nature human behaviour

ARTICLES

https://doi.org/10.1038/s41562-019-0782-4

Economic status cues from clothes affect perceived competence from faces

DongWon Oh[1*], Eldar Shafir[2,3] and Alexander Todorov[2]

Impressions of competence from faces predict important real-world outcomes, including electoral success and chief executive officer selection. Presumed competence is associated with social status. Here we show that subtle economic status cues in clothes affect perceived competence from faces. In nine studies, people rated the competence of faces presented in frontal headshots. Faces were shown with different upper-body clothing rated by independent judges as looking 'richer' or 'poorer', although not notably perceived as such when explicitly described. The same face when seen with 'richer' clothes was judged significantly more competent than with 'poorer' clothes. The effect persisted even when perceivers were exposed to the stimuli briefly (129 ms), warned that clothing cues are non-informative and instructed to ignore the clothes (in one study, with considerable incentives). These findings demonstrate the uncontrollable effect of economic status cues on person perception. They add yet another hurdle to the challenges faced by low-status individuals.

Impressions of competence from faces predict important real-world outcomes[1–5]. These impressions have been shown to be non-deliberate and triggered by brief exposure[1,6–8] (for review, see ref. [9]). For instance, the perceived competence of faces of winners and runners-up predicted election outcomes when participants were exposed to the faces for 100 ms, 250 ms or for an unlimited time[1]. The visual context around a face also affects person perception: bodily gestures influence the perception of facial emotion[10,11] while cues of social status or culture influence the perception of facial ethnicity[12,13]. Verbally provided social-status information also affects face perception[14], and clothes suggestive of high-status professions (for example, a medical doctor) elicit more attention and better face recall than do clothes suggestive of low-status professions (for example, a fry cook)[15].

The present research tests whether subtle economic status cues, in the form of 'richer'- or 'poorer'-looking clothing, otherwise clean and intact, influence perceived competence. People expect individuals of higher socioeconomic status to be more competent[16]. Social status is linked to economic success, job prestige and competence across cultures[17], and clothing is often diagnostic of socioeconomic status[18]. The following studies presented images of faces appearing with upper-body clothes that were independently rated as richer or poorer (Fig. 1a; see Supplementary Figs. 1 and 2 for stimuli): participants rated the faces for competence (studies 1–8) or chose the more competent in a pair of faces (study 9). Importantly, the clothing manipulation was subtle: when asked to describe them, no one in an independent group of participants described these clothes as notably rich or poor. Moreover, quantitative analyses found no differences in how positively or negatively these clothes—rich or poor—were described ($F(2,4893) = 1.57$, $P = 0.208$), suggesting that the clothing-status effect cannot be attributed to a general valence of impressions from these clothes (see Methods and Fig. 1d).

Results

To test whether economic status cues from everyday clothes affect competence as judged from faces, participants were presented with pictures of individuals' faces paired with different upper-body clothing. Each face occurred once with clothes rated by an independent group of judges as richer and once with clothes rated as poorer (although not obviously so, as judged by a separate group of respondents; see Methods for details). Participants were asked to judge the competence of each face (studies 1–8) or to compare competence between faces (study 9). Faces were judged significantly more competent when seen with richer clothes than with poorer clothes. We employed various measures in an attempt to attenuate the effect of clothing cues, but to no avail (see Methods and Fig. 2 for details).

We first tested whether clothing influences perceived facial competence and whether this influence might vary with length of exposure to the stimulus (study 1). We then investigated whether participants were able to avoid the influence of clothing cues on their competence judgements, having been explicitly instructed to disregard components "such as clothes" (studies 3 and 6–8). We ascertained that the observed effects cannot be attributed to deliberate inferences based on the clothing (studies 3–8), appear effortless since they arise at extremely short intervals (see, for example, Supplementary Table 3) and are hard for respondents to control (study 8). We replicated the effects with older participants (studies 2 and 3c) and reproduced the original effect in the realm of choice rather than judgement (study 9).

Across studies 1–8, we found a significant main effect of clothing type on competence judgements (Fig. 2 and Supplementary Table 1; for the effect for the shortest duration see Supplementary Results and Supplementary Table 3). This effect was highly consistent across studies and across faces, as shown in Fig. 3. In each study, most faces (>83%, median across studies = 94%) were perceived as more competent when seen with richer than with poorer clothes (for a discussion of why faces may have benefitted to different extents from clothing cues, see Supplementary Results and Supplementary Fig. 6). We report the full results and the motivation behind each study in what follows. In addition to a main effect of clothing type, we observed effects of face race (Supplementary Figs. 3 and 4 and Supplementary Table 2), presentation time (Supplementary Fig. 3 and Supplementary Table 3) and participant age (Supplementary Fig. 4). Because these effects were not predicted before data collection,

[1]Department of Psychology, New York University, New York, NY, USA. [2]Department of Psychology, Princeton University, Princeton, NJ, USA. [3]Woodrow Wilson School of Public and International Affairs, Princeton University, Princeton, NJ, USA. *e-mail: dongwon.oh@nyu.edu

and were smaller and less consistent than the main effect of clothing, we discuss these in the Supplementary Results.

To assess the effect of clothing-status cues on perceived competence in studies 1–8, we conducted a repeated-measures analysis of variance (ANOVA) with clothing type (richer/poorer) as a within-subject variable. Face race (black/white) and presentation duration (129/553/1,059, 300/600/1,100 or 250/750 ms) were included as within-subject variables (except in study 7, where duration was set at 750 ms). The dependent variable was the mean competence rating. Undergraduate student participants ($n=24$) rated faces appearing in richer clothing as more competent than the same faces in poorer clothing (study 1, $F(1,23)=70.52$, $P<.001$, generalized η-squared $\eta_G^2=0.37$ (90% CI = 0.31; 0.45); Figs. 2 and 3 and Supplementary Table 1). To ascertain that students were not particularly attuned to economic cues in clothing, older participants were recruited ($n=52$, mean age = 39.16 years) and showed the same effect: faces seen in richer clothing were rated as more competent than when seen in poorer clothing (study 2, $F(1,51)=73.95$, $P<0.001$, $\eta_G^2=0.15$ (90% CI = 0.12; 0.18); for further analysis of the effect of participant age, see Supplementary Results and Supplementary Fig. 5). Even when participants ($n=36$) were asked to "focus on the person, and ignore other features such as the clothes", the effect for clothing type persisted (study 3a, $F(1,35)=56.52$, $P<0.001$, $\eta_G^2=0.14$ (90% CI = 0.11; 0.19)). Then, retaining the "ignore the clothes" instructions, we applied Fourier phase-scrambled images of the faces as visual masks to better control the presentation duration (Fig. 1c). The effect for clothing type persisted for both young (study 3b, $n=36$, mean age = 19.78 years, $F(1,35)=82.82$, $P<0.001$, $\eta_G^2=0.18$ (90% CI = 0.15; 0.22)) and older participants (study 3c, $n=51$, mean age = 39.39 years, $F(1,50)=51.35$, $P<0.001$, $\eta_G^2=0.07$ (90% CI = 0.05; 0.09)).

