Sorry to say, but pilots' decisions were not irrational

By Jose Perezgonzalez, 2016, Dec. 16
(Massey University, New Zealand)

https://digest.bps.org.uk/2016/05/16/sorry-to-say-but-your-pilots-decisions-are-likely-just-as-irrational-as-yours-and-mine/comment-page-1/#comment-10868
doi: 10.6084/m9.figshare.4460078

Fradera's Digest (2016, https://digest.bps.org.uk/2016/05/16/sorry-to-say-but-your-pilots-decisions-are-likely-just-as-irrational-as-yours-and-mine/; also *The Psychologist*, 29[7]:511) makes for interesting reading both for aviators and cognitive psychologists alike (actually for all scientists, I would say), although not for the reason which seems the most obvious. Fradera reports on a research article by Walmsley and Gilbey (2016, doi: 10.1002/acp.3225) and, except for the use of the word 'irrational', which the original article never used, and the conclusions, Fradera's Digest seems pretty accurate to the contents commented upon. In a way, thus, whatever praises or criticisms are raised towards or against Fradera's Digest apply equally to the article by Walmsley and Gilbey [1].

The reason why the Digest is interesting is because what is said is quite relevant in principle, but rather misleading in practice. Before entering into details, a disclaimer, paraphrasing Moynihan: "*Everyone is entitled to their own opinions but not to their own facts*". Fradera is entitled to his own opinions and I respect that. But the facts underlying such opinions—i.e., the actual results reported by Walmsley and Gilbey, and I mean the results not their interpretation—do not seem to support them. The portrayal of pilots as biased and irrational are not (fully) warranted neither by the methods nor by the results described in the original article—which may prove a relief for pilots. Instead, it originates in the interpretation of those results by Walmsley and Gilbey, an interpretation chiefly based on inappropriate reliance on a flawed statistical technique—null hypothesis significance testing, or NHST, something which should act as a call of attention to those scientists still using it (for a quick solution see, for example, Perezgonzalez, 2015, doi:10.3389/fpsyg.2015.00223). In a nutshell, Fradera opted to summarise the interpretation of (selected) outputs made by Walmsley and Gilbey instead of re-interpreting those outputs anew within the context of the methodology and the results described in the original article, as I shall argue below.

Fradera correctly summarizes the methods used, including the division of the sample into two groups, and portrays an interpretation of the selected results similar to that given by Walmsley and Gilbey for Study 1, on the 'anchoring effect': "*the pilots tended to rate the atmospheric conditions as better – higher clouds, greater visibility – when they'd been told earlier that the weather forecast was favourable*". Fradera, then, concludes that "*old and possibly irrelevant information was biasing the judgment [pilots] were making with their own eyes*". Such conclusion, however, can only be applicable with some degree of confidence to a group of pilots—those who had been told earlier that the weather forecast was poor—but not to all pilots—i.e., not to the group who had been told earlier that the weather forecast was good. Walmsley and Gilbey certainly wrote that "*pilots reported a higher assessment of cloud height after been exposed to the [good forecast]…compared to when they were exposed to the [poor forecast]*" and "*a higher assessment of visibility after exposure to the [good forecast]…compared to exposure to the [poor forecast]*" (p. 535). However, to me, it is

doubtful what is there to be gained from mere group comparison when a better standard exists: that of actual weather conditions.

Indeed, Fradera fails to realize that the good-forecast group perceived the atmospheric conditions accurately and that it was only the poor-forecast group that seemed affected by the bad forecast. Walmsley and Gilbey's study did not use a control group, so it is not possible to ascertain to which degree pilots exposed to a good forecast were also affected by such exposure. However, and this is important, even if they had been so affected, they still reported an accurate perception of the atmospheric conditions presented to them, so it is not reasonable to conclude that their judgement of such conditions—i.e., their perception— was cognitively biased by "*old and possibly irrelevant*" information.

Furthermore, the relationship between expertise and anchoring effect also needs to be put in the 'interpret with caution' category. On the one hand, there was no reportable (a.k.a., statistically significant) interaction between level of expertise and assessment of cloud base, but only between level of expertise and visibility assessment. On the other hand, it is not very clear where the relevant effect lies in the latter case. Walmsley and Gilbey reported an "*increase from low [poor forecast condition] to high [good forecast condition] anchors being significantly greater for experts than for novices*" (p. 535). Yet again, it is doubtful what is there to be gained from mere group comparison when a better standard exists: that of actual weather conditions.

When compared against actual conditions, expert pilots exposed to good forecasts did report visibility levels to be 2.2 km higher than actual conditions while novice pilots exposed to good forecasts reported visibility levels to be 1.6 km lower; that is, expert pilots overestimated visibility by about the same amount than novice pilots underestimated it. On the other hand, both groups were affected by poor forecasts in about the same degree, reporting between 4 km and 5 km lower visibility than actual conditions showed, respectively.

