Cognition 230 (2023) 105288

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Contextual modulation of appearance-trait learning

Harriet Over [a,*], Ruth Lee [a], Jonathan Flavell [a], Tim Vestner [c], Richard Cook [a,b]

[a] Department of Psychology, University of York, York, UK
[b] Department of Psychological Sciences, Birkbeck, University of London, London, UK
[c] School of Psychology and Therapeutic Studies, Leeds Trinity University, UK

ARTICLE INFO

Keywords:
First impressions
Trait-inference mapping
Renewal
Contextual modulation
Appearance-trait learning

ABSTRACT

When we encounter a stranger for the first time, we spontaneously attribute to them a wide variety of character traits based on their facial appearance. There is increasing consensus that learning plays a key role in these first impressions. According to the Trait Inference Mapping (TIM) model, first impressions are the products of mappings between 'face space' and 'trait space' acquired through domain-general associative processes. Drawing on the associative learning literature, TIM predicts that first-learned associations between facial appearance and character will be particularly influential: they will be difficult to unlearn and will be more likely to generalise to novel contexts than appearance-trait associations acquired subsequently. The study of face-trait learning *de novo* is complicated by the fact that participants, even young children, already have extensive experience with faces before they enter the lab. This renders the study of first-learned associations from faces intractable. Here, we overcome this problem by using Greebles – a class of novel synthetic objects about which participants had no previous knowledge or preconceptions – as a proxy for faces. In four experiments (total $N = 640$) with adult participants we adapt classic AB-A and AB-C renewal paradigms to study appearance-trait learning. Our results indicate that appearance-trait associations are subject to contextual control, and are resistant to counter-stereotypical experience.

Humans spontaneously attribute a wide range of traits to strangers based on their facial appearance (Todorov, Olivola, Dotsch, & Mende-Siedlecki, 2015; Zebrowitz, 2017). These first impressions include judgements about apparent likeability, honesty, intelligence, competence, and aggression. While a wealth of spontaneous attributions have been studied, observers' judgements appear to load on two principal dimensions commonly described as trustworthiness and dominance (Oosterhof & Todorov, 2008). While our first impressions are typically inaccurate, they can have serious real-world consequences (Olivola, Funk, & Todorov, 2014; Rule, Krendl, Ivcevic, & Ambady, 2013). For example, first impressions are thought to affect criminal sentencing (Wilson & Rule, 2015), financial decisions (Hooper et al., 2018) and voter preferences (Todorov, Mandisodza, Goren, & Hall, 2005). More recently, it has been shown that first impressions also influence how adults interact with children, with potentially long term consequences for developmental outcomes (Thierry & Mondloch, 2021).

Some first impressions – so-called consensus impressions – are widely shared within a culture or community. For example, many individuals in the U.K. and U.S. attribute naivety and trustworthiness to faces with large eyes and round features (Zebrowitz McArthur & Berry, 1987; Zebrowitz & Montepare, 1992). The same observers tend to judge short, squat faces to be more aggressive than faces that are tall and thin (Geniole, Molnar, Carré, & McCormick, 2014; Stirrat & Perrett, 2010), and attribute to beautiful faces a range of positive traits including trustworthiness, intelligence, and generosity (Dion, Berscheid, & Walster, 1972; Eagly, Ashmore, Makhijani, & Longo, 1991). Other face-trait judgements are idiosyncratic – they differ between individuals living within the same culture (Sutherland et al., 2020). There is growing evidence that learning plays an important role in the emergence of both types of first impression (Cook, Eggleston, & Over, 2022). This view is supported by several lines of evidence.

Studies conducted in the lab indicate that adults readily acquire new first impressions following periods of training (Chua & Freeman, 2022; Cone, Flaharty, & Ferguson, 2021; Falvello, Vinson, Ferrari, & Todorov, 2015; FeldmanHall et al., 2018; Lee, Flavell, Tipper, Cook, & Over, 2021; Todorov & Uleman, 2002; Verosky & Todorov, 2010). For example, training procedures might pair unfamiliar faces with positive (e.g., "Gave his balloon to a child who had let hers go") or negative (e.g., "Stole money and jewellery from the relatives he was living with") behaviours. At test, faces that have previously been paired with positive

* Corresponding author at: Department of Psychology, University of York, Heslington, York YO10 5DD, UK.
E-mail address: harriet.over@york.ac.uk (H. Over).

https://doi.org/10.1016/j.cognition.2022.105288
Received 12 November 2021; Received in revised form 25 August 2022; Accepted 14 September 2022
Available online 24 September 2022

behaviours are judged to be more trustworthy than those paired with negative behaviours (Cone et al., 2021; Falvello et al., 2015; Todorov & Uleman, 2002). Crucially, newly acquired face-trait associations generalise to novel faces of similar appearance (Chua & Freeman, 2022; FeldmanHall et al., 2018; Lee et al., 2021; Verosky & Todorov, 2010). These findings suggest that our first impressions of strangers are likely influenced by our knowledge of familiar others and their traits.

Consistent with a learning account, there are systematic cultural differences in first impressions (Chen, Jing, Lee, & Bai, 2016; Jones et al., 2021; Lakshmi, Wittenbrink, Correll, & Ma, 2021; Over, Eggleston, & Cook, 2020a; Sofer et al., 2017; Sutherland et al., 2018; Walker, Jiang, Vetter, & Sczesny, 2011; Zebrowitz et al., 2012). For example, in so-called WEIRD cultures (Western, Educated, Industrialized, Rich, Democratic), straight white teeth are associated with attractiveness, social status, and a host of other positive characteristics (Dion et al., 1972; Eagly et al., 1991). Positive associations with straight white teeth are not culturally universal, however. Various forms of dental modification including the deliberate creation of gaps through the removal of teeth, the filing of teeth to modify their shape, and teeth-blackening are common in other cultures (Over, Eggleston, & Cook, 2020b). Within these cultures, dental modification enhances impressions of the individual by signalling group membership, social status, and desirable character traits (e.g., sobriety and restraint).