Perhaps participants deliberately infer competence from clothing cues? Poorer clothes may signal lower socioeconomic status—perhaps a less successful career—whereas richer clothing may signal success. To discourage such inferences, participants in the next study ($n=36$) were provided with information intended to equalize expectations. They were told that the pictures were of people who "work in sales at a mid-size firm in the Midwest, and earn around US$80,000 a year". This information notwithstanding, the effects persisted unchanged (study 4, $F(1,35)=79.32$, $P<0.001$, $\eta_G^2=0.35$ (90% CI = 0.30; 0.40)).

Might some of the effect be attributable to the formal aspects of some of the clothes? Suits and ties occurred with some regularity in the higher-economic status condition and may have signalled greater competence. We replaced all formal attire (suit or tie) with plain, non-formal shirts and topwear (Fig. 1b; see Supplementary Fig. 2 for stimuli). The wardrobe change notwithstanding, the effect of clothing type persisted (study 5, $n=36$, $F(1,35)=36.62$, $P<0.001$, $\eta_G^2=0.09$ (90% CI = 0.07; 0.12)). To test for robustness even further, we used the explicit instructions to ignore the clothes (as in study 3) in combination with non-formal clothing (as in study 5) and the post-stimulus phase-scramble masking (as in studies 3b,c, 4 and 5), and we increased the sample size ($n=200$). With all these added measures, we again found the same effect (study 6, $F(1,199)=151.96$, $P<0.001$, $\eta_G^2=0.03$ (90% CI = 0.03; 0.04)).

Perhaps our observed effect somehow depends on the unique combination of faces and clothes used. Because the same face–clothes combinations were used in studies 1–6 (see Methods), the clothing-status effect might be attributable to the specific faces and clothes rather than to richer and poorer status cues more generally. To rule this out, we combined faces and clothes randomly, generating new person images that existed neither in the stimulus-preparation pilot nor in studies 1–6. The new stimuli were 80 images consisting of every combination of eight faces (four black, four white) and ten clothes (five poorer, five non-formal richer). Participants ($n=50$) were randomly presented with these new, non-pretested images, and

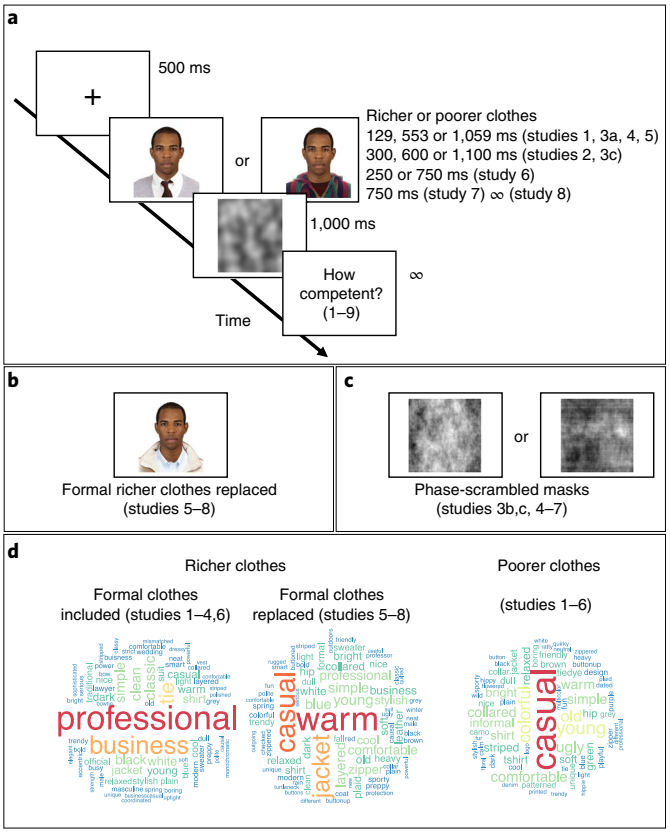

Fig. 1 | Experimental procedure and stimuli in the competence-rating task. **a**, Studies 1–8 presented in each trial an image of a face paired with with richer or poorer clothes, followed by a mask and a rating question. **b**, Studies 5–8 replaced all suits and ties with non-formal clothing. **c**, Studies 3b,c and 4–7 replaced a cloud-image mask with phase-scrambled masks customized to each image to better control the presentation duration. **d**, Use of both visual inspection and quantitative analyses of the descriptions of all clothing items in the stimulus set, elicited from independent raters, found that the status-cue manipulation was subtle (see Methods). The word clouds represent the description of richer clothes including formal clothes (left), richer clothes with formal clothes replaced by non-formal clothes (middle) and poorer clothes (right). Descriptions given at least five times are shown (>4,700 total responses), with size of presentation corresponding to the frequency of mention. The face image in **a** was adapted with permission from ref. [29].

the effect of status cues in clothes persisted (study 7, $F(1,49)=35.84$, $P<0.001$, $\eta_G^2=0.09$ (90% CI = 0.06; 0.12); for the effect's generalizability across images, see Supplementary Fig. 7).

How controllable is the effect of clothes-status cues on competence judgements? Although the effect persisted through various manipulations (including explicit instructions to ignore the clothes), one might argue that these manipulations were simply not sufficiently strong to motivate participants to suppress the immediate bias. To better assess the controllability of the effect, we introduced an added incentive to be accurate. Specifically, in addition to the advice to ignore the clothes (as in studies 1–7), participants ($n=63$) were told that "(the) participant whose ratings are the most accurate will receive, in addition to their standard pay, an additional US$100 reward". Accuracy, they were told, would be determined by how close a participant's ratings are to those of participants who saw the faces without the same clothes. Despite the added incentive to ignore the clothes, participants' competence ratings persisted in being influenced by clothing cues, as shown in Fig. 3b (study 8,

