Warsaw, January 3, 2025

Michał Krawczyk
Wydział Nauk Ekonomicznych UW
ul. Długa 44/50, 00-241 Warszawa

Review of Doctoral Dissertation "Decision Making and Numeracy: The Role of Context in Adaptive Strategy Selection" by Supratik Mondal

Overall assessment

The dissertation consists of three interrelated published papers. One of them, published in a Polish journal, is a single-authored replication of an earlier study. The other two are original studies, co-authored by the advisor and published in good international journals.

To my reading, it is a mixed bag. The questions posed are interesting, given the existing literature. The studies are well designed, reasonably powered, pre-registered. The econometric analysis is state-of-the-art and clearly reported. Then again, several times I found myself scratching my head. Below I list some issues I find problematic (although my overall assessment is positive).

Construction of the dissertation

For me, the crux of the dissertation is the papers. Because they are published, they are (or: should be) self-contained. Yet, they are preceded by the abstract, "overview" and "summaries" here. As a result, there are a lot of unnecessary repetitions. Moreover, in the table of contents, the actual papers can easily be missed – they are merely subsections of Chapter 10, following the bibliography (?).

Abstract

The abstract is rather long and not too exciting as, among others, it includes definitions (later to be repeated in two other places).

Introduction and Theoretical Overview

This part meanders a bit. It seems to me that the links between numeracy, chess and abstract art are tenuous at best. Likewise, it is hard to say what is the link between the paragraphs on p. 14. It is very surprising to learn that "maximisation of expected utility" and "feelings" are examples of "actors". (p. 11). It is not clear what does a one-point increment in numeracy score represent (p. 11).

In what remains I discuss key aspects of the three papers.

Theoretical framework

Papers 1 and 2 make a key distinction between "high payoff" and "low payoff" conditions and choice situations. Now, a choice between 50% to get 50 euro and 51% to get 49 euro could intuitively be called "high payoff" (because the amounts are higher than typically observed in short experiments) whereas here it would be "(very) low payoff" because the difference in expected values is miniscule. When reading Paper 1, I got an impression it would have been better to label "low payoff choices" as "low stakes" choices or "small valuation difference choices" or, quite simply, "hard choices" and analogously for the "high payoff choices". Currently, the reader is puzzled to discover (if they every do, because it is hidden in the appendix), that, on average, participants took less time to choose in "high payoff" choices. I would tend to think much longer when payoffs are high (say, choosing between jobs) than when they are low (say, choosing between snacks). It would not have been surprising that choices are faster when *differences in valuations* are high, i.e. one option is much better than the other and so the choice is easy (one job being much better than the other) than when it is hard (both options have their pros). This is logically orthogonal to the value of payoffs themselves – the choice between snacks can also be hard or easy. There is voluminous literature e.g. using the Drift Diffusion Models (DDM), on why choices are slower when the decision maker is closer to indifference. In fact, there are papers that claim precisely that we typically devote too much time to cases in which both options are nearly equally good (e.g. Oud, B., Krajbich, I., Miller, K., Cheong, J. H., Botvinick, M., & Fehr, E. (2016). Irrational time allocation in decision-making. Proceedings of the Royal Society B: Biological Sciences, 283(1822), 20151439.); it is a natural hypothesis in this context that highly numerate individuals are better at recognising this trap and, instead, devoting their attention to cases in which it really matters. It would be interesting to model this in the context of DDM (or maybe it has already been done?). Overall, it seems that suboptimal choice of labels not only confuses the reader but also misguides the Candidate in his literature search.

In Paper 2, the Authors explicitly consider two dimensions: absolute EV difference (AED) and relative EV difference (RED), controlling each of them. That is a clear methodological improvement (which actually shows that the conclusion from Paper 1 was premature). However, the Candidate continues to call low RED "low payoff" and high RED "high payoff". In this design, this becomes even more confusing because to keep the AED the same, the "low-payoff" choices must actually involve much larger EVs – the EVs must be high for given AED to be relatively small. I am also confused by the last paragraph of section 2.2 of the second paper (btw, some page numbering, consistent throughout the dissertation, would make the reviewer's life easier). The Authors seem to suggest that AED and difficulty are different concepts whereas they admit in footnote 1 that – at least in their sample – it is precisely low AED that makes some choices more difficult.

