CSC2552: Review 6, Paper 2

Due on March 06

Ashton Anderson

496 words

Thomas Hollis

Paper 2

This paper, by Riederer, Hofman and Goldstein is a *digital virtual lab experiment* where turkers on Amazon's MTurk are used to investigate methods of generating analogies for estimating country populations and areas. The main result of this paper is the revelation that all analogy methods improve estimations and these analogies could have long term benefits for users.

The main weakness of this paper stems from its weak *external validity* which results in significantly reduced *generalisation*. Indeed, no attempt has been made to suggest that the same effects would hold true for tasks other than population and area estimation. To make matters worse, there are *confounds* with the incentive structure used here to compensate the turkers. Turkers care little about the quality of the research as long as they stay within the rules and complete the task in the shortest amount of time possible. Mitigation of this was attempted by choosing individuals with high approval ratings only, but rushed behaviour is difficult to detect and may result in *sample bias*. In addition, this combined with the *non-response bias* present in Experiment 3 casts doubt on the *internal validity* of the experiment. The main compromise here arises from the prohibitive cost and lack of scalability of the alternative of doing this as an analogue experiment under full laboratory conditions. Another notable weakness of this paper is its vulnerability to *coverage bias* of the turker methodology. Indeed, by only considering humans that are users of Amazon MTurk, we are only measuring the effect of a particular slice of the population of the US that would not necessarily be representative, especially for tasks correlated to IQ [1].

Conversely, a significant strength of this paper arises from its *controlled trial* format [2]. Since the use of controls was carefully implemented, this paper does show beyond reasonable doubt a causal improvement in estimation brought along by using analogies. This is also in line with exceptions from clearly identified prior studies and benchmarks [3]. However, this method does have its drawbacks. Most notably, it requires an experiment that is costly to scale to more than a few thousand participants as each needs to be paid. An observational alternative, while less realistic, would perhaps have helped address the issue of external validity present in this paper. Alternatively, a *natural experiment* seems possible in the context of education in schools using standardised tests.

While the implications of this paper seem to at first glance provide significant tangible improvement to Bing search, this feature has now been revoked from Bing [4]. In addition, my interpretation of the results is more skeptical than those of the authors, perhaps biased by their employment at Microsoft. While I agree with the major benefits of using analogies for estimation, I don't think enough evidence has been presented to suggest that "beneficial effects of perspectives can remain significant for at least six weeks after", as this benefit could be due to better memorisation due to increased thinking about the task.

[1] Ross, J., Zaldivar, A., Irani, L., and Tomlinson, B. (2009). Who are the Turkers? Worker Demographics in Amazon Mechanical Turk. Department of Informatics, University of California, Irvine, USA, Tech. Rep.

[2] Salganik, M. J. (2017). Bit By Bit: social research in the digital age. Princeton University Press.

[3] Brannon, E. M. (2006). The representation of numerical magnitude. Current opinion in neurobiology 16, 2 (2006), 222229. http://doi.org/10.1016/j.conb.2006.03.002

[4] Bing, Microsoft (2019). Area of Pakistan. Available at: goo.gl/SZh5zs (Accessed: 18/02/2019)