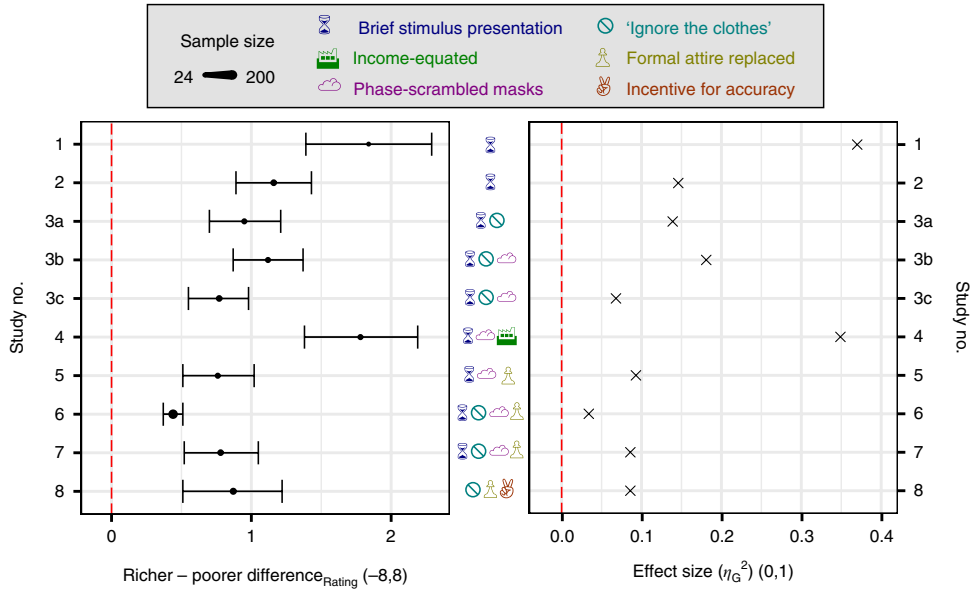

Fig. 2 | Effect of economic status cues from clothes on competence ratings. Left: mean differences in the competence ratings of faces paired with poorer versus richer clothes and their 95% confidence intervals (CI_{diff}) for all rating studies. The red dashed line represents no-clothing bias. Across studies, faces paired with richer clothes were rated as more competent than the same faces paired with poorer clothes ($P < 0.001$). For illustrative purposes, study 8 ratings (obtained using a different scale) were rescaled; study 1, difference = 1.84 (95% CI = 1.39; 2.29); study 2, difference = 1.16 (95% CI = 0.89; 1.43); study 3a, difference = 0.95 (95% CI = 0.70; 1.21); study 3b, difference = 1.12 (95% CI = 0.87; 1.37); study 3c, difference = 0.77 (95% CI = 0.55; 0.98); study 4, difference = 1.78 (95% CI = 1.38; 2.19); study 5, difference = 0.76 (95% CI = 0.51; 1.02); study 6, difference = 0.44 (95% CI = 0.37; 0.51); study 7, difference = 0.78 (95% CI = 0.52; 1.05); study 8, difference = 0.78 (95% CI = 0.46; 1.09). Right: effect sizes of economic status cues from clothes (see Supplementary Table 1 for full statistics). The red dashed line represents no-clothing bias. The effect was significant across all studies; study 1, $F(1,23) = 70.52$, $P < 0.001$, $\eta_G^2 = 0.37$ (90% CI = 0.31; 0.45); study 2, $F(1,51) = 73.95$, $P < 0.001$, $\eta_G^2 = 0.15$ (90% CI = 0.12; 0.18); study 3a, $F(1,35) = 56.52$, $P < 0.001$, $\eta_G^2 = 0.14$ (90% CI = 0.11; 0.19); study 3b, $F(1,35) = 82.82$, $P < 0.001$, $\eta_G^2 = 0.18$ (90% CI = 0.15; 0.22); study 3c, $F(1,50) = 51.35$, $P < 0.001$, $\eta_G^2 = 0.07$ (90% CI = 0.05; 0.09); study 4, $F(1,35) = 79.32$, $P < 0.001$, $\eta_G^2 = 0.35$ (90% CI = 0.30; 0.40); study 5, $F(1,35) = 36.62$, $P < 0.001$, $\eta_G^2 = 0.09$ (90% CI = 0.07; 0.12); study 6, $F(1, 199) = 151.96$, $P < .001$, $\eta_G^2 = 0.03$ (90% CI = 0.03; 0.04); study 7, $F(1,49) = 35.84$, $P < 0.001$, $\eta_G^2 = 0.09$ (90% CI = 0.06; 0.12); study 8, $F(1,62) = 24.29$, $P < 0.001$, $\eta_G^2 = 0.09$ (90% CI = 0.06; 0.12). Symbols denoting measures taken to attenuate the effect of clothing cues appear between the left and right plots.

$F(1,62) = 24.29$, $P < 0.001$, $\eta_G^2 = 0.09$ (90% CI = 0.06; 0.12)). In fact, the US$100 reward had no perceptible effect on participants' judgements: the effect size was almost identical in studies 7 and 8, which used the same image set (Fig. 2 and Supplementary Table 1). This finding supports the notion that the effect of status cues on perceived competence is indeed hard to control, at least under the conditions tested here.

Study 9 ($n = 64$) extended our investigation further. Here, we employed a choice task: participants saw pairs of faces and chose the face that appeared more competent. Faces were drawn from three distinct competence levels (16 high-, 16 medium- and 16 low-competence pairs). Each pair consisted of faces that had been rated equally competent. One face appeared with clothes rated richer than those of the other. Furthermore, half of the participants (at each competence level) were explicitly warned that clothing, although not diagnostic of competence, may affect their judgement. Presented with faces originally rated equally competent, participants chose the face with richer clothes as more competent than its lower-status counterpart 69% of the time (Fig. 4; $t(63) = 11.56$, $P < 0.001$ (95% CI = 66%; 72%)). To assess the effect of the warning—regarding no relation between clothing and competence—we ran a mixed-effect ANOVA with the warning condition (no warning/warning) as a between-subject variable and facial competence level (high/medium/low) as a within-subject variable. The dependent variable was the proportion of richer faces chosen. We found no significant effects of warning (Fig. 4; $F(1,62) = 0.49$, $P = 0.486$), level of competence ($F(2,124) = 1.07$, $P = 0.349$) or an interaction between the two ($F(2,124) = 0.03$, $P = 0.967$). Clothing had a significant effect on which face was seen as more competent at all three competence levels, irrespective of the warning regarding clothing's biasing effects.

Discussion

Across studies, we found that economic status clothing cues influenced competence judgements of faces. The effect persisted when faces were presented very briefly (that is, 129 ms), when information was provided related to the person's profession and income, when formal clothing was replaced by more casual clothing, when participants were advised to ignore the clothing, when they were warned that there was no relationship between clothing and competence before choosing rather than rating faces, and when participants were offered a monetary reward for accuracy. These findings support the notion of uncontrollable effects of minor contextual cues in face perception, and are consistent with a large body of research that finds people spontaneously encode the context surrounding a face when making social judgements[10,11,13,19] (while our focus has been on competence judgements, similar, if attenuated, effects can be observed for other traits, such as trustworthiness; see Supplementary Results and Supplementary Fig. 8).

