between\_day\_reliability

Giacomo Bignardi

2023-10-18

## S8. Between-day reliabilities for pairwise aesthetic agreement, taste-typicality, and evaluation-bias.

As an additional sensitivity analysis to the one carried out in S7, since memory effects can inflate reliabilities computed on repeated images, we further quantified the stability of taste typicality and evaluation-bias (following 2). We analysed a third openly available sample for which repeated ratings of images were obtained across days (CITE). We used data from n = 78 unrelated individuals with a mean age of 34 y (sd = 7 y, ranging from 21 to 49 y; 47 women, 30 men, 1 other) available at <https://osf.io/35zf8/?view_only=e76c6755dcea4be2adc5b075cae896e8>. These individuals followed the same experimental procedure undertaken by the twins in the validation sample, rating 150 images of faces (including 50 repeats) and 74 images of scenes (including 24 repeats), with the addition of taking part in the experiment a second time, after a 3 days delay (More details are given in (CITE)). This allowed us to compute conservative for individual taste-typicality and evaluation-bias scores and test its stability over time, further complementing the taste-typicality and evaluation-bias reliability analysis with an estimate of pairwise for aesthetic agreement.

On the one end, after removing 3 individuals with < .5 (only for faces, as for scenes there were no individuals with <. 5), Pearson correlations were *r* = 0.8, 95% CI [0.7, 0.87], and *r* = 0.83, 95% CI [0.74, 0.89], for taste-typicality and *r* = 0.83, 95% CI [0.75, 0.89], and *r* = 0.91, 95% CI [0.86, 0.94] for evaluation-bias, for scenes and faces, respectively. On the other, pair-wise correlations for aesthetic agreement computed on random pairs were equal to *r* = 0.72, 95% CI [0.52, 0.84], and *r* = 0.68, 95% CI [0.46, 0.83], for scenes and faces, respectively, indicating a small decay in repeatability in pairwise aesthetic agreement compared to taste-typicality and evaluation bias. Taste-typicality correlations between days were higher than pair-wise correlations between days, with the difference amounting to = 0.08 and = 0.14, for scenes and faces, respectively However, these differences were not significant (Fisher’s *z* = 1.61, *p* = 0.11 and Fisher’s *z* = 1.61, *p* = 0.11). As such, more research using larger samples, more time points, and larger time differences between time points is needed to test whether taste-typicality is more stable than aesthetic agreement over longer periods.

For comparison, Intra Class Coefficients (ICC(2,1)) computed within pairs and individuals between scores across days were ICC(2,1) = .80, ICC(2,1) = .83, and good to excellent reliability for taste-typicality ICC(2,1) = .81, ICC(2,1) = .91, and evaluation-bias for scenes and faces, respectively. Overall, these results show good reliability between days and relatively good stability of the metrics used to assess variability in aesthetic value and indicate little room for unreliable measurement to confound E estimates.