The key notion of "adaptive choices" is presented in a somewhat confusing, contradictory way. We are told "a group of decisions can be considered "adaptive" when, as a collective, it outperforms the existing normative prediction or matches it without utilizing a similar amount of resources." With decisions under risk, involving low rewards, maximisation of EV is the strongest normative prediction (and so the Candidate is right to use it throughout the dissertation) and, for the same reason, to "outperform" would mean to yield more in expectation. To outperform EV maximisation would thus mean to earn, in expectation, more than is possible, a feat of magic thus, not "adaptability". By contrast, in 10.2 (again, page numbers missing) we are told that "participants' choices would be considered adaptive if participants, on average, followed the EV consistent strategy in the high-payoff condition and changed their strategy to EV inconsistent choices in the low-payoff condition". This is quite imprecise but actually seems to imply that an adaptive decision maker must make LESS in expectation than a EV maximiser (because they are EV-inconsistent in some choices). This becomes

clearer when the Candidate discusses time restrictions. The benchmark when defining "adaptability" is thus not a theoretical, instant EV maximiser but an actual slow (and imperfect?) EV maximiser. This should perhaps be explicitly said when adaptability is defined. But this makes the definition even more blurry because it first requires an operational definition of a (real-life, slow, imperfect) EV maximiser for reference.

At times, the Candidate seems to present results as if they were general whereas in fact, they (seem to me to) depend dramatically on the stimuli used in given study. For example, in the third paper, p. 3, he says, "For instance, after drawing just one sample, Hertwig and Pleskac (2008) demonstrated that the probability of choosing an option with a higher EV was approximately 60%. Drawing an additional five samples increased accuracy by 18%, while further increments in the number of draws—from five to 10 samples and from 10 to 20 samples—improved accuracy by 6% and 4%, respectively." – this finding seems to be very specific to the (arguably: arbitrary) chose of parameters of the gambles used in that study.

Empirical methods

The Candidate uses the Berlin Numeracy Test extensively, providing a "representative" item: "Imagine we are throwing a five-sided die 50 times. On average, out of these 50 throws, how many times would this five-sided die show an odd number ?" – this question tells me more about the mathematical skills of the designer of the test than those of the takers. I do not think it is even possible to construct a fair five-sided die, because, as high-school math tells us, there is no five-sided platonic solid (convex regular polyhedron). You can have a fair die with 4, 6, 8, 10, 12 or 20 faces, but not 5. Why label willingness to make mathematically nonsensical assumptions "numeracy"? While a thousand citations that this method received represent some justification for yet another scholar (here: the Candidate) using it, can't one find a less confusing test?

In fact, that is not the only important assumption that is omitted here: the Candidate seems to assume that the sides are numbered 1 through 5. In fact, not all (actually existing) n-sided dice have sides numbered 1 through n. For example, 10-sided dice are often numbered 00 through 90. Here, the source paper (Cokely et al) actually helpfully add that numbers "1, 3 or 5" are meant, it is the Candidate who dropped it, I do not know why.

One could think these are trivial issues because most participants will easily make the default assumption that the five sides are equally probably are numbered 1-5. Yet for some this may sound like a three-sided coin – a good reason to become distracted, confused, mistrustful or annoyed. In experiments with human subjects (especially online when they cannot ask a clarifying question and are constantly tempted to simultaneously play Fortnite or check their Facebook account) details matter, so one cannot be too careful.

In Papers 1 and 2, it would seem natural to suppose that reaction times mediate the effect of BNT-payoff interaction on choice: in the high-payoff condition, high BNT participants choose to think relatively hard and that is what makes them choose the option with higher EV. Thus, when RT is included in the model, the direct effect is reduced. Why wasn't it explicitly posed and tested?

"In future studies, I intend to conduct further experiments with between-participant design to establish the causal relationship between adaptive behavior and numeracy" – I have no idea how using between-participant design per se would allow identifying such a causal effect.

I don't know how to test normality between groups (table 8 of 10.1.1). Perhaps it is just the caption of the table that confuses me.

My understanding is that the main difference between experiments 1 and 2 of Paper 2 is that the former (later) uses within-(between-)subject design. The pros and cons of either design are well known and it would seem natural to explain how they apply in this particular domain. By contrast, the Authors' motivation seems to be the "non-inferiority test". I hand not been familiar with this approach, possibly because vast majority of its applications seem to be in medical science. It seems somewhat natural to make sure that a new drug or procedure is, in the worst case, not substantially (or: unacceptably) inferior to an already existing treatment. I suspect that at least in some medical applications there is reasonable consensus on what is unacceptably worse, thereby providing some support for this approach. Why would we here think in terms of non-inferiority of one experimental condition over another? And are the "acceptable difference" levels here not completely arbitrary? In the end, this approach seems to add little to the interpretation of the results. By contrast, it adds to the confusion of the reader when (s)he encounters phrases such as "mean EV consistency in the low-payoff condition is not non-inferior to the mean EV consistency in the high payoff condition". It seems a very complicated way of saying that, not surprisingly, the share of subjects choosing the better of two options is higher when the difference between the options is large.