The strong and persistent effects we observed are consistent with theoretical work[16,18] and empirical findings[16,17], showing a robust tendency for people of lower economic status to be perceived as less competent and to be disrespected[20], often leading to social exclusion with detrimental effects on physical and emotional health[21]. Poverty is a place where many challenges—physical, social and psychological—converge: being perceived as of lower competence and disrespected adds to those challenges, and can exacerbate

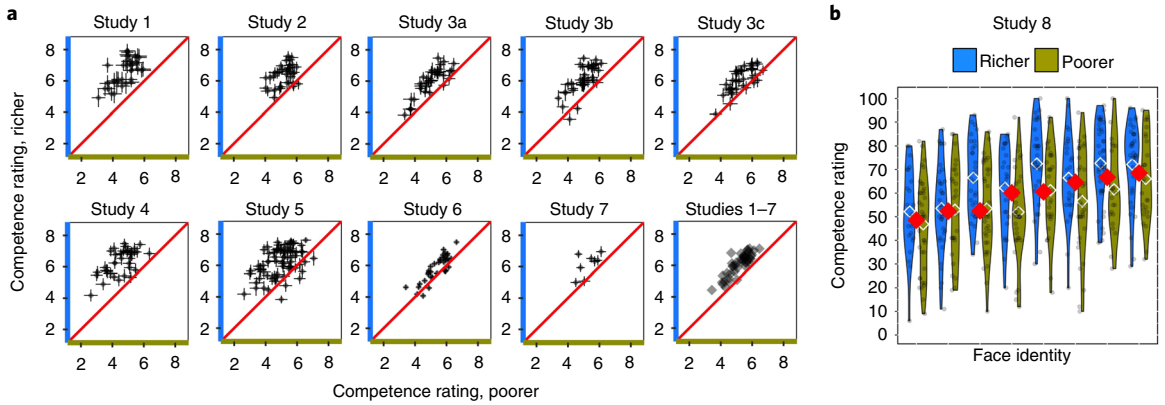

Fig. 3 | Competence ratings of faces. a, Competence ratings of faces averaged across participants in studies 1, 2 (plain versions), 3a–c ("Ignore clothes" warnings), 4 ("Equal income" information), 5 (no formal attire), 6 (no formal attire, "Ignore clothes" larger sample size) and 7 (novel face × clothes combinations, no formal attire, "Ignore clothes"), and studies 1–7 combined (mean ratings weighted by sample size). Each dot represents the ratings of a face presented with poorer (x) or richer clothes (y). Horizontal and vertical error bars denote the s.e.m. for the corresponding rating. Dots above the red diagonal line ($y = x$) represent faces rated more competent with richer than with poorer clothing. Richer clothes yielded higher competence ratings in almost all cases (median across studies 1–7 = 94.44% of faces). **b**, Competence ratings of faces in study 8 (monetary incentives). Faces are horizontally ordered by facial competence score. Each dot contributing to a violin plot corresponds to a single participant. Each diamond within a violin plot represents the mean rating of a face presented with poorer (olive) or richer (blue) clothes. Red diamonds represent the 'pure' facial competence score—the mean ratings ($n = 30$) of the face presented without clothes. Dots above red diamonds represent faces rated as more competent when seen with than without clothes; dots below red diamonds represent faces rated as less competent when seen with than without clothes. Monetary incentives notwithstanding, people's judgements of competence were swayed by status cues in clothing relative to when those same faces appeared with no clothes, $F(1,62) = 24.29$, $P < 0.001$, $\eta_G^2 = 0.09$ (90% CI = 0.06; 0.12).

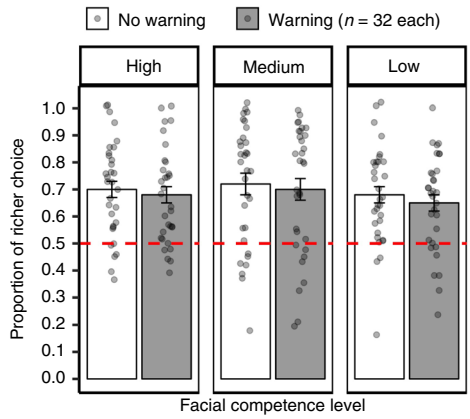

Fig. 4 | Proportion selecting face with richer clothing as being the more competent in a choice task. Participants in study 9 were presented with pairs of faces drawn from three distinct competence levels—low, medium and high. Faces within each pair had originally been rated equally competent on average. Participants selected which face appeared more competent. They were warned that there was no relationship between clothes and competence, or were given no warning. Across all three competence levels, faces in richer clothing were selected significantly more often than those in poorer clothing; high competence, mean = 69%, $t(63) = 8.52$, $P < 0.001$ (95% CI = 64%; 73%); medium competence, mean = 71%, $t(63) = 7.31$, $P < 0.001$ (95% CI = 65%; 77%); low competence, mean = 66%, $t(63) = 7.35$, $P < 0.001$ (95% CI = 62%; 71%). The warning ($F(1,62) = 0.49$, $P = 0.486$), the level of competence ($F(2, 124) = 1.07$, $P = 0.349$) and their interaction ($F(2, 124) = 0.03$, $P = 0.967$) did not affect participant choices. Each dot corresponds to a single participant; the red dashed line represents no-clothing bias. Error bars denote s.e.m.

cognitive load and hamper performance, thereby potentially proving self-fulfilling[22,23].

To overcome a bias, one needs not only to be aware of it but to have the time, attentional resources and motivation to counteract the bias[24]. In our studies, we warned participants about the potential bias, presented them with varying lengths of exposure, gave them additional information about the targets and offered financial incentives, all intended to alleviate the effect. None of these interventions were effective, however. While it is possible that higher incentives and greater experience could reduce the bias, its persistence in the face of our various manipulations is impressive.

The present findings demonstrate that economic status cues from clothes naturally intervene in people's assessments of competence. This is consistent with research showing that people associate status with competence in stereotypes of social groups[16,25]. This strong status–competence association suggests that any attempt at independen manipulation of the apparent competence and economic status of a person may need to resort to explicit and salient manipulations, rather than fairly subtle cues.

The poor clothing cues in our studies were benign compared to real-world poverty signals. Recent work has shown that people can accurately guess others' social class from brief exposure to photos or speech recordings[26]. We might thus expect people wearing truly impoverished clothes or exhibiting other peripheral signs of poverty to encounter substantial low-competence stereotyping, both when perceivers think fast as well as when they have more time to deliberate.

Stereotypes about rich and poor individuals are common, prominent and consequential. Just as the clothing cues in our studies led to differential disambiguation of facial competence, views about a person's economic background can lead to notably different interpretations of what is otherwise ambiguous performance[27]. Beyond their immediate impact, an important question for future research concerns the extent to which we might be able to transcend first impressions.