Convergent developmental research suggests that consensus impressions emerge around 3–5 years of age (Cogsdill, Todorov, Spelke, & Banaji, 2014), and continue to develop throughout childhood, showing adult-like patterns between 10 and 13 years of age (Siddique et al., 2022). Some have argued that the emergence of consensus judgements at around 3–5 years is early enough to preclude a social learning account of their origin (Cogsdill et al., 2014; Ewing, Sutherland, & Willis, 2019). Contrary to this view, however, the attribution of intelligence to those who wear glasses (Fleischmann, Lammers, Stoker, & Garretsen, 2019) also emerges at this point in development (Eggleston, Flavell, Tipper, Cook, & Over, 2021). Given that glasses have been in existence for less than 800 years (Ilardi, 2007), this trait inference cannot possibly be a genetic adaptation; rather, it must be learned, either through exposure to cultural messages or via first-hand experience (Over & Cook, 2018).

More recently, attention has turned to the question of how first impressions are learned. A particular challenge for learning accounts is to explain how many people within the same culture can acquire similar but inaccurate first impressions. The Trait Inference Mapping (TIM) account asserts that first impressions are learned through exposure to cultural messages about the appearance of certain types of characters (e.g., heroes and villains, jocks and geeks, leaders and followers). By exposing many members of a community to the same (erroneous) face-trait associations, cultural messages have the potential to explain the emergence of widely-held but inaccurate impressions (Cook et al., 2022; Cook & Over, 2020; Over et al., 2020b; Over & Cook, 2018). Face-trait stereotypes may be learned through exposure to film and television, illustrated storybooks, art, and iconography. For example, action films (e.g., those from the Superman and James Bond franchises) and animated movies (e.g., those produced by Disney and Pixar) have taught millions of viewers worldwide clear lessons about the appearance of heroes and villains; for example, villains are more likely than heroes to be depicted with pallid complexion, facial disfigurements, and baldness (Croley, Reese, & Wagner, 2017). Similar messages may also be conveyed by cultural rituals (Over et al., 2020b) and visual propaganda (Keen, 1991).

According to the TIM framework, mappings between 'face space' and 'trait space' are acquired through correlated face-trait experience; i.e., learning episodes where individuals' facial appearance is predictive of their traits and characteristics (Cook et al., 2022; Cook & Over, 2020; Over et al., 2020b; Over & Cook, 2018). This account is heavily influenced by the associative learning literature. Specifically, TIM hypothesises that correlated face-trait experience induces face-trait mappings through domain-general associative processes, such as those responsible for conditioning phenomena. If this view is correct, then insights from the literature on associative learning may aid efforts to understand the development and acquisition of first impressions. In the present study, we explored the possibility that appearance-trait learning exhibits evidence of renewal.

In renewal paradigms, participants first learn that a stimulus predicts one outcome (e.g., a reward) in one context, before being transferred to another context where they learn that the same stimulus no longer predicts that outcome. Crucially, when returned to the original context (AB-A renewal), or introduced to a novel context (AB-C renewal), participants once again expect the stimulus to predict the original outcome (Bouton & King, 1983; Nelson, Sanjuan, Vadillo-Ruiz, Pérez, & León, 2011; Peck & Bouton, 1990). Renewal effects are important because they demonstrate that what we learn first about a stimulus can be hard to unlearn. Rather than over-write existing associations, subsequent learning in a different context appears to establish parallel associations that manifest in that particular learning environment. Insights from renewal have proved important in understanding human behaviour; for instance, why people addicted to drugs are prone to relapse when they leave clinical settings and return to their home environment (e.g., Bouton, 2002).

The study of appearance-trait learning *de novo* is complicated by the fact that participants, even young children, already have extensive experience with faces before they enter the lab. This renders the study of first-learned associations from faces challenging. In this study, we circumvent this problem by studying appearance-trait learning using Greebles – a class of novel synthetic objects (Gauthier & Tarr, 1997; Gauthier, Tarr, Anderson, Skudlarski, & Gore, 1999; Gauthier, Williams, Tarr, & Tanaka, 1998). Like faces, Greebles can be categorised into different 'families' based on their parts and configuration. The presence of this inter-exemplar structure makes them an ideal proxy for faces in studies of appearance-trait learning. Recent research suggests that participants quickly learn that individual Greebles have particular trait profiles and will generalise that learning to other Greebles from the same family (Lee et al., 2021).

In the present study, we used the Greeble paradigm to examine whether appearance-trait learning exhibits evidence of renewal. In total, we describe four experiments, each of which was pre-registered. Our preregistration documents and the data for each experiment can be found in the OSF repository at https://osf.io/a65fc/.

1. Experiment 1 (AB-A)

In our first experiment, we sought evidence of a classic AB-A renewal effect. Participants first observed two families of Greebles in a particular context. Whereas one family of Greebles was paired with trustworthy behaviours in this context, the other family was paired with untrustworthy behaviours. Participants then encountered the same Greebles in a second context. In this context, the contingencies were reversed - the Greebles previously paired with trustworthy behaviours were paired with untrustworthy behaviours. At test, participants were presented with 12 *novel* Greebles from the two families, similar to those presented during training, and were asked to judge their apparent niceness. Crucially, this test took place in the first learning context.

If new appearance-trait learning that contradicts old appearance-trait learning overwrites the original learning (the unlearning account), one might expect little or no systematic bias in participants' ratings of the novel Greebles. However, if new appearance-trait learning that contradicts old appearance-trait learning creates a second set of parallel associations that manifest selectively in the learning environment (the context-specific learning account), one might expect evidence of AB-A renewal; i.e., Greebles from the family paired with trustworthy behaviours in the first context should be judged more trustworthy than Greebles from the family presented as trustworthy in the second context.

