Reply to Data Colada [128]
Yulia Evsyukova, Wladislaw Mill, Felix Rusche
June 16, 2025

We are thankful for the thorough and constructive criticism of our design. While it's important to highlight that the points put forward by Uri do not affect our results or the interpretation thereof, we generally agree that variation in stimuli is important for external and internal validity. Below, we discuss our experimental choices and express our thoughts on the trade-offs we faced when working with visual stimuli. These thoughts may also be relevant to future work using visual – or really any kind of – stimuli.

We also want to highlight that we generally agree with the premise that having little variation in stimuli can hurt internal and external validity. This is why we opted to create hundreds of different CVs and a unique AI-generated profile image for each of our profiles. At the same time, we agree that the images do share a lot of commonalities. On the spectrum between having two images only and having 200 twin-pairs with fully differentiated images, our study is probably closer to the former option. There are two reasons for this.

The first one is technical. When we generated the images, we relied on a collection of 100k images provided by StyleGAN2 (FFHQ). Unfortunately, Black individuals are extremely underrepresented in the data, such that we could only identify about 40 Black males in our age group of interest (which needed to be rather young to be realistic for just creating a LinkedIn profile). This meant that we had to effectively create “grandchildren” of these pictures to obtain a sufficient number of unique images for our profiles. Because of this, the resulting images look somewhat more similar than if we had had 200 Black images to begin with.

The second reason for keeping characteristics across images relatively stable was a deliberate choice as part of the experimental design: while we do generally agree that creating more variation across images would further strengthen the study’s external and internal validity, moving to fully different images across twin-pairs also comes at a substantial cost to statistical power. For instance, our study only included eight profiles per US state. Had we gone closer to the other extreme and introduced a lot of variation in images, chances are that, e.g., Republican states would have received more attractive-looking images than Democratic ones (if we had kept the sample size and not stratified further). This would have strongly undermined any heterogeneity analyses by geography. Simultaneously, we were a lot more interested in understanding heterogeneous differences on the target level than differences between profile images with varying characteristics. Of course, an obvious solution would have been to create tens of thousands of fake profiles to gain enough statistical power for these differences between states not to play a role. In a study like ours, which involved creating hundreds of fake accounts on LinkedIn, this is not

viable, though. Specifically, it would have strongly raised the (already high) costs of running the experiment, increased the risk of being systematically detected and blocked by LinkedIn, and also thrown up ethical and legal questions of flooding the platform with tens of thousands of fake profiles.

In other words, the sweet spot lies somewhere in-between the extremes of having two images only and having fully differentiated images across twin-pairs. Specifically, we believe that in a study like ours, the sweet spot is where the researcher retains full control of an image's characteristics while still introducing some variation between images (i.e., using more than two images). This, however, means that the researcher needs to choose which dimensions of interest she/he is simultaneously able to cleanly vary. For example, in our study, we were interested in varying race. Other characteristics of obvious interest – such as age or gender – were ruled out by design. Beyond these demographics, it is unclear which dimensions would have been of immediate interest. Specifically, we were interested in characteristics assigned by birth, meaning that hairstyle, facial expression, or professionalism were not too interesting to us for this study.

To sum up, we agree with the premise that variation in stimuli is important for external and internal validity. While our study includes a lot of such variation (e.g., through hundreds of different CVs and profile images) and, as such, arguably more so than most comparable studies, the images used are somewhat similar. This is in part explained through deliberate experimental choices and in part explained through limitations in the training data.