I very much appreciate the use of time pressure. I am wondering however it was an optimal choice that participants did not know how many decisions they can maximally make. They may have formulated some (unobservable) beliefs in this regard, with those guessing it is a very large number tending to go faster (sample less). This could have added noise or even, arguably, bias (if guesses were correlated with variables of interest, such as numeracy).

A very problematic aspect of empirical analysis in this dissertation is the use of pre-registrations. Of course, it is great that the Candidate decided to use them in the first place. Sadly, the delivery leaves much to be desired.

First, the key point of pre-registration is to limit authors' leeway in selection of models and hypotheses. The idea is that they rigorously stick to the pre-registered data analysis plan and, should they ever deviate from it or expand on it, they explicitly warn the reader and explain why they are doing so. Here, the Candidate does not indicate which analyses had been pre-registered and which not. The reader should then assume that he sticks to the analysis plan meticulously, with no changes or additions. Yet, as far as I can tell, the data analysis plan for Paper 2 does not specify the "acceptable differences" in the non-inferiority test and does not mention Johnson-Neyman intervals at all. Thus, important decisions concerning data analysis seem to have been made after the data was collected. Still, the reader is led to believe otherwise.

Second, in Paper 3, we are told that "The analysis of qualitative data was preregistered (Mondal & Traczyk, 2023b)." In the list of references, this entry involves the link https://doi.org/ 10.17605/OSF.IO/U59FC which seems to lead nowhere (doi foundation error 404). How can a pre-registration link expire within a few months since the publication? It also seems to say that study 2 of paper 3 was not pre-registered although the Authors do not seem to mention that in the paper.

Overall, pre-registered analysis plan which is a) partial b) seemingly only partially available to the reader c) modified in the actual study without warning is worse than no pre-registration at all – it gives a less inquisitive reader a completely false picture.

When discussing different potential benchmarks in Paper 2, the authors say they "did not find any statistical difference in the prediction between the [models]", referring the reader to figure 1 in the appendix which shows the share of choice problems in which given pair of models yield identical predictions, with some sort of whiskers (we are not told what they represent but one would suspect some sort of confidence interval). I am also not sure what "[no] statistical difference" is supposed to mean here. I suspect this: if we define a binary variable consistency_{XY} taking the value of 1 if model X and model Y yield the same prediction for given choice problem and the value of 0 otherwise, then the mean of consistency_XY is not statistically different from 1. There is more than one problem with this. First, another way to say it is that in the sample, predictions of models X and Y always coincide, which is, for some pairs, apparently *not* true. Second, and perhaps more importantly, I believe any notion of statistical significance and confidence intervals is meaningless here. As far as I can tell, the 72 choice problems are not a random sample from any well-defined population. Statistical significance and confidence intervals describe our knowledge of the parent population or parameters of some probability distribution based on the sample drawn. They are thus not relevant for the case at hand.

In experiment 2 of Paper 2, points earned in the two conditions have different values. Yet the authors do not take it into account, as if they assumed their subjects cared about points per se, rather than about actual money.

Moreover, as the Authors explain, "in the Low-payoff condition, participants were told that for every 1000 points, they would receive an additional 0.80 GBP on top of the flat fee. However, in the High-payoff condition, participants were told that for every 500 points, they would receive an additional 0.65 GBP on top of the flat fee.". it is very clear from existing literature that if you want to facilitate comparisons between ratios (e.g., probabilities), you want to keep denominators equal. For example, it is much easier to recognise that 16/100 is larger than 15/100 than to realise that 8/50 is larger than 15/100. I understand that monetary values of points were not explicitly compared across conditions here, but still it is not clear why they were presented differently.

Worse still, comparing 11.2 and 11.3 in 10.2.1, we see that participants were told directly (…for every 500 points…) and by means of an example (… so, if you collect 1500 points…) how much a point is worth in the high-payoff condition. By contrast, in the low-payoff condition – only an example is provided. This looks like a serious mistake, making the two conditions different in an additional, unplanned way.