2. Methods

2.1. Participants

One hundred and sixty participants (M_{age} = 35.99 years, SD_{age} = 12.50 years, range: 18–73 years, 78 female, 81 males, 1 prefer-not-to-say) were recruited through http://www.prolific.co. Participants received a small honorarium. Participants had to be at least 18, speak English as a native language, and reside in the UK. Thirteen further participants were tested but were replaced having failed one or both attention checks. Sample size was determined by an a priori power analysis. Power and alpha were set at 0.8 and 0.05. Pilot data yielded a Cohen's *d* of 0.3. This analysis indicated that a sample size of 160 provided adequate power for a paired *t*-test.

2.2. Materials

We used Greebles from two different families (Fig. 1). Three Greebles from each family were presented during training. Six novel Greebles from the same two families were presented at test. The experiments were conducted online using Gorilla (Anwyl-Irvine, Massonnié, Flitton, Kirkham, & Evershed, 2020).

2.3. Design

Participants completed a training procedure during which they encountered Greebles from two different families in two different contexts - the Forest Planet and the Mountain Planet (Fig. 2a). On each of the training trials, participants were shown a pair of Greebles and asked to guess which acted in a trustworthy or untrustworthy manner. What differed between the training contexts was the contingency between Greeble family and trustworthy / untrustworthy behaviour. In the first context, there were 36 trials in which one family of Greebles was paired with trustworthy behaviour (e.g., in Context A, Family-1 is trustworthy and Family-2 is untrustworthy). In the second context, there were 36 training trials in which the contingency was reversed (e.g., in Context B, Family-1 is untrustworthy and Family-2 is trustworthy). The test phase took place in Context A after a brief distractor task. Participants saw 12 novel Greebles from the same two families that were perceptually similar to those encountered in the first two phases. Participants rated how nice they seemed on a scale from −50 (not at all nice) to 50 (extremely nice), in keeping with common practice in this field (e.g., Cogsdill et al., 2014; Lee et al., 2021).

2.4. Procedure

The first context was introduced by a screen saying: "This is the [Forest/Mountain] planet. Greebles live on the [Forest/Mountain] Planet". Participants completed 36 training trials in the first context. Training trials depicted two Greebles side-by-side, one from each family, on the appropriate background below a text description of an action. Eighteen of the training trials involved trustworthy behaviours: e.g., 'One of these two Greebles helped another Greeble to find a lost pet and made the other Greeble feel pleased'. The other 18 involved untrustworthy behaviours: e.g., 'One of these two Greebles refused to share a boat with another Greeble, so the other Greeble got stuck on an island'. Participants were asked to click on the Greeble they thought performed the action. Regardless of whether or not participants answered correctly, a green tick was displayed over the correct Greeble and a red cross over the incorrect Greeble. Training trials were almost identical in the second context, but crucially, the trustworthiness of the two Greeble families was reversed.

The Greeble family (1 or 2) associated with trustworthy behaviours in the first encountered context was counterbalanced. We also counterbalanced which context (Forest or Mountain) was presented first and which of two sets of events – closely matched for content – were presented in the first and in the second stages of training. Within the two stages of training, trials were presented in a random order.

An attention check was presented between the 13th and 14th trials of each training block. This check was structured in the same way as the training trials but, rather than describing an event, the text accompanying the picture of two Greebles asked participants to click on a specific Greeble (right or left). Participants saw a feedback screen indicating a correct or incorrect answer to the attention check.

Immediately following training, participants completed a distractor task measuring divergent thinking for 6 min. In this task, participants were asked to think of as many uses as possible for a brick (2 min), a blanket (2 min), and a pencil (2 min).

Participants then completed 12 test trials. Test trials were preceded by an image of the planet associated with the initial training context seen by the participant. This image was accompanied by the words "Now you are back on the [initial context] Planet". Each test trial presented a single Greeble in the centre of the display superimposed on the planet image. Participants rated Greebles from the same two families as those seen during training (6 novel exemplars from each family) on a scale ranging from 'Not at all' nice (−50) to 'Extremely' nice (+50). The order of the 12 test trials was randomized.

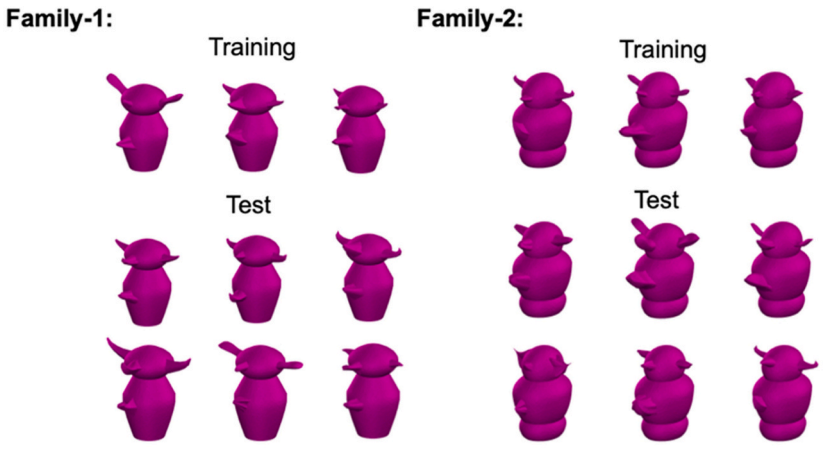

Fig. 1. Images of the Greebles from the two families used during training and at test in all experiments. All Greeble images are courtesy of Michael J. Tarr, Carnegie Mellon University, http://www.tarrlab.org/

