Qeios

Peer Review

# Review of: "The Monty Hall Problem: Does Performance Improve After the Acquisition of Some Probability Knowledge?"

**Ulrich Hoffrage**[1]

1. Faculty of Business and Economics, University of Lausanne, Switzerland

This manuscript reports the performance of some undergraduate students who were asked to provide a probability judgment and a decision for the Monty Hall Problem, a task that involves reasoning with probabilities. Performance was measured at two points in time: before students took a class on Biostatistics, and afterwards. It turned out that the percentage of correct responses was virtually the same before and after this class.

I have three major concerns. One is about the dependent variable, or more specifically, how the rationality of participants' responses has been assessed. I think the paper has weaknesses when it comes to the normative benchmark. Second, I take issue with the dependent variables, that is, participants' responses. Participants were first asked to provide an assessment of the chances of winning when opting for either switching or staying, and second, they were asked to indicate what they would do. But the response options they were given did not match the questions they were asked. And third, I think the intervention (the course in which students were supposed to acquire some probability knowledge, as the author calls it) leaves room for improvement. After having elaborated on these issues in more detail, I add some minor issues.

**First major issue: normative benchmark.**

The problem description as cited from Parade magazine in the introduction, "Suppose you're on a game show --- choice of doors?" is, in my opinion, underspecified. It lacks a crucial piece of information, namely whether Monty Hall always has to open – and will open – one door in this situation, or whether his choice set also includes the possibility of opening no door at all. If he always has to open one door,



doi.org/10.32388/K55VRJ

then Marilyn vos Savant is, in my view, right: in the long run, it is better to switch. This has been convincingly explained by, for instance, Krauss and Wang (2003), who simply created all possible combinations between three variables: (1) behind which door the car is, (2) which door the candidate initially selects, and (3) which door Monty Hall opens. Looking at the arrangement that results from cross-tabulating the levels of the three variables, one can easily count the number of cases when it is better to switch and when it is better to stay. Note that one would need to make some assumptions, namely that (1) the car is with equal probability behind the left, the middle, and the right door, and (2) that the candidate initially selects the three doors with equal probability, and (3) that this selection is independent of where the car actually is. This set of assumptions results in a probability of 1/3 for initially wanting to open the door with the car behind it. (As a side note: upon closer inspection, it becomes clear that either the car's position being a random variable plus independence of car position and candidate's initial selection, OR the candidates' initial selection being a random variable plus independence of car position and candidate's initial selection, is sufficient for the probability of initially picking the door with the car to be 1/3.)

To reiterate, such a cross-tabulation of car position, the candidate's initial door selection, and Monty Hall's door opening shows that switching is, in the long run, better in 2/3 of the cases. But note that this is only true if all possible cases and their respective probabilities are considered. If this complete cross-tabulation is considered, then this can be described as a situation under risk, with one particular realization of this game being a random draw from a lottery with known outcomes and known probabilities.

If, however, Monty Hall has the option to NOT open any door at all, then one cannot conclude that it is better to switch in 2/3 of the cases. By opening a door only in a subset of the 9 possibilities of where the car is and which door the candidate initially wants to open, Monty Hall can easily manipulate this probability (with which switching would be better for the candidate after Monty Hall had opened a door). For instance, Monty Hall could have it as a policy to only open a door when it would be better for the candidate to switch (this would be the case when the candidate initially picks a door with a goat behind it). Alternatively, Monty Hall could have it as a policy to only open a door when it would be better for the candidate to stay (this would be the case when the candidate initially picks the door with the car behind it). If Monty Hall is not forced to always open a door, but can, at his own discretion, at any given new run of the show decide whether to open any door or to open none of them, then this would be a situation under uncertainty – for which there is no normative correct answer. And where even Monty Hall's

behavior in previous shows would be of limited use, as any inferences based on this previous behavior would hinge on assumptions about how his policy in the past is related to his policy for the present game.

I know that the distinction between risk and uncertainty (see also Mousavi & Gigerenzer, 2014) is usually not made for the Monty Hall problem, and one can hence also close an eye here, for the present manuscript. But strictly speaking, it is problematic to use an underspecified problem description, treat it as a situation under risk (even though it could also be a situation under uncertainty), and use a norm (that is only valid if this was a situation under risk) to evaluate the rationality of participants' responses. The basic question here is whether the sloppiness in the present manuscript should be ignored given that this sloppiness is widespread. The good news here is that this could easily be repaired: Just add that the following analysis is based on the understanding that Monty Hall always has to open one of the two doors that the candidate did not initially pick.

**Second major issue: judgment versus choice (the two performance measures, and hence dependent variables, before and after the acquisition of some probability knowledge).**

I find it very smart and useful to ask two questions. The first calls for a judgment, or assessment: What do you think has more chances? The second calls for a decision: What would you do? One may think that the answer to the first question would be a perfect predictor for the answer to the second question, but, well, at the end of the day, this is an empirical question, and everyone who has already made some experience with human participants has presumably already encountered some surprises. So, fine to ask these two questions. BUT: the response options for the first question should be assessments, such as "I think chances of winning are higher if I keep the first door." This would, in fact, be a probability judgment. But the option "Keep the first door" is not a judgment; it is a decision. It is a possible answer to the question "What would you do?" And in fact, the response options among which participants can choose are identical for the judgment question and the decision question. In my view, this is a fundamental flaw in the methodology, and the discussion of the comparison/consistency between participants' answers to the first versus the second question below suffers from this flaw, and this flaw makes any findings hard to interpret.