Therefore, it is not correct to conclude that "*[expert] pilots were especially prone to being influenced by the earlier weather forecast*". (Incidentally, above estimates would largely be dismissed as practically irrelevant in terms of degree of inaccuracy in real flying, as it is quite difficult to appreciate a difference of two, even five, kilometres when horizontal visibility is as large as 16 km, the actual conditions presented to pilots; indeed, pilots would rather ask for a weather update than rely on their own perception of visibility—verbal comment provided by a pilot, Nov. 2016).

Regarding Study 2, on 'confirmation bias', Fradera correctly summarizes the methods used, and portrays an interpretation of the selected results similar to that given by Walmsley and Gilbey. The study, however, is doubtful in terms of methodology, as the scenarios seem not to have been well designed for purpose (see Walmsley, 2016, http://mro.massey.ac.nz/handle/10179/8275). The disconfirmatory items were phrased rather ambiguously while the confirmatory items were phrased with certainty, and all scenarios prompted quick landing rather than offer some flexibility between landing and diverting, all of which may partly account for pilots discounting the disconfirming evidence. As García-Pérez and Alcalá-Quintana put it elsewhere (2016, doi:10.3389/fpsyg.2016.01042), "*Inferring states of knowledge from item responses requires items worded unambiguously and whose content relates exclusively to the piece of knowledge being assessed and not to something else*" (p. 5). Because the main failures reside in the methods—something easier to

spot in hindsight but difficult to correct once spotted—rather than in the interpretation of results, I shall comment on these particular results no further.

Regarding Study 3, on 'outcome bias', Fradera correctly summarizes the methods used (save for the sample being divided into three groups, including a control group), and portrays an interpretation of the selected results similar to that given by Walmsley and Gilbey: "*participants tended to rate their decision making much more harshly when the [third-party] flight ended in disaster [the negative outcome] than when all went well [the positive outcome]*". Such interpretation may be technically correct, yet again it is doubtful what is there to be gained from mere group comparison when a better standard exists: that of responses against the actual scale used. Indeed, when queried whether they would fly safer under similar conditions, all groups responded with the equivalent of a "Don't know / Can't decide", although the positive outcome group were closer to a "Somewhat agree".

What is more interesting (and here I compare responses between items but within groups) is that all groups gave a somewhat optimistic assessment compared to other questions asked, irrespective of condition; that is, they responded with the equivalent of being capable of flying somewhat better than a pilot who landed successfully, somewhat better than a pilot who crashed, and somewhat better than any pilot, in general. In a nutshell, their assessment was not dependent on outcome but on risk perception.

The thing is, Walmsley and Gilbey had also asked each group to rate the quality of decision making by third-party pilots—"Don't know", according to the positive outcome and control groups; "Somewhat poor", according to the negative outcome group—and to rate the riskiness of the third-party pilots' behaviour—"Don't know", according to the positive outcome and control groups; "Somewhat risky", according to the negative outcome group. Therefore, participants rated third-party decision making and behaviour more poorly when the third-party flight ended in disaster, which may have tainted their own perception of weather conditions (something which would be coherent with the first study's results, thus a more plausible hypothesis). The simpler, more parsimonious explanation for how participants rated their own decision making is that risk perception, not outcome, informed their responses. Said otherwise, that any effect of outcome bias on pilots' self-assessment was mediated by risk perception.