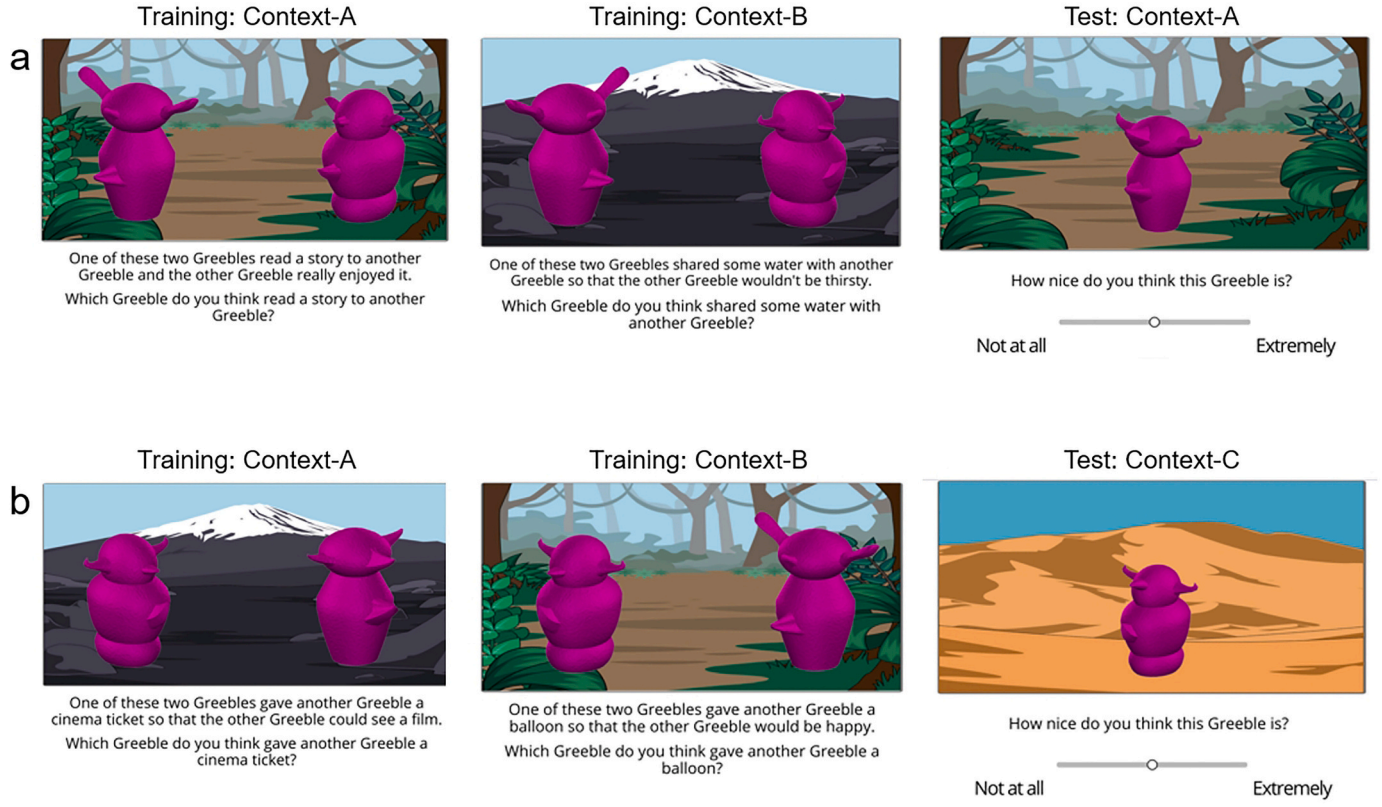

Fig. 2. (a) Example training and test trials from Experiment 1 (AB-A renewal) (a) Example training and test trials from Experiment 3 (AB-C renewal).

3. Results

3.1. Training performance

The proportion of correct responses seen across the first and second phase of the training procedure are plotted in Fig. 3a. In the first training phase, participants gave significantly more correct responses in the second half (trials 19–36, $M = 84.0\%$; $SD = 20.5\%$) than in the first half (trials 1–18, $M = 70.3\%$; $SD = 20.4\%$) [$t(159) = 10.30$, $p < .001$, $d = 0.81$]. This was also true of the second training phase: participants gave significantly more correct responses in the second half (trials 19–36, $M = 85.1\%$; $SD = 21.5\%$) than in the first half (trials 1–18, $M = 78.0\%$; $SD = 21.9\%$) [$t(159) = 7.10$, $p < .001$, $d = 0.56$]. Together, these results suggest discovery and learning of the Greeble feature rule in Context A, and re-learning of the flipped rule in Context B.

3.2. Test performance

For each participant, we computed average niceness ratings for the Greeble family who acted in a trustworthy manner in the context that participants encountered first and second. Following our pre-registered analysis plan, average trust ratings were analyzed using a paired *t*-test (Fig. 4a). Greebles from the family paired with trustworthy behaviours in the first context ($M = 15.20$; $SD = 20.18$) were rated as significantly nicer than were Greebles from the family paired with trustworthy behaviours in the second context ($M = -12.20$, $SD = 22.49$) [$t(159) = 8.72$, $p < .001$, $d = 0.69$]. This was the case even though both families of Greebles were paired with an equal number of trustworthy and untrustworthy behaviours overall.

The results from Experiment 1 demonstrate that a classic finding from the associative learning literature – AB-A renewal can be replicated using an appearance-trait learning paradigm. These results suggest that contradictory learning that occurs in Context B sets up parallel associations that exist alongside – i.e., that do not overwrite – the original associations acquired in Context A. Because the learning in Context B is context-specific, the initial learning is protected from subsequent unlearning. Thus, when participants are returned to the original learning context, participants' responses follow the first-learned contingency.

4. Experiment 2 (AB-B)

An unlearning account of the results from Experiment 1 might still be plausible if one assumes that the learning that occurred in Context B was weaker than that seen in Context A. If the original learning in Context A was only partially overwritten, this might explain why participants responded according to the Context A contingencies when returned to Context A. It is possible, for example, that participants paid less attention in Context B than in Context A due to boredom or fatigue, and thus required more than 36 training trials to fully unlearn the rule acquired in the first context.

