Making Items Matter More

Thomas Weskott (Georg-August-Universität Göttingen)

The turn towards controlled experimentation in linguistics in the last 25 years has lead to a surge of experimental studies. This trend has been boosted by the advent of crowdsourcing technology which has increased the availability of experimental data. While these developments are certainly desirable, they have lead to a certain sloppiness in the treatment of experimental materials. Many reported studies are (i) severely underpowered with respect to the item samples; (ii) tend to overestimate the generalizability of the results beyond the sample of linguistic expressions; and (iii) exhibit a lack of control over the task performance by the participants: in many experiments, the authors forego the use of fillers and benchmarking items, thereby jeopardizing the validity of their results. I will first give a few example of this malpractice and then go on to show by means of a simulation study how underpowered studies can lead to type I error inflation even in the case of large participant samples. I will conclude with a discussion of "intelligent benchmarking" as a solution to the validity problem.