Methods

The experimental protocol for all studies was approved by the Institutional Review Board for Human Subjects of Princeton University (protocol no. 7212). We obtained informed consent from all participants. We determined sample size before data collection in every study, and we stopped data collection once the predetermined sample size was reached. Individuals of different self-identified races—white, black and Asian were the most frequent—and genders participated in

each study. Data distribution was assumed to be normal and equal in variance, but this was not formally tested. All datasets, which include participant demographics and all stimuli, are publicly available at Open Science Framework: https://osf.io/v2j43/). All analyses are based on two-sided tests. Data collection and analysis were not performed blind to the conditions of the experiments. Confidence intervals (CI) for the main effect sizes (η_G^2) were calculated via bootstrapping.

Studies 1–5. *Participants.* Princeton University undergraduates and community members participated, for either course credit or payment ($n = 271$ in total; study 1, $n = 24$, mean age = 24.88 years, 17 female, 7 male; study 3a, $n = 36$, mean age = 19.89 years; 25 female, 11 male; study 3b, $n = 36$, mean age = 19.78 years, 24 female, 11 male, 1 other; study 4, $n = 36$, mean age = 19.42 years, 23 female, 12 male, 1 no response; study 5, $n = 36$, mean age = 21.97 years, 25 female, 11 male). To gauge the replicability of the findings with a non-student, older, population, adults at a shopping mall in central New Jersey, USA participated for payment (study 2, $n = 52$, mean age = 39.16 years, 35 female, 16 male, 1 no response; study 3c, $n = 51$, mean age = 39.13 years, 36 female, 12 male, 3 no responses). Power analysis using G*Power v.3.1.9 (ref. [28]), based on a within-subject ANOVA design ($\eta^2 > 0.26$) with a moderate level of correlation between the within-subject measures ($r = 0.5$, Pearson correlation between richer and poorer face ratings calculated at the participant level), suggested more than $n = 20$ participants. We decided to stop collecting data at natural stopping points, such as end of day, at somewhere between $n = 30$ and 60. The observed effect size in studies 1–5 proved smaller than initial estimates, and informed sample-size decisions in subsequent rating studies (see secction Studies 6–8 for details). The observed subject-level correlations between the ratings of richer and poorer faces in studies 1–5 were indeed around 0.5 (mean $r = 0.53$, median $r = 0.62$).

Materials. We used 36 images of faces. To arrive at this final set of faces, we began with 50 photographic images—25 self-identified and universally perceived as Black men and 25 self-identified and universally perceived as white men—taken from a standardized face stimulus set[29]. The faces were of amateur actors with no distinct facial hair, accessories or make-up. We combined each face once with richer clothes selected from images displayed by online modern US-style apparel retailers, and again with poorer clothes, selected from product images of online vintage clothing stores. The 50 faces, combined with richer and poorer clothes, were rated (How rich or poor does this person look?) by a group of judges ($n = 31$, mean age = 22.8 years, 24 female, 7 male) before the studies. Each judge saw each face once, combined with either richer or poorer clothing, counterbalanced across participants. Based on those ratings, we selected the 18 black and 18 white face–clothing pairs (see Supplementary Fig. 1 for stimuli) that showed the largest rich–poor rating differences ($t > 3.00$, $P < 0.001$).

To ascertain the subtlety of the manipulation (that is, that no clothes portrayed excessive wealth or poverty), we collected verbal descriptions of the clothing items in the stimulus set. A separate group of judges ($n = 15$, mean age = 37.13 years, 7 female, 8 male) were asked to provide three descriptive words for each item of clothing presented with no face. The descriptions showed only mild differences (see Fig. 1d): extremely positive or negative words were rare, and negative descriptions, although infrequent, were found for both richer (for example, ugly, cheap, dull, rugged) and poorer clothes (for example, ugly, cheap, dull, dated). The words 'rich' or 'poor', or their synonyms (determined by WordNet[30], a large English-language lexical database) occurred exactly once out of total 4,725 words. To further test for potential differences across clothing types, we computed the 'valence score' of all descriptive words. The valence score was computed using R package affectr[31], a validated English-language dictionary developed on fastText, a text analysis tool that represents words as the sums of vectors[32]. The affectr dictionary spans two million words and provides the emotional valence score of any given word on a continuous scale (for example, adventurous 0.54, comfortable 0.38, tweed −0.03, unstylish −0.11, ugly −0.64—sample words from the description responses). This allowed us to calculate the levels of positivity and negativity expressed in the clothes' descriptions, not just the frequency of the positive versus negative words. Before the analysis, we corrected all typos and removed unintelligible responses as well as names (for example, 'Mr Rogers'). We found no evidence of difference in valence across clothes descriptions (formal richer, $n = 1{,}617$, mean = 0.13, s.d. = 0.25; informal richer, $n = 1{,}663$, mean = 0.15, s.d. = 0.30; poorer, $n = 1{,}616$, mean = 0.13, s.d. = 0.31; $F(2,4893) = 1.57$, $P = 0.208$). In summary, clothing showed no apparent difference in general positivity or negativity, despite its ensuing significant effect on perceived competence.

Procedure. Participants were told that the researchers were interested in how people evaluate others' appearance and were encouraged to rely on their 'gut feeling'. Following two practice trials with faces that were not part of the experimental set, each participant rated the 36 target faces in random order, each occurring exactly once. For each participant, half the faces were presented with richer upper-body clothing and the other half with poorer clothing (Fig. 1). Clothing was not mentioned, and clothing type and presentation duration were counterbalanced across raters. Faces were presented at three durations, randomly assigned via a Latin square design: in each duration, every face occurred once with richer clothes and once with poorer clothes, randomly divided between two groups and counterbalanced across participants.

Following a 500-ms fixation point presentation, each face was presented for either 129, 553 or 1,059 ms on an 85-Hz cathode ray tube (CRT) monitor (studies 1, 3a,b, 4 and 5, which took place in a behavioural study laboratory setting at a university) or for about 300, 600 or 1,100 ms on a liquid crystal display (LCD) monitor (studies 2 and 3c, which took place outside the laboratory on portable laboratory computers). Presentation on the CRT monitors occurred in multiples of 11.75 ms, the duration of a single frame on an 85-Hz screen (that is, 1/85 × 1,000). On the LCD monitors, presentation durations were increased because display time cannot be controlled precisely: stimulus presentation as brief as 129 ms, for instance, would not result in an exact 129-ms presentation on an LCD. Each face presentation was followed by a 1-s greyscale mask. Following the mask, participants were asked, "How competent is this person?" on a scale ranging from 1 (Not at all competent) to 9 (Extremely competent). The question and the scale remained on the screen until the participant responded with a keyboard press (see Fig. 1a).