In order to assess this possibility, we conducted a second experiment that was identical to the first, but this time participants were tested in Context B (AB-B). If individuals are learning rival sets of parallel associations that manifest selectively in different contexts, one would expect to see evidence of the Context A learning when individuals are tested in Context A, but also evidence of the Context B learning when individuals are tested in Context B. An unlearning account that assumes only one set of context-insensitive associations predicts similar effects when individuals are tested in Context A and Context B.

5. Methods

One hundred and sixty participants were recruited through http://www.prolific.co ($M_{age} = 39.20$ years, $SD_{age} = 12.23$ years, range: 18–65 years, 77 female, 83 male) using the same criteria as Experiment 1, and received a small honorarium. Ten further participants were tested but were replaced having failed one or both attention checks. The

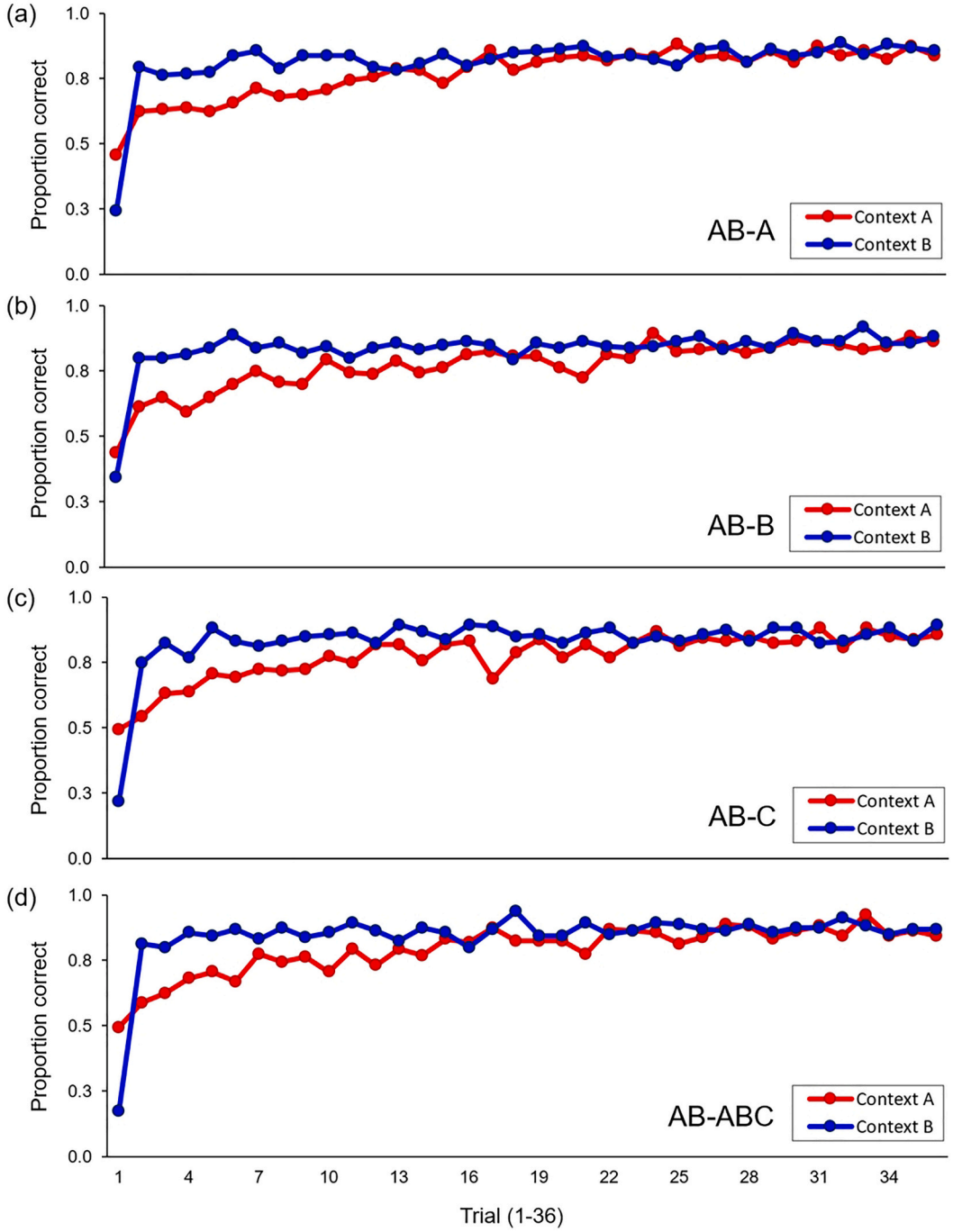

Fig. 3. Performance across the training trials in Experiment 1: AB-A (a), Experiment 2: AB-B (b), Experiment 3: AB-C (c), and Experiment 4: AB-ABC (d).

experimental design and procedure were identical to the first experiment, with the exception of the test phase which was conducted in Context B.

6. Results

6.1. Training performance

The proportion of correct responses seen across the first and second phase of the training procedure are plotted in Fig. 3b. The pattern of responding during training was very similar to that observed in the first experiment. In the first training block, participants gave significantly more correct responses in the second half (trials 19–36, $M = 83.1\%$; $SD = 21.6\%$) than in the first half (trials 1–18, $M = 71.2\%$; $SD = 19.0\%$) [$t(159) = 10.03$, $p < .001$, $d = 0.79$]. This was also true of the second training block: participants gave significantly more correct responses in the second half (trials 19–36, $M = 86.0\%$; $SD = 19.7\%$) than in the first half (trials 1–18, $M = 80.7\%$; $SD = 19.1\%$) [$t(159) = 5.76$, $p < .001$, $d = 0.46$]. Together, these results suggest discovery and learning of the Greeble feature rule in Context A, and re-learning of the flipped rule in Context B.