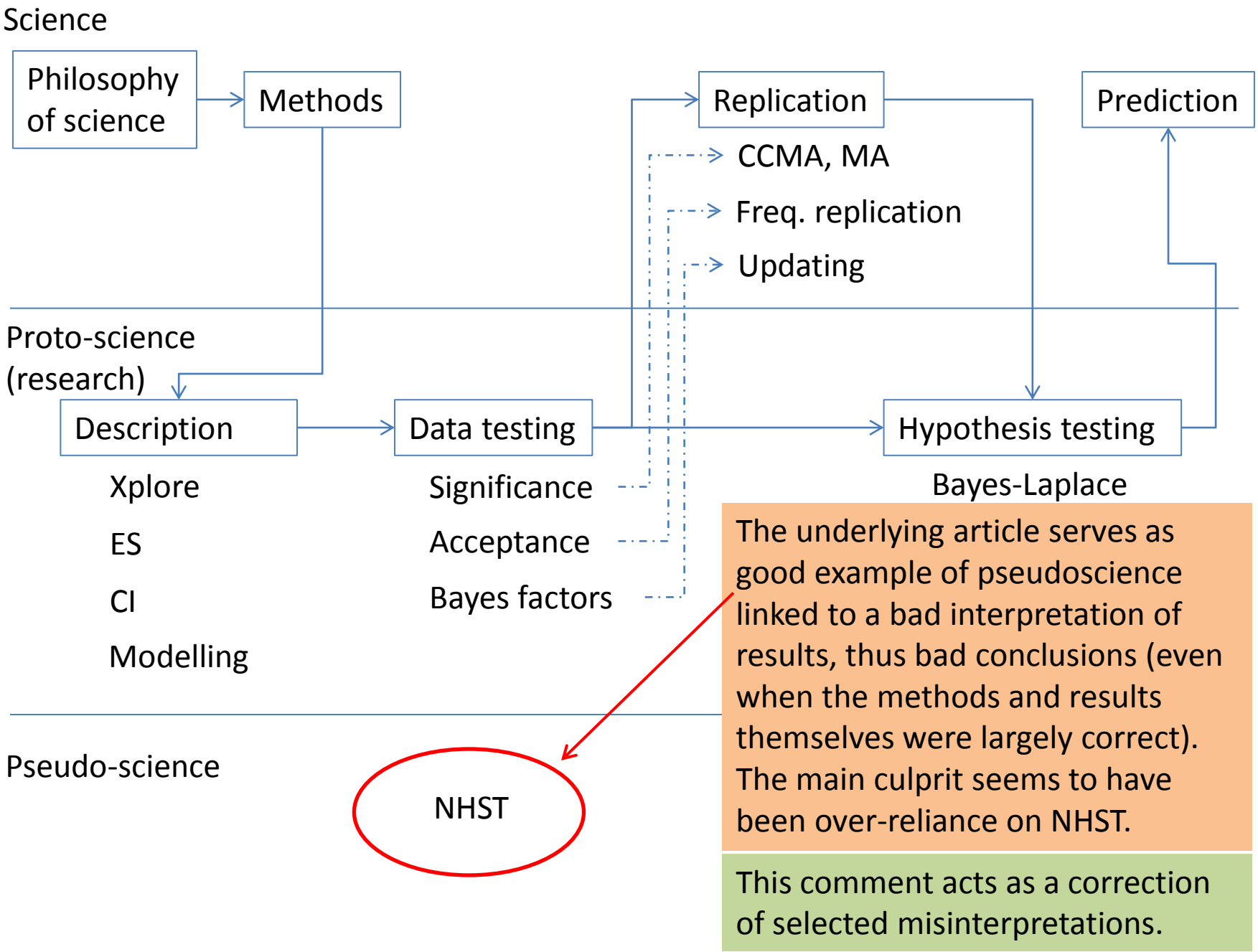
Furthermore, the relationship between expertise and outcome bias also needs to be interpreted with caution. There was no reportable (a.k.a., statistically significant) interaction between level of expertise and third-party ratings, but only between level of expertise and self-assessment; yet, as discussed above, self-assessment does not seem to have been affected by outcome bias in a direct manner. If we draw confidence intervals for each group and condition in the self-assessment variable, we would observe that the main difference between novices and experts lies in the positive outcome condition, with expert pilots reporting somewhat more self-confidence than novices (they also did so for the negative outcome condition but the differences seem to be barely statistically nonsignificant). Thus, although expert pilots may have shown differentially more confidence than novices, this is neither unexpected (it could be partly accounted for by their level of expertise) nor really important: expert pilots on the positive outcome condition were somewhat confident—they arguably had no opinion regarding decision making and risk propensity of the third pilots who landed safely—while expert pilots on the negative outcome condition showed neither confidence nor lack of confidence—they arguably perceived the decision making and risk propensity of the third pilots who crashed as somewhat poor and somewhat risky, respectively. In any case, it

is not really clear what could have been the better decision in any of Walmsley's scenarios, so it is not possible to ascertain whether "*[expert pilots] did [not]…make better decisions than other pilots*", as Fradera concludes.

In brief, Fradera's Digest suffers from bias against the actual results described in Walmsley and Gilbey's article. Two biases seem apparent: on the one hand, an anchoring effect on statistical significance and group comparisons in preference to the actual instruments used, when interpreting results. On the other hand, a confirmation bias, where Fradera may have sought to confirm his own expectations —or, perhaps, those of Walmsley and Gilbey's—of an effect of cognitive biases on pilots'—especially on expert pilots'— perception and decision-making despite evidence to the contrary. As such, the Digest makes no justice neither to the population of pilots which was the target of the original study nor to the larger literature on cognitive biases and rational thinking.

---

[1] Walmsley and Gilbey's article (2016) was itself an adaptation of Walmsley's PhD thesis (2016, http://mro.massey.ac.nz/handle/10179/8275), even if not so cited in their article. Authorship conventions dictate that authors are solely responsible for the contents they publish, irrespective of sources or influences. This said, I was co-supervisor of Walmsley's thesis and, therefore, may have had some influence—or lack of, thereof—on Walmsley's work. As far as I may bear some responsibility as a 'ghost' author of Walmsley's thesis, the criticisms laid here may equally apply to me retrospectively, and shall serve as a partial correction of the research literature cited.


Science
Philosophy of science
Methods
Replication
Prediction
CCMA, MA
Freq. replication
Updating
Proto-science (research)
Description
Data testing
Hypothesis testing
Xplore
ES
CI
Modelling
Significance
Acceptance
Bayes factors
Bayes-Laplace
Pseudo-science
NHST
The underlying article serves as good example of pseudoscience linked to a bad interpretation of results, thus bad conclusions (even when the methods and results themselves were largely correct). The main culprit seems to have been over-reliance on NHST.
This comment acts as a correction of selected misinterpretations.