Study 6. *Participants.* We observed a small main effect of stimulus duration in two studies, and a fairly modest clothing-status effect size, ranging from $\eta_G^2 = 0.07$ to $\eta_G^2 = 0.37$ in studies 1–5 (Fig. 2 and Supplementary Table 1). This raises the possibility that earlier studies might have failed to detect a main effect of stimulus duration, or an interaction with clothing status. To test the robustness of our effects, we substantially enlarged our sample size. Power analysis using G*Power indicated that 199 participants would afford a power of 0.8 for a 2 × 2-interaction across within-subject variables (Status × Duration) with an extremely small effect size ($\eta^2 = 0.01$). We did not consider a three-way interaction (Status × Duration × Race), because we had no relevant hypothesis and any effect would be hard to interpret. Participants recruited on Amazon Mechanical Turk (MTurk) participated for payment ($n = 405$, mean age = 38.16 years; 178 female, 224 male, 3 other). To maintain precision in stimulus duration (over which we had limited control), 202 participants who saw the stimuli for durations shorter or longer than specified by the study were excluded from the main analysis. Three participants who used a single rating response throughout were also excluded. Participants were recruited until the intended sample size was reached ($n = 200$, mean age = 38.16 years, 86 female, 111 male, 3 other). Inclusion of all 405 participants' responses yields the same results (see Supplementary Results).

Materials. We used the same set of 36 faces from study 5 (formal attire removed).

Procedure. Participants followed the same instructions as in study 3. They were explicitly asked to ignore the clothing when rating the competence of the target individuals. To simplify the study design, images were presented at two durations rather than three, and participants performed the study on their computers. Within each presentation duration, every face occurred once with richer clothes and once with poorer clothes, randomly divided between two groups and counterbalanced across participants. In each trial, following a 500-ms fixation point, each face was presented for either 250 or 750 ms. Each face presentation was followed by a 1-s phase-scramble mask crafted specifically for each face image, as in studies 3–5.

To assure control over stimulus presentation, we used Inquisit 5 Web. Before launching an experiment, the Inquisit Web system downloads the experiment on the participant's computer and records the computer's native timing information to obtain stimuli presentation and response times in milliseconds. Based on those time records, we excluded from the formal analysis all participants who were presented with any stimuli for an inaccurate time duration (see Participants, above).

Study 7. *Participants.* Having conducted power analyses based on participant-level means, s.d.s in competence ratings (Fig. 2 and Supplementary Table 1) and the correlation between richer and poorer face ratings in study 6 ($r = 0.84$), we decided to collect data from 50 participants; collection was stopped after we obtained about $n = 50$. We did not consider an interaction effect because we did not find any in study 6, which had a sample size sufficiently large to detect any existing interaction. Paid participants were recruited on MTurk ($n = 52$, mean age = 34.25 years, 19 female, 32 male, 1 other). Two participants who used a single rating response throughout were excluded from further analysis ($n = 50$, mean age = 34.20 years, 19 female, 31 male).

Materials. We used a new set of 80 person images that featured all combinations of eight face images (four black, four white) and ten clothing images (five poorer, five richer). To generate the new stimuli, we randomly chose eight faces used previously. We then selected the five richer and five poorer clothes, excluding jackets, ties and dress shirts, that elicited, respectively, the highest and the lowest mean competence ratings in study 5.

Procedure. Each participant saw all eight faces (four black, four white), four of which were randomly paired with richer clothing and four with poorer clothing, evenly across face races. Stimuli were randomly assigned to participants. As in studies 3 and 6, participants were told to ignore the clothes. In each trial, following

a 500-ms fixation point, each face was presented for 750 ms. As in studies 3–6, a phase-scramble mask customized for each face image followed the stimulus.

Study 8. *Participants.* Having conducted power analyses based on participant-level means, s.d.s of competence ratings (Fig. 2 and Supplementary Table 1) and the correlation between richer and poorer face ratings in study 7 ($r = 0.63$), we decided to collect data from about 60 participants; collection was stopped after we obtained $n = 63$. Study 8 used the same stimuli and a procedure similar to that in study 7—it differed only in the 'incentive' provision and a slightly revised scale (see Procedure, below). Paid participants were recruited on MTurk ($n = 63$, mean age = 33.53 years, 25 female, 38 male).

Materials. We used the same 80 images (8 faces × 10 clothes) as in study 7.

Procedure. As in study 7, each participant saw all eight faces (four black, four white), four randomly paired with poorer clothing and four with richer clothing. Participants were instructed to rate the competence of target faces while ignoring the clothes, and told, "the participant whose ratings are the most accurate will receive, in addition to their standard pay, an additional US$100 reward". To suppress the temptation to cheat (by trying to simulate the viewing conditions of the reference group), accuracy was defined somewhat vaguely ("(another participant) group rated these faces under slightly different conditions, without the same clothes. Your accuracy will be determined by how close your ratings are to those of people who saw the faces under those different conditions"). The monetary reward was mentioned twice before the main task, with the first incentive instruction slide staying up for 20 s before participants were able to proceed.

As in previous studies, on each trial participants were asked, "How competent is this person?" The stimulus remained until the participant responded. To allow for more fine-grained responses, participants used a mouse click-and-drag on a slide bar scale ranging from 1 (Not at all competent) to 100 (Extremely competent) before clicking to continue to the next trial.

A separate group of MTurk participants ($n = 23$, mean age = 32.52 years, 9 female, 14 male) rated the same faces, cropped around the neck, showing no clothes. The procedure was the same as in the main study, except that no extra reward was offered and faces were presented twice so as to diminish measurement error. We excluded three participants whose test-retest reliability (calculated by correlating faces' repeated ratings) was ≤0. Using the remaining data ($n = 20$, mean age = 32.75 years, 9 female, 11 male), we calculated each face's mean ratings to compute its 'pure' perceived competence. We then Pearson-correlated the eight faces' pure ratings with those of each participant (participants on average performed poorly—median $r = 0.20$). The participant with the highest coefficient ($r = 0.90$) received the US$100 reward.

Study 9. *Participants.* Power analyses based on a between-subject ANOVA design with a large main effect ($d = 0.8$) suggested a total of 30 participants. We aimed for twice that number and decided to recruit about 60 subjects (30 for the warning condition and 30 for the no-warning condition). Data collection was stopped after we obtained $n = 64$. Princeton University undergraduates and community members participated for either course credit or for payment ($n = 64$, mean age = 19.87 years, 45 female, 18 male, 1 other).

Materials. We averaged each face's competence ratings across richer and poorer clothing conditions in studies 1–4, to obtain each face's 'average competence'. The eight faces with the highest average competence scores were categorized as high competence, the eight in the middle of the range as medium competence and the eight with lowest average competence scores as low competence. Ratings across competence levels were significantly different ($t > 10.70$, $P < 0.001$).