6.2. Test performance

Following our pre-registered analysis plan, average niceness ratings were analyzed using a paired *t*-test (Fig. 4b). Greebles from the family

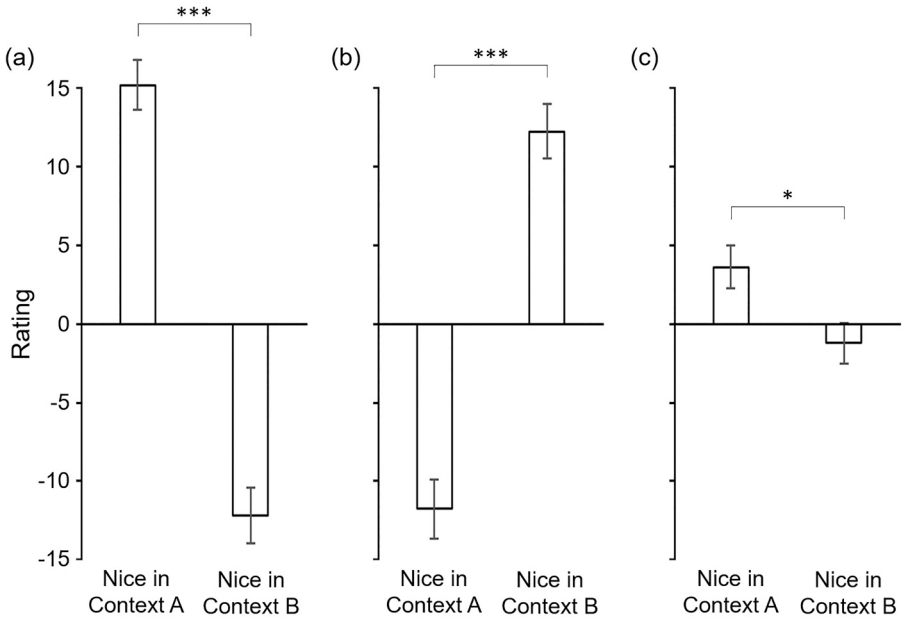

Fig. 4. (a) Results from Experiment 1: AB-A. (b) Results from Experiment 2: AB-B. (c) Results from Experiment 3: AB-C. Error bars represent ±SEM. *** denotes $p < .001$, * denotes $p < .05$.

paired with trustworthy behaviours in the first context ($M = -11.75$; $SD = 23.63$) were rated as significantly less nice than Greebles from the family paired with trustworthy behaviours in the second context ($M = 12.24$, $SD = 21.86$) [$t(159) = 7.13$, $p < .001$, $d = 0.56$].

Together, the results of the AB-A and AB-B experiments provide compelling evidence that participants acquire sets of parallel associations in the first (Context A) and second (Context B) training phases that selectively manifest at Test in Context A and Context B, respectively. While a partial unlearning account might explain the results of the AB-A experiment in isolation, it cannot also explain the results of the AB-B experiment.

7. Experiment 3 (AB-C)

In our third experiment, we investigated which set of associations – those acquired in Context A or Context B – dominate participants' responding when the test phase in conducted in a novel context. Based on previous findings from the associative learning literature (e.g., Nelson et al., 2011), we predicted that the associations acquired in Context A would be more likely to generalise to novel contexts, than those acquired in Context B. In the associative learning literature, this is referred to as AB-C renewal.

8. Methods

One hundred and sixty participants were recruited through http://www.prolific.co ($M_{age} = 35.40$ years, $SD_{age} = 13.25$ years, range: 18–72 years, 93 female, 66 male, 1 non-binary) using the same criteria as in the first two experiments. Once again, participants received a small honorarium. Twelve further participants were tested but replaced having failed one or both attention checks.

Participants completed the same training procedure as in Experiments 1 and 2. In this experiment, however, the test took place in a novel context - the Desert Planet (Fig. 2b). Immediately prior to the test trials, participants viewed an image of the Desert Planet accompanied by the words "Now you are on the Desert Planet. Greebles live on this planet".

9. Results

9.1. Training performance

The proportion of correct responses seen across the first and second phase of the training procedure are plotted in Fig. 3c. The pattern of responding during training was very similar to that observed in the first two experiments. In the first training block, participants gave significantly more correct responses in the second half (trials 19–36, $M = 83.3\%$; $SD = 20.9\%$) than in the first half (trials 1–18, $M = 71.8\%$; $SD = 18.8\%$) [$t(159) = 8.47$, $p < .001$, $d = 0.67$]. This was also true of the second training block: participants gave significantly more correct responses in the second half (trials 19–36, $M = 85.4\%$; $SD = 21.0\%$) than in the first half (trials 1–18, $M = 80.7\%$; $SD = 18.4\%$) [$t(159) = 5.57$, $p < .001$, $d = 0.44$]. Together, these results suggest discovery and learning of the Greeble feature rule in Context A, and re-learning of the flipped rule in Context B.

9.2. Test performance

Following our pre-registered analysis plan, average niceness ratings were analyzed using a paired *t*-test (Fig. 4c). Greebles from the family paired with trustworthy behaviours in the in first context ($M = 3.63$; $SD = 17.05$) were rated as significantly nicer than Greebles from the family paired with trustworthy behaviours in the second context ($M = -1.22$, $SD = 16.54$) [$t(159) = 2.11$, $p = .037$, $d = 0.17$].

The results of Experiment 3 are suggestive of an AB-C renewal effect ($d = 0.17$), but one that is considerably weaker than the AB-A renewal effect seen in Experiment 1 ($d = 0.69$). The mean difference in ratings seen in Experiment 1 ($M = 27.40$; $SD = 39.76$) was significantly greater than the mean difference in ratings seen in Experiment 3 ($M = 4.85$; $SD = 29.11$) [$t(318) = 5.79$, $p < .001$, $d = 0.65$].

