour)$ can *never* be equal to 0.5 for the IST. As such, the authors' rationale for retaining the original $P(correct)$ measure is incorrect even under the most generous assumptions about participants' knowledge of the generative properties of the IST.

In general, the prior-dependence of $P(correct)$ means that there is an infinite number of possible $P(correct)$ measures, each corresponding to a different prior belief. Our previous commentary (1) detailed a modified $P(correct)$ measure that respects the statistical structure of the IST while making minimal assumptions about participants' prior beliefs, statistical inferences, and knowledge of parameters encoded in task software. We therefore continue to recommend that researchers use the revised measure. Matlab code to compute this

revised measure is freely available code for researchers (5).

More broadly, we agree with Axelsen and colleagues that the prior-dependence of $P(correct)$ calls into question the idea that the IST is a unitary measure of a single latent cognitive construct. For instance, participant groups who assume that all trials are driven by the same underlying probability, and attempt to learn this probability, are likely to differ systematically in task performance from groups who do not. As such, group differences might reflect differences in probabilistic sophistication or statistical inference as much as reflection impulsivity. Such confounds are insoluble on the basis of choice data from the IST alone, and should be addressed by the development of behavioural tasks and mathematical models that explicitly dissociate the roles of different cognitive processes (see, e.g., (6)).

References

1. Bennett D, Oldham S, Dawson A, Parkes L, Murawski C, Yucel M (2016): Systematic overestimation of reflection impulsivity in the Information Sampling Task. *Biological Psychiatry.*

2. Clark L, Robbins TW, Ersche KD, Sahakian BJ (2006): Reflection impulsivity in current and former substance users. *Biological Psychiatry.* 60: 515–522.

3. Axelsen MC, Jepsen JRM, Bak N (2017): The choice of prior in Bayesian modeling of the Information Sampling Task. *Biological Psychiatry.*

4. Huq S, Garety P, Hemsley D (1988): Probabilistic judgements in deluded and non-deluded subjects. *The Quarterly Journal of Experimental Psychology.* 40: 801–812.

5. Bennett D, Oldham S, Dawson A, Parkes L, Murawski C, Yucel M (2016): https://github.com/danielbrianbennett/ist.

6. Hauser T, Moutoussis M, Iannaccone R, Brem S, Walitza S, Drechsler R *et al.* (2017): Increased decision thresholds enhance information gathering performance in juvenile Obsessive-Compulsive Disorder (OCD). *PLoS Computational Biology.* 13: e1005440.