Further, it is not clear if, say, in the high-payoff condition, earning 400 points meant a) making 0 (because 400<500) or b) it meant making 400/500 times 0.65. Similarly, it is not clear if the 1500th point, turning 1499 to 1500, represent an increment of a) 0.65 or b) just 0.65/500 etc. The examples do not give any help in this regard. To the extent that some participants favoured interpretation a), we have two potential problems. The first problem is that depending on the number of points they (believe to) have accumulated, participants may have incentives to be risk-seeking or risk-averse; for example, if I believe I have about 300 points so far and I only have a few gambles left, I may be risk-seeking to get any chance to reach 500 points. The second problem is that the different denominators (/500 points in the high-payoff condition and /1000 points in the low-payoff condition) imply yet another unplanned strategic difference between the conditions (again, with the additional caveat that participants in the low-payoff conditions, have not even been explicitly told what their numerator was, so who knows what they were thinking).

Overall, one gets an impression that not only the Authors made a surprising assumption that the participants did not care about monetary value of the points but also the Authors themselves did not care about it too much, which resulted in some sloppiness in this regard.

It can also be noted that in instructions used in Paper 2, participants are told they will face "gambles" with two options. To the best of my judgment calling a choice between two gambles a "gamble" is potentially very confusing. (Similarly, in Paper 3, participants are told they will face "decision problems in the form of lotteries"; and indeed, in the course of instructions, they are encouraged to "choose" from a set of one (sic) lottery). Again, these are small details but, when preparing experimental stimuli, details matter and cannot be fixed in a revision (unless, of course, a new study is run).

I do not know what the value added is of the simulation study of 10.3. The Authors write "we found that the "blue" decision maker would face 30 choice problems, given that they are following a more energy-intensive EV maximization strategy 75% of the time. However, the "orange" decision maker would earn more reward, compared to the "blue" decision maker, by facing 10 (i.e., 40 choice problems in total) problems more by randomly choosing between options (i.e., EV consistency of 50%)." – If I am not mistaken, the first "finding" was just an assumption and the second could easily be calculated based on the assumptions (parameters of the gambles), without simulating 60,000 observations.

It is unclear to me why the Authors are interested in the "switching ratio" – why is this improtant and why do we expect more numerate people to switch less?

Clarity

I found quite a few statements in the dissertation difficult to decipher. Here are some of the more important cases.

"More numerate individuals were better at maximizing expected reward following the frame of reference of absolute difference (as captured by the absolute expected value difference distribution) and not the relative difference in reward (as captured by both payoff conditions)." (p. 2) is unclear to me.
"elaborate and in-depth processing of information, utilizing heuristics (Cokely & Kelley, 2009)" (p. 12) sounds rather surprising. "in-depth" and "heuristics" rarely go hand-in-hand.

"While heuristics can lead to biases and errors, more often than not, they often approximate optimal choices sufficiently well and sometimes even outperform them." (p. 13) By definition, it's impossible to outperform an optimal choice.

Priority heuristic does not seem to be precisely defined when it is first used. How can I know if given difference "is significant"?

10.1

What does it mean that "participants … on average maximized EV" (p. 5 of 10.1; the rest of that sentence is not clear to me either).

CCA is only sketched on p. 15 of 10.1. In Table 1, why does the reader have to guess what payoff=1 or choice=1 mean? I find Table 2 very confusing. Was indeed the test run for high BNT different from that run for low BNT as the table title suggests? Why? And why does the note suggest the opposite?

How are "long/medium/short" RT defined? (p. 16 of 10.1)

In 10.1, why are result of models 1 and 2 not presented together?

Figure 4 is a complete mess. Was it badly OCRed from a pdf?

Supplementary materials of 10.1.1 often left me puzzled. What does it mean exactly if "CPT/EV choices…" equals, say .2?

10.2
"the presence of the two conditions [low-payoff and high-payoff I suppose] together" – how could they be present "together"?

In 3.1.2. the Authors say the "followed the same principle" and "control[led] as many factors as possible (i.e. variance, AED)" immediately after a paragraph criticising (their own) previous papers which had not controlled for AED.

My understanding is that "restricting them to 100 and 50, respectively" mean that AED cannot be larger 50 points. How does this restriction cause (as "as a consequence" would imply) that "participants will earn … less if they [make] EV inconsistent choices"?

If the authors counterbalanced option placement so that "option A is not option A for each participant" (sic), what is even the point of checking if option A was on any dimension significantly different from option B?

