"The (Statistical) Power of Amazon Mechanical Turk"

Experimental linguists have long had the need to recruit subjects—to make grammaticality judgments, to participate in experiments, or to help in data analysis. At the outset of most research, linguists must determine how many subjects will let them test their hypothesis in a statistically valid way. In an ideal world, when determining the number of subjects to use researchers would avoid underpowered or overpowered studies by estimating effect size and conducting a statistical power analysis to determine the proper number of subjects. In practice, it may be difficult to estimate effect size, and experimenters have historically been constrained by equipment limited subject availability, funding, and amount of man hours available to analyze data. In this talk, I will argue that the advent of Amazon Mechanical Turk (AMT), an online marketplace of paid workers who may be used as subjects, lessens these concerns for many types of experiments and allows researchers to greatly increase the power of their studies quickly and with minimal funding. I will show that (despite some obvious limitations of using distant subjects) with properly designed experiments data from AMT is trustworthy, cheap, and much faster than traditional face-to-face data collection. Not only this, but AMT workers may help with data analysis. These two facts greatly increase the scope of research that one researcher may carry out.

In my talk I will first give an overview of research that indicates that data collected on AMT is as good or better than data collected in a lab for several types of tasks [1-4]. Next I will give information about the demographics of participants on AMT, and how to recruit only specific demographic groups. I will follow with an example experiment of my own and practical information about how to set up studies on AMT that provide the cleanest possible data and are in keeping with IRB requirements. I will also elucidate the potential uses of AMT workers as participants in data analysis, coding, and annotation. I will close with an argument that because of the availability of AMT and other resources, researchers should turn their attention to statistical analyses to determine the correct number of subjects to create a statistically valid experiment.

[1]Paolacci, G., J. Chandler, and P. Ipeirotis. (2010.) "Running Experiments on Amazon Mechanical Turk." *Judgment and Decision Making* 5 (5): 411–419.

[2]Buhrmester, M. D., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6, 3-5.

[3] Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon. com's Mechanical Turk. *Political Analysis*, *20*(3), 351-368.

[4]Crump, M.J.C., McDonnell, J.V., Gureckis, T.M. (2013). Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *PLoS ONE* 8(3).