Procedure. Each participant saw all eight faces within a competence level, four arbitrarily paired with richer clothing and four with poorer clothing, counterbalanced across participants. This yielded 16 binary choices within competence level (4 faces × 4 faces) and 48 choice trials per participant overall (16 face pairs × 3 competence levels), each trial presenting a choice between two faces of equal competence level, one with richer and one with poorer clothing (the logic underlying this construction of stimuli was not known to participants).

Each choice trial was displayed for 1 s around the vertical centre of the screen. The horizontal distance between the centres of the faces was about 410 pixels or 15 visual degrees. The display of the faces was followed by a question, "Which face looked more competent?", which remained on the screen until the participant responded with a keyboard press. Following two practice trials with faces that were not part of the experimental set, a random one-half of the participants ($n = 32$, mean age = 19.87 years, 22 female, 10 male) received the following warning text: "One of the things that have been found to affect our judgements of competence are the clothes that people wear. When we see people wearing certain clothes, we think of them as more competent. There is, however, absolutely no evidence that a person's clothes are related to their actual competence. So please keep in mind: the clothes a person wears may not reflect how competent they are". The remaining participants ($n = 32$, mean age = 19.88 years, 23 female, 8 male, 1 other) did not receive this warning.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All data and stimuli are available on Open Science Framework: https://osf.io/v2j43/.

Received: 12 October 2018; Accepted: 4 November 2019;
Published online: 9 December 2019

References

1. Ballew, C. C. & Todorov, A. T. Predicting political elections from rapid and unreflective face judgments. *Proc. Natl Acad. Sci. USA* **104**, 17948–17953 (2007).
2. Graham, J. R., Harvey, C. R. & Puri, M. A corporate beauty contest. *Manag. Sci.* **63**, 3044–3056 (2017).
3. Rule, N. O. & Ambady, N. The face of success: inferences from chief executive officers' appearance predict company profits. *Psychol. Sci.* **19**, 109–111 (2008).
4. Stoker, J. I., Garretsen, H. & Spreeuwers, L. J. The facial appearance of CEOs: faces signal selection but not performance. *PLoS One* **11**, e0159950 (2016).
5. Todorov, A. T., Mandisodza, A. N., Goren, A. & Hall, C. C. Inferences of competence from faces predict election outcomes. *Science* **308**, 1623–1626 (2005).
6. Bar, M., Neta, M. & Linz, H. Very first impressions. *Emotion* **6**, 269–278 (2006).
7. Borkenau, P., Brecke, S., Möttig, C. & Paelecke, M. Extraversion is accurately perceived after a 50-ms exposure to a face. *J. Res. Personal.* **43**, 703–706 (2009).
8. Willis, J. & Todorov, A. T. First impressions: making up your mind after a 100-ms exposure to a face. *Psychol. Sci.* **17**, 592–598 (2006).
9. Todorov, A. T., Olivola, C. Y., Dotsch, R. & Mende-Siedlecki, P. Social attributions from faces: determinants, consequences, accuracy, and functional significance. *Annu. Rev. Psychol.* **66**, 519–545 (2015).
10. Aviezer, H. et al. Angry, disgusted, or afraid? Studies on the malleability of emotion perception. *Psychol. Sci.* **19**, 724–732 (2008).
11. Aviezer, H., Trope, Y. & Todorov, A. T. Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science* **338**, 1225–1229 (2012).
12. Freeman, J. B. et al. The neural basis of contextual influences on face categorization. *Cereb. Cortex* **25**, 415–422 (2015).
13. Freeman, J. B., Penner, A. M., Saperstein, A., Scheutz, M. & Ambady, N. Looking the part: social status cues shape race perception. *PLoS One* **6**, e25107 (2011).
14. Santamaría-García, H., Burgaleta, M. & Sebastián-Gallés, M. Neuroanatomical markers of social hierarchy recognition in humans: a combined ERP/MRI study. *J. Neurosci.* **35**, 10843–10850 (2015).
15. Ratcliff, N. J., Hugenberg, K., Shriver, E. R. & Bernstein, M. J. The allure of status: high-status targets are privileged in face processing and memory. *Personal. Soc. Psychol. Bull.* **37**, 1003–1015 (2011).
16. Fiske, S. T., Cuddy, A. J. C., Glick, P. & Xu, J. A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *J. Personal. Soc. Psychol.* **82**, 878–902 (2002).
17. Durante, F. et al. Nations' income inequality predicts ambivalence in stereotype content: how societies mind the gap. *Br. J. Soc. Psychol.* **52**, 726–746 (2013).
18. Mattan, B. D., Kubota, J. T. & Cloutier, J. How social status shapes person perception and evaluation: a social neuroscience perspective. *Perspect. Psychological Sci.* **12**, 468–507 (2017).
19. Carroll, J. M. & Russell, J. A. Do facial expressions signal specific emotions? Judging emotion from the face in context. *J. Personal. Soc. Psychol.* **70**, 205–218 (1996).
20. Fiske, S. T. *Envy Up, Scorn Down: How Status Divides Us* (Russell Sage Foundation, 2011).
21. WHO Europe. *Poverty, Social Exclusion and Health Systems in the WHO European Region* (WHO Regional Office for Europe, 2010).
22. Mani, A., Mullainathan, S., Shafir, E. & Zhao, J. Poverty impedes cognitive function. *Science* **341**, 976–980 (2013).
23. Mullainathan, S. & Shafir, E. *Scarcity: Why Having Too Little Means So Much* (Macmillan, 2013).
24. Bargh, J. A. in *Dual-process Theories in Social Psychology* (eds Chaiken, S. & Trope, Y.) 361–382 (Guilford Press, 1999).
25. Phalet, K. & Poppe, E. Competence and morality dimensions of national and ethnic stereotypes: a study in six Eastern-European countries. *Eur. J. Soc. Psychol.* **27**, 703–723 (1997).
26. Kraus, M. W., Park, J. W. & Tan, J. J. X. Signs of social class: the experience of economic inequality in everyday life. *Psychol. Sci.* **12**, 422–435 (2017).
27. Darley, J. M. & Gross, P. H. A hypothesis-confirming bias in labeling effects. *J. Personal. Soc. Psychol.* **44**, 20–33 (1983).

28. Faul, F., Erdfelder, E., Lang, A.-G. & Buchner, A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* **39**, 175–191 (2007).
29. Ma, D. S., Correll, J. & Wittenbrink, B. The Chicago face database: a free stimulus set of faces and norming data. *Behav. Res. Methods* **47**, 1122–1135 (2015).
30. Miller, G. A. WordNet: a lexical database for english. *Commun. ACM* **38**, 39–41 (1995).
31. Thornton, M. A. affectr: R package for 3-D sentiment analysis. R version 3.0 https://github.com/markallenthornton/affectr(2018).
32. Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. Enriching word vectors with subword information. *Trans. Assoc. Comp. Ling.* **5**, 135–146 (2017).

