Survey Inventory Reliability Analysis

Since we modified the number of scale points on some surveys and used selected survey inventories that are not validated in the presented work, we performed a reliability analysis on relevant opening and post-stimulus survey inventory data to confirm that scales were working as intended.

In the initial video-based study with university students, for the three NARS scales (negative attitudes toward *interactions with robots*, *social influence of robots*, and *emotions in robots*), the Cronbach's alpha values were $\alpha = 0.724$, $\alpha = 0.758$, and $\alpha = 0.670$, respectively. For RoSAS *warmth*, *competence*, and *discomfort*, the results were $\alpha = 0.964$, $\alpha = 0.977$, and $\alpha = 0.945$, respectively. For IOS, a reliability analysis does not apply, or in other words, $\alpha = 1.000$ since there is only one question. For JRS, $\alpha = 0.952$, and for Godspeed *anthropomorphism* and *likeability*, $\alpha = 0.975$ and $\alpha = 0.977$, respectively. The reliability analysis for our custom profanity-tolerance scale yielded a value of $\alpha = 0.72$. These results show that the NARS scales demonstrated moderate internal consistency, as did the profanity-tolerance scale. All other scales showed high internal consistency; thus, we were confident of the scales despite the removal of the "Neutral" anchor point.

In the follow-up Prolific study, for the NARS scales, the Cronbach's alpha values were reported as $\alpha = 0.825$, 0.823, and 0.724, for scales 1, 2, and 3, respectively. For RoSAS *warmth*, *competence*, and *discomfort*, the alpha values were $\alpha = 0.971$, 0.979, and 0.945, respectively. For IOS (inevitably), $\alpha = 1.000$. For JRS, the alpha was $\alpha = 0.948$, and for Godspeed *anthropomorphism* and *likeability*, the alpha values are $\alpha = 0.979$ and 0.961, respectively. Reliability analysis for our profanity-tolerance scale revealed $\alpha = 0.825$. These results show similar or better consistency for all scales (compared to in the initial study), so again we felt confident that scales were functioning as intended.

We did not perform a reliability analysis for the in-person deployment since each of the survey scales used in that phase of the work was single-item.