Section 3, results. We read "27 believed that keeping the first door had greater chance of winning." Hm, maybe. But who knows? What we do know is that 27 respondents checked "keep the first door." But as I just explained, this option is an answer to the question "what would you do?" and NOT "What do you believe?" One may speculate that someone who believes it is better to switch may still decide to stay, maybe because of different amounts of regret for the two errors one could make here (switching where

staying would be better and staying where switching would be better). The point is that the sentence "27 believed..." is not trustworthy, simply as participants did not have a chance to indicate what they believed. They could only indicate what they would DO.

**Third major issue: Transfer of "probability knowledge" (independent variable).**

The independent variable was, and here I cite from the title, "Acquisition of Some Probability Knowledge," with the levels "before" and "after" this acquisition. The present study is by far not the first that evaluates the effectiveness of some didactical intervention, in general, and teaching some statistics (here, biostatistics), in particular. Result in the present study: virtually no improvement in performance after the course. This is perfectly fine and methodologically sound (maybe except for the fact that a between-subject design would have been stronger as it would not have to deal with the problem of participants' memory of their own answers before the course, possibly combined with the desire to answer consistently). One could criticize the study on this ground and ask for a replication with a between-subject design (which would then have, of course, other problems). Maybe a combination of the two designs would be ideal.

But there is still another possibility to extend this study, and this would be not just to implement some "Acquisition of Some Probability Knowledge" but to design a tutorial that would be expected to be more effective based on theoretical grounds. I already mentioned the study of Krauss and Wang (2003). These authors did not use Bayes' theorem to compute the probability of getting the car when switching (versus when staying) conditioned on Monty Hall's action, but they chose a visual display that allowed readers to count cases. More importantly, they used the same display to allow their participants to gain some insight and to see why switching is the better option. What these authors did came close to one condition that Sedlmeier and Gigerenzer (2001) and Kurzenhäuser and Hoffrage (2002) implemented in their tutorial programs that were designed to improve performance in Bayesian inference problems. Both G&S and K&H compared two different training programs: one in which the use of Bayes' rule was explained and trained (rule training), and one in which participants learned how to translate the information that is typically provided in a Bayesian inference problem (prior probability, hit rate, and false-alarm rate) into what Gigerenzer and Hoffrage (1995) called frequency formats and Hoffrage and Gigerenzer (1998) later called natural frequencies. Result: performance increase was higher and lasted longer for representation training (where participants were taught how to represent probabilities in terms of natural frequencies) compared to rule training. By now, it should be clear that the way Krauss and Wang (2003) explained the correct solution in the Monty Hall problem corresponds to what G&S and K&H implemented as

representation training. The author of the manuscript under review did not provide any details, but given my knowledge of how statistics is typically taught at universities, I suspect that the Biostatistics class that formed the intervention in this study on the usefulness of "the acquisition of some probability knowledge" came close to what G&S and K&H had as their control condition: namely rule training. This discussion could inspire a new study in which the effectiveness of two different ways of teaching statistics is compared. My (admittedly bold) prediction would be that a statistics course that consistently describes distributions as probability distributions would not lead to any performance improvement in the Monty Hall Problem (basically the finding of the present study), whereas a statistics course that explains the same distributions as distributions of cases (or frequencies) would help students to improve their performance in the Monty Hall Problem.

**Minor, copy-editing stuff:**

Abstract, line 2: "in the field"  Which field?

"study how a group of people" This could be misread as a study in group decision making, implying that those people have to agree on one assessment that they then provide as a group. But in fact, these were individuals who provided their assessments alone. I'd simply say "how people." (This would leave no doubt that n>1 anyway, and hence "group" would not be necessary anymore.)

The formulation of the three objectives leaves room for improvement. "Some probability knowledge" is very unspecific. Why "potential" decision-making? What is meant by the consistency between an assessment (of what?) and (decision-making) processes?

Section 2.1. "not included in the responses" – I guess what is meant here is "not included in the analyses reported below." Next sentence: "The survey included 47 respondents." No, a survey does not include respondents; it rather includes questions. And then it is given to some respondents. And some of them return it. And some of those are included in the analysis. Hence, the sentence should read "The analysis used the data from 47 respondents" (or something like this).

First sentence below Table 2 (on page 7 of the manuscript): "To compare how each group responded in the same opportunity…" is not clear. What is meant by opportunity?

Section 4, Discussion, second sentence "…47 subjects answered correctly that it would be more convenient to change…" First, they have not been asked about convenience – this word did not exist in the questionnaire. And second, it was good that it was not in the questionnaire, as the issue here is not convenience but chances of winning.

Much of the discussion is not a discussion of the results but a repetition of the results.

3rd paragraph of discussion: so far, you often said “first occasion” or “second occasion.” Now you say “second instance.” In my view, neither of these terms is ideal. Maybe pre-course and post-course – and make sure that this refers to the point in time before and after they took the Biostatistics course?

References:

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychological review*, *102*(4), 684.

Hoffrage, U., & Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Academic medicine*, *73*(5), 538-40.

Krauss, S., & Wang, X. T. (2003). The psychology of the Monty Hall problem: discovering psychological mechanisms for solving a tenacious brain teaser. *Journal of Experimental Psychology: General*, *132*(1), 3.

Kurzenhäuser, S., & Hoffrage, U. (2002). Teaching Bayesian reasoning: an evaluation of a classroom tutorial for medical students. *Medical teacher*, *24*(5), 516-521.

Mousavi, S., & Gigerenzer, G. (2014). Risk, uncertainty, and heuristics. *Journal of Business Research*, *67*(8), 1671-1678.

Sedlmeier, P., & Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of experimental psychology: general*, *130*(3), 380.

My overall score is 2 (without my second major point, I would have given a 3).

## Declarations

**Potential competing interests:** No potential competing interests to declare.