Acknowledgements

We thank B. Labbree, R. Drach, S. Anjur-Dietrich, A. Duker and K. Solomon for help with running the experiments. This work was supported by the National Science Foundation (award no. 1426642) and the Sloan Foundation (grant no. 2014-6-16). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

E.S. and A.T. devised the study concept. All authors designed the experiments and wrote the manuscript. D.O. collected and analysed data.

Competing interests

Authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41562-019-0782-4.

Correspondence and requests for materials should be addressed to D.O.

Peer review information Primary handling editor: Aisha Bradshaw

Reprints and permissions information is available at www.nature.com/reprints.

nature research

Corresponding author(s): DongWon oh

Last updated by author(s): Sep 15, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Experiments in studies 1-5 and 9 were built and run on E-Prime 2.0. Experiments in studies 6 and 7 were written and run on Inquisit 5 Web. Two experiments in study 8 were built and run on Qualtrics. Two experiments for stimulus preparation (described in Methods) were written in and run on Python 2.7. One online experiment for stimulus description data collection (described in Methods) was written in and run on JavaScript with the jsPsych library. One experiment for the face-only stimulus rating (described in Methods) was built and run on Qualtrics.
Data analysis	All analyses were carried out in the R 3.4 Environment with the dplyr, plyr, reshape2, ez, stringr, tm, SnowballC, wordcloud, RColorBrewer, pwr, lme4, lmerTest, lsmeans, geepack, FactoMineR, factoextra, affectr, and ggplot2 packages, or in Python 2/3 Environment with numpy, pandas, and fastText.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data analyzed and all stimuli used in the research are available on Open Science Framework: https://osf.io/v2j43/

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	All data that were collected and analyzed were quantitative. Human participants responded to visual stimuli presented to them on the computer screen after giving informed consent.
Research sample	For studies 1-5 and 9 and two experiments for stimulus preparation, convenience samples of Princeton students and community members and visitors of a shopping mall in central New Jersey were recruited. For studies 6-8, the experiments for stimulus description data collection, and the experiment for the face-only stimulus rating (described in Methods), Amazon Mechanical Turk users were recruited. All participants were adults. These facts are clearly described in the main text.
Sampling strategy	Convenience samples of Princeton students and community members, visitors of a shopping mall in central New Jersey, and Amazon Mechanical Turk users were recruited (see Research Sample). For studies 1-5, power analysis using G*Power 3.1.9 based on a within-subject analysis of variance (ANOVA) design (η^2 > .26) with a moderate level of correlation between the within-subject measures (r = .5, Pearson correlation between "richer" and "poorer" face ratings calculated at the participant level), suggested over 20 participants. We decided to stop collecting data at natural stopping points, such as end of day, at somewhere between N = 30 and 60. The observed effect size in studies 1–5 proved smaller than initial estimates, and informed sample-size decisions in subsequent rating studies (see studies 6–8 Methods sections for details). The observed subject-level correlations between the ratings of "richer" and "poorer" faces in studies 1–5 were indeed around .5 (mean r = .53, median r = .62). We observed a small main effect of stimulus duration in two studies, and a fairly modest clothing status effect size, ranging from ηG^2 = .07 to ηG^2 = .37 in studies 1–5. This raises the possibility that earlier studies might have failed to detect a main effect of stimulus duration, or an interaction with clothing status. So, in study 6, to test the robustness of our effects, we substantially enlarged our sample size. Power analysis using G*Power indicated that 199 participants would afford a power of .8 for a 2 × 2 interaction across within-subject variables (Status × Duration) with an extremely small effect size (η^2 = .01). In study 7, having conducted power analyses based on participant-level means, s.d.'s in competence ratings, and the correlation between "richer" and "poorer" face ratings in study 6 (r = .84), we decided to collect data from 50 participants, and collection was stopped after we obtained about N = 50. We did not consider an interaction effect, because we did not find any in study 6, which had a sample size large enough to detect any existing interaction. In study 8, having conducted power analyses based on participant-level means, s.d.'s of competence ratings and the correlation between "richer" and "poorer" face ratings in study 7 (r = .63), we decided to collect data from about 60 participants, and collection was stopped after we obtained N = 63. Study 8 used the same stimuli and a similar procedure as study 7. In study 9, power analyses based on a between-subject ANOVA design with a large main effect (d = .8) suggested 30 total participants. We aimed for twice that and decided to recruit about 60 subjects (30 subjects for the warning condition and 30 subjects for the no-warning condition). Data collection was stopped after we obtained N = 64.
Data collection	Participants responded to visual stimuli presented to them on the computer screen. In all cases, a researcher was not in the same space with participants during the experiment except for studies 2 and 3c, in which the studies were run in a semi-open space in a shopping mall and a researcher sat across each participant during the experiment. Even in these two studies, the researcher could not see what was happening on the computer screen, nor did they engage in any conversation with the participants during the study.
Timing	Data were collected from during the following time periods: July - November 2015, May - June 2016, November - December 2016, and January - March 2019.
Data exclusions	In studies 1-5 and 9 (in-person participation), no data points were excluded except for the data with incomplete responses (presumably, the participants left without noticing that they didn't complete their task). In studies 6-8 (online participation), no data points were excluded except for the data with identical single rating response throughout all trials. Additionally, in study 6, data points from participants who were presented with any stimulus for an inaccurate amount of time were excluded; however, the results with all participants' data are reported in Supplementary Results. In any rate, all data exclusion criteria and the number of excluded participants per study are reported in details in the main text.
Non-participation	No in-person participants dropped out or declined participation, to our knowledge. We do not have access to how many online participants dropped out or declined participation, because the study advertisements were publicly available and participants could voluntarily quit the experiments any time in the middle of participation on their local computer, and this was not logged.
Randomization	Studies 1-6 and 9 assigned each subject in either of the between-subject stimulus group (one of the six conditions in a Latin square design in studies 1-5; one of the four conditions in a Latin square design in study 6; one of the two conditions (warning/no warning) in study 9). Studies 7-8 randomly selected eight stimuli per subject.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	See above. Demographic data, which include participant age, self-identified race, and sex, are made available alongside with participant response data for every study for any interested reader on Open Science Framework: https://osf.io/v2j43/).
Recruitment	Convenience samples of Princeton students and community members and visitors of a shopping mall in central New Jersey were recruited in studies 1-5 and 9 and two experiments for stimulus preparation. Amazon Mechanical Turk users were recruited in studies 6-8 and the experiments for stimulus description data collection. We do not expect any biases induced by a demographic variable that would seriously impact our main results, although according to recent empirical data regarding online study participants, these samples might be more highly educated and consist of more individuals from middle socioeconomic statuses than the general population in the US.
Ethics oversight	Princeton University IRB has approved the study protocol (protocol no. 7212).

Note that full information on the approval of the study protocol must also be provided in the manuscript.