10. Experiment 4 (AB-ABC)

Direct comparison of the effects seen in the first 3 experiments is complicated by the fact that different participants completed each experiment. In our final experiment, we therefore sought to replicate the three effects described above using a within-subjects paradigm. Rather,

than being tested in one of the three contexts (A, B or C), each participant was tested in all three contexts.

11. Method

One hundred and sixty participants were recruited through http://www.prolific.co (M_{age} = 37.82 years, SD_{age} = 11.72 years, range: 19–63 years, 84 female, 73 male, 1 prefer-not-to-say, 2 non-binary) using the same criteria as Experiment 1 and received a small honorarium. Twelve further participants were tested but replaced having failed one of both of the attention checks.

Participants completed the same two training phases used in the experiments described above. During the test phase, however, each of the twelve test Greebles was rated three times, once in each context. The order of the test trials was randomized.

In our first three experiments, a notification was presented to participants at the start of the test phase. In the first two experiments, this notification was either "Now you are back on the Forest Planet" or "Now you are back on the Mountain Planet" depending on which counterbalancing condition they were in. In the third experiment, this notification was always "Now you are on the Desert Planet. Greebles live on this planet." In our final experiment, no such notification was presented.

12. Results

12.1. Training performance

The proportion of correct responses seen across the first and second phase of the training procedure are plotted in Fig. 3d. The pattern of responding during training was very similar to that observed in the previous experiments. In the first training block, participants gave significantly more correct responses in the second half (trials 19–36, M = 85.1%; SD = 18.9%) than in the first half (trials 1–18, M = 73.3%; SD = 19.4%) [$t(159) = 8.83$, $p < .001$, $d = 0.70$]. This was also true of the second training block: participants gave significantly more correct responses in the second half (trials 19–36, M = 87.1%; SD = 17.9%) than in the first half (trials 1–18, M = 81.5%; SD = 18.2%) [$t(159) = 6.00$, $p < .001$, $d = 0.48$]. Once again, these results suggest discovery and learning of the Greeble feature rule in Context A, and re-learning of the flipped rule in Context B.

12.2. Test performance

Following our pre-registered analysis plan, average niceness ratings (Fig. 5) were analyzed using ANOVA with Test Context (A, B, C) and Greeble Family (trustworthy in first context, trustworthy in second context) as within-subjects factors. The analysis revealed no main effect of Test Context [$F(1,318) = 0.60$, $p = .551$, $\eta_p^2 = 0.004$] or Greeble Family [$F(1,318) = 1.46$, $p = .228$, $\eta_p^2 = 0.009$], but did show a significant Test Context × Greeble Family interaction [$F(1,318) = 44.68$, $p < .001$, $\eta_p^2 = 0.22$].

When presented in Context A, Greebles from the family paired with trustworthy behaviours in the first context ($M = 8.51$; $SD = 20.88$) were rated as significantly nicer than Greebles from the family paired with trustworthy behaviours in the second context ($M = -10.05$, $SD = 20.00$) [$t(159) = 6.40$, $p < .001$, $d = 0.51$]. When presented in Context B, Greebles from the family paired with trustworthy behaviours in the in first context ($M = -8.76$; $SD = 20.15$) were rated as significantly less nice than Greebles from the family paired with trustworthy behaviours in the second context ($M = 8.66$, $SD = 20.53$) [$t(159) = 6.04$, $p < .001$, $d = 0.48$]. When presented in Context C, Greebles from the family paired with trustworthy behaviours in the first context ($M = 1.63$; $SD = 16.82$) tended to be rated as nicer than Greebles from the family paired with trustworthy behaviours in the second context ($M = -2.82$, $SD = 17.04$), however this trend failed to reach significance [$t(159) = 1.97$, $p = .051$, $d = 0.16$].

Next, we sought to compare the strength of the AB-A, AB-B, and AB-C effects observed. The strength of the AB-A and AB-C effects were estimated by subtracting the ratings for Greebles that were presented as trustworthy in the second context from the ratings of Greebles that were presented as trustworthy in the first context. The strength of the AB-B effect was calculated by subtracting the ratings for Greebles that were presented as trustworthy in the first context from the ratings for Greebles that were presented as trustworthy in the second context. The strength of the AB-A ($M = 18.55$; $SD = 36.70$) and AB-B ($M = 17.41$; $SD = 36.45$) effects did not differ significantly [$t(159) = 0.38$, $p = .707$, $d = 0.03$]. However, both the AB-A [$t(159) = 5.13$, $p < .001$, $d = 0.41$] and AB-B effects [$t(159) = 3.40$, $p < .001$, $d = 0.27$] were significantly greater than the AB-C effect ($M = 4.44$; $SD = 28.54$).

In our first two experiments, we observed strong, highly significant AB-A and AB-B effects that we were able to replicate in Experiment 4. However, the AB-C effect observed in Experiment 3 was much weaker and we were unable to replicate this effect in Experiment 4. During the

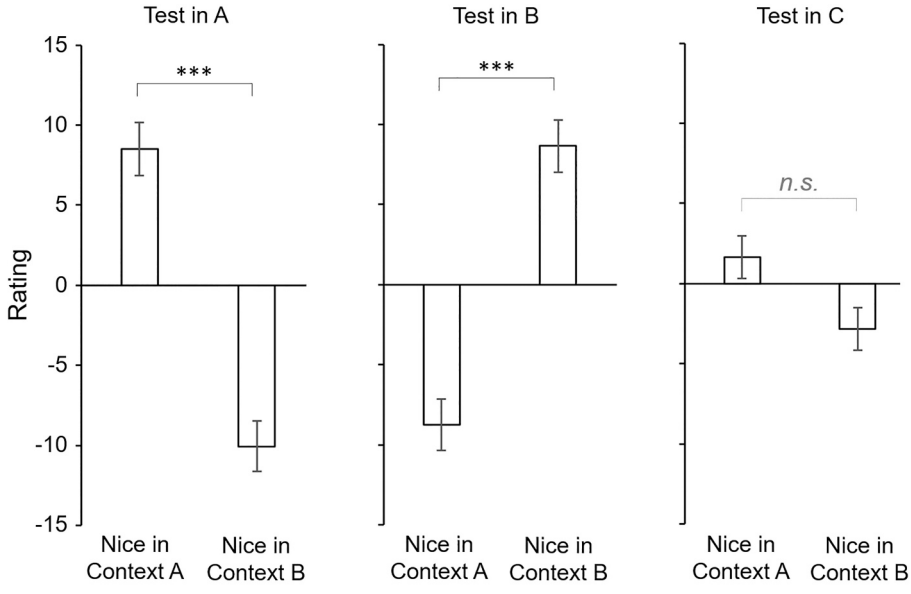

Fig. 5. (a) Results from Experiment 4: AB-ABC. Error bars represent ±SEM. *** denotes $p < .001$.