Panel B of Figure 1: how is that possible that "possible earning" associated with EV-consistent choices in the low payoff condition is lower than that in EV-*in*consistent choices in the same condition?

I do not understand Figure 3. These straight lines cannot directly represent empirical data. I suppose they are best linear fits but this is not explained.

I also have great difficulty understanding the discussion of the hypothesis of diminishing value sensitivity. I am told that "as the psychological interpretations of the law of diminishing returns dictate, preferences become more inconsistent when the difference in value is relatively similar." Similar to what? More importantly, what does the simple fact that probability of choosing the better option goes down as options become more similar (I am assuming this is meant here) have to do with the law of diminishing returns (also known as the law of diminishing marginal productivity) or its, undisclosed, "psychological interpretation"? Assuming almost any stochastic element in the model makes mistakes more likely when difference in utility is smaller. And even if by "diminishing returns" the Candidate means diminishing marginal utility (of money?), it is still unclear to me how it is necessary here. The Candidate then says that "when the absolute difference between two offers is 50$ and the relative difference is small (i.e., option A is 400$ and option B is 450$), preference for the higher value item is relatively less consistent and in accordance with the hypothesis of diminishing value sensitivity". Is "diminishing value sensitivity" the same thing as "diminishing returns"? Is preference for the higher value "in accordance"? Or is the fact that it is "relatively less consistent" "in accordance"? Are we talking about expected values here? Because I am quite sure preference for 450$ over 400$ (no other differences between the options) will be very "consistent". Overall, I find this paragraph very confusing.

I also do not know why aspiration levels are mentioned (in the same section). We do not seem to learn anything about them from the experiment.

10.2.1

Footnote 1: why was simulation needed? Why not just add up EV of the better option across all choice problems?

In subsection 5: What does “we divided 1.6 GBP based on reward earned following EV consistent choices mean Divided by what? In the same paragraph, where do 20 and 12 come from?

In figure 1 of the Appendix, what “Similarly between different models in being plotted.” is supposed to mean?

10.3

I don’t understand what Figure 1 represents and why.

I can only guess what a “binned relative block/trial” mean.

Why should we “speculate that skilled decision makers in the loss or mixed domains in a tie-constrained environments should make slower decisions, ultimately preventing them from making any choice at all”? It looks rather speculative indeed.

Language

In general, the text is reasonably well written. Occasionally, the Candidate succumbs to wordiness. For example:
on p. 1, “In the real world” is redundant. Would the reader typically be interested in a fake world?
on p. 8. “essential” would do, “most essential” is too much.

This includes cases of tautology:

“Also, we counterbalanced… as well”; “continue the study further” (both in 10.2)

Here are a handful of other minor issues:
on p. 9, I don’t know what “efficient mathematics” means
on. p. 12, “variance… between” is unclear
On. p. 5 of 10.1: “data col lection was done online by appointing 73 participants” (??)
On. p. 7 of 10.1: “Every moment of our life is bombarded with information condensed in a statistical shell (??)
“a statistician… would continuously buy lottery ticket year after year…” – something really went wrong here in more than one way.

“Financial decision-making in the last decade” – it was probably meant that evidence has been accumulated in the last decade.

“CV1 is a synthetic predictor variable consist of…” (p. 15 of 10.1)

“lastly it is assumed that each participant to have varied…” (p. 18 of 10.1)

Summary and conclusion

To summarise, there are multiple issues that would require clarification. Perhaps three of them are most important:

1) the way the points are translated into money (and what participants knew about it)
2) which analyses followed pre-registration exactly, which did not and why

3) how does a more careful look at the DDM literature affect interpretation of the results and the contribution.

I probably have a general tendency to focus on potential problems and weaknesses. Here, I also recognise, as mentioned before, the ambitious plans, interesting research questions and designs, and high-quality data analysis. Obviously, the fact that all three studies have been published also speaks to their quality. I would thus be happy to learn that some of my critique was misguided. In any case, I believe the Candidate should be allowed to clarify the potentially problematic issues during a defence.

To be on the safe side, I provide the formal motion in Polish:

Recenzowana rozprawa spełnia warunki określone w art. 187 Ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (Dz.U. z 2018 r. poz. 1668). Rozprawa stanowi oryginalne rozwiązanie problemu naukowego, prezentuje ogólną wiedzę teoretyczną Kandydata w dyscyplinie Psychologia. Wnioskuję o dopuszczenie magistra Supratika Mondala do dalszych etapów przewodu doktorskiego.