Notes on K&T review

 Introductory. The representativeness heuristic proposed by Kahneman and Tversky has been and continues to be hugely influential. Not only has the 1973 paper inspired much further development of the research topic and been cited many times, it is a standard component of undergraduate modules on thinking and reasoning. It forms part of a general thesis developed by the authors that much thinking and reasoning is guided by heuristics rather than by, for example, strict rules of inference or knowledge of statistics and probabilities. The present authors propose a slightly altered replication of the set of studies reported in the 1973 paper. The registered report manuscript is nothing if not thorough, with detailed specification of the methods and analyses that would be used in the research. The thoughtful preparation for the research is admirable. However I do have some concerns which I will do my best to express.

 The opening paragraph does get across the prediction that prior probabilities will (often) be disregarded, but it does not state the qualifications to this, nor does it define representativeness. In fact there is no unambiguous definition of the representativeness heuristic and just characterising it as a heuristic draws a veil over the kind of processing that is actually going on when people make judgments of the sort exemplified in the research studies. Kahneman and Tversky (1973) (Hereafter "K&T") say that people "select or order outcomes by the degree to which the outcomes represent the essential features of the evidence" (pp. 237 - 238). That is the closest thing I can find to a definition in the 1973 paper. But the phrase "select or order" is odd, given that the research is concerned with judgments of likelihood. And what are "essential features" of evidence? Why are prior probabilities and other statistical information not part of that? People do judge by prior probabilities when they have no other information, or when the information they have seems not to be informative or relevant. Judging from the studies, it is really a contest between statistical information and individuating personal information and the latter usually wins. So perhaps they should have said, "judgments of likelihood about individuals are determined by relevant individuating information when available, not by prior probabilities". That makes it look less universal and less like a heuristic, more like a statement of people's ignorance about probabilities and how to use them in judgment. People more expert than I have written more extensively on these issues - Gigerenzer, whom the authors cite, is an example - and I think their work is very relevant to this manuscript and merits closer attention. As the authors are proposing a replication study they don't have to do a thorough critique of the representativeness heuristic and they don't have to agree with my analysis of it, but I do think they should address the problematic issue of what the representativeness heuristic is, how far its use generalises beyond the topics of the studies in the 1973 paper, and whether an unambiguous definition of it can be formulated. That much is important to understanding what is going on in the studies. Readers should be given a clear idea of what is really being tested, if possible. I should think a paragraph or two should suffice.

 The remainder of the introduction does a good job of reviewing relevant replications and critiques of the representativeness heuristic research. However my main concern is that I'm not sure what is the point of doing a replication of the studies, given the amount of water that has flowed under the bridge since they were carried out. The research literature has moved on, as the brief summary of relevant subsequent research makes clear, so what could we learn from replication of the original studies that would make a real contribution to the literature? If the proposed studies do indeed replicate the results reported by K&T, that just confirms that they suffer from the problems identified in subsequent research. If the proposed studies don't replicate the original findings, what would that mean? The authors should give some thought to that. In general, they need to make a case that there really is a need for the proposed replication.

 The various justifications for the replication given on p. 11 struck me as rather vague. On "the potential for improvements in methodology", that seems to me to be contradicted by the need for a replication study to use the same methods as were used in the original research. If there are going to be "further extensions examining the effect of consistency in numerical predictions", what do the authors hypothesise about consistency and why, and how does that issue fit into the existing literature? Extensions should be theoretically motivated and should test hypotheses. They mention "the absence of direct replications" but, unless replicability is likely to be a serious issue here, does that really matter, given the multiple studies that have added to or critiqued the research literature on representativeness? So I think there needs to be a stronger justification for replication, given the extent to which the field has moved on in the last 50 years.

 On p. 13 I thought the section "Selection of studies..." was unnecessary. The preceding section could just conclude with a sentence saying that all seven studies in the paper would be replicated. Table 1 is entirely adequate as a description of the studies, and the numbering is very useful given that the studies in K&T were not numbered. Table 2 is also a clear and useful summary of the results.

 p. 14 The rationale for the extension to study 4. I understand why the authors would want to investigate confidence more explicitly. However, saying "which we theorized would give a more straightforward estimate of participants' confidence" is a bit vague. Confidence judgments have been used in large numbers of studies on various topics, but they are explicit judgments, which are not always trustworthy. They could, for example, be prone to response biases such as self-presentation. I wonder if the authors should check whether explicit confidence ratings are generally regarded as valid. Also, if it turns out that the confidence ratings don't predict the standard deviations in judgments, how would that result be interpreted?

 On p. 21, "In Study 4, participants were given either adjectives and reports...". Should the "and" be "or"? If not, there is an "or" missing.

 In the next paragraph, again "either" is used but no "or" appears later in the sentence, so something is not right with that.

 On p. 31 it is stated that the studies will be run as an online Qualtrics survey. My experience of supervising final year project students over the COVID period, when Qualtrics was a common option, has not impressed me: many participants do not engage with the tasks and the data have been very noisy. I am concerned about this and I think the authors should discuss data quality. For example, what would count as evidence that a participant had not engaged with the task and what rules would there be about excluding participants that don't engage properly with it? What would be evidence of lack of engagement?

 In the participants section the authors mention recruiting people on Prolific. I had to google that to find out what it was and I think the authors should add some information about it for the benefit of readers in the same state of ignorance as me. What demographic information can they provide about samples obtained using Prolific? And in particular for study 7, do they not need students or ex-students who have done modules or courses on statistics? This needs to be sorted out.

 p. 26: "We ran the seven studies together in a single unified collection". It appears that this is the plan for the real data collection. Even though the authors say they have done this before, I do not think it is a good idea for all participants to take part in all the experiments. First, it is not the way K&T did it, so it compromises the fidelity of the replication. More important, the danger is that participants' knowledge of the experiments will accumulate as they go through, and that could have effects on their responses in the later experiments. They might, for example, be induced to reflect on what they are doing by the repeated presentations of personality information, and might change their thinking about its relevance. It would be much better to have separate samples for each experiment - but then I am uncertain whether there could be a target n of 800 for each one, because that would entail a total sample size of 5600. This needs to be clarified. The issue is discussed on p. 49. My response to that is that the authors should examine order as a moderator regardless of the results they get; I do not think this analysis should be contingent on the data.

 p. 30 The measure of statistical knowledge is rather vague and subjective. Would it be better to ask the Ps what education they have had in statistics - e.g. what modules at university and at what level - and how well they did? I see on p. 34 and p. 37 that they will be asked directly about their knowledge of confidence intervals. How will they answer that question? Will it be a free verbal report? What will the authors do about Ps who report that they don't know what a confidence interval is?

 Table 4 in the manuscript appears to have combined studies 1 and 2 from K&T, but it doesn't resemble the study 2 reported in K&T. In study 2 in K&T "the experimental materials consisted of five thumbnnail personality sketches of ninth-grade boys" (p. 240). Participants were divided into high and low accuracy conditions, on the basis of a statement saying how often people like themselves make correct predictions. Are the authors planning to do that? Possibly more information is needed there.

 On p. 49 the authors state that they set alpha to .005. How did they arrive at this decision? Did they use the Bonferroni correction? I think some sort of rationale should be given because there is a happy medium to be found between the risks of type 1 and type 2 errors and an arbitary choice might not be in the right place for that.

 I confess I don't understand why analyses were run with simulated data and I have nothing to say about that section of the report. To the extent that the fabricated data illustrate the kinds of analyses that will be run and the kinds of tables and graphs that will be generated, I think it all looks O.K.

Peter White