test phase of Experiment 4, test context was randomly interleaved. By shifting participants from the Mountain Planet, to the Forest Planet, to the Desert Planet, and then back again, within the space of a few trials, we may have undermined the extent to which participants ever felt 'in' a given context. This aspect may have hindered our ability to detect the relatively weak AB-C effect. It is noteworthy that the AB-A and AB-B effects seen in Experiment 4, while highly significant, are also a little weaker than seen in Experiments 1 and 2.

13. General discussion

There is increasing consensus that learning plays a key role in the emergence of first impressions (Chua & Freeman, 2022; Cook et al., 2022; Falvello et al., 2015; FeldmanHall et al., 2018; Lee et al., 2021; Sutherland et al., 2020; Sutherland, Burton, et al., 2020; Verosky & Todorov, 2010). At present, however, little is known about the nature of the learning processes responsible. According to the TIM framework, mappings form between points in face-space – representations of the facial appearance of others – and points in trait space – representations of the likely trait profile of others (Cook et al., 2022; Cook & Over, 2020; Over et al., 2020b; Over & Cook, 2018). Importantly, TIM argues that face-trait mappings are products of domain-general associative processes, like those revealed through the study of conditioning phenomena (Cook et al., 2022; Over & Cook, 2018). If correct, then the associative learning literature can provide useful insights into the mechanisms underlying first impressions and their development. In keeping with this approach, the present study sought to examine whether renewal – a phenomenon revealed through the study of conditioning – might be evident in the learning of first impressions.

In total, we conducted four experiments testing whether appearance-trait learning about Greebles exhibits evidence of renewal. The first two phases of these experiments were identical. In the first phase, participants learned that Greebles from one family were trustworthy while Greebles from another family were untrustworthy. There was a perfect contingency between the appearance of the Greebles (the features associated with their family) and their character traits. This first phase of the experiment took place in Context A (e.g., a mountain environment). In the second phase, participants encountered the same Greebles again, this time in Context B (e.g., a forest environment). In this second phase, the family-trait contingency was reversed: the family that had been trustworthy in Context A was now untrustworthy, and vice versa.

In each case, the key test occurred in the third phase of our experiments. In Experiment 1, the third phase took place back in Context A. When asked to evaluate the likely traits of novel Greebles from the two families, participants tended to base their judgements on the contingencies previously encountered in Context A, thereby demonstrating AB-A renewal. In Experiment 2, the third phase took place in Context B. When asked to evaluate the likely traits of novel Greebles from the two families, participants tended to base their judgements on the contingencies previously encountered in Context B. These effects were replicated in Experiment 4, during which participants rated the novel Greebles in Contexts A and B, as well as a novel Context C.

Together, these findings provide compelling evidence that during the AB training procedure participants acquired sets of parallel associations in the first (Context A) and second (Context B) training phases that selectively manifested at test in Context A and Context B, respectively. During these experiments, participants had equal opportunity to learn an appearance-trait association in phase 1, and then unlearn that association in phase 2. Thus, one might reasonably expect the trait judgements of participants in phase 3 to show no systematic pattern. Crucially, this is not what we observed. Rather than overwrite existing associations, new appearance-trait learning that contradicts old learning appears to create parallel associations that exist alongside the existing associations. Once acquired, the activation of these competing associations is subject to contextual control.

Considered in isolation, the results of the first experiment might be explained by partial overwriting or unlearning. If one assumes that the learning that occurred in Context B was weaker than that seen in Context A (e.g., due to fatigue or boredom), the original learning in Context A may have been only partially overwritten. This might explain why participants still responded according to the Context A contingencies when returned to Context A. However, while a partial unlearning account might explain the results of the AB-A experiment in isolation, it cannot also explain the results of the second experiment. A partial unlearning account that assumes one set of context-insensitive associations cannot explain how the same AB training procedure produces opposite patterns of responding when participants are tested in Context A (Experiment 1) and Context B (Experiment 2).

The results of the first two experiments indicate that the AB training procedure left participants with two sets of rival associations, one that manifested in Context A and one that manifested in Context B. In our third experiment, we sought to determine which set of associations dominated participants' responding when tested in a third novel context (C). Consistent with previous studies from the associative learning literature (e.g., Nelson et al., 2011), we found that participants tended to respond according to the contingencies acquired in the first context – evidence of so-called AB-C renewal. However, the AB-C renewal effect seen in Experiment 3 was much smaller than the AB-A renewal effect seen in Experiment 1. A similar trend was seen in Experiment 4, however the AB-C renewal effect failed to reach significance ($p = .051$). In light of the findings from Experiment 3 and 4, it seems likely that the associations acquired in Context A tend to generalise slightly better to novel contexts than those acquired in Context B. However, the fact that the AB-C effect is so weak suggests that learning from Context B also affected responding in the novel context.

13.1. Implications for first impressions from faces

These findings have several implications for our understanding of first impressions from faces and the learning processes responsible for face-trait mappings. First, our results suggest that context matters. The first impressions we form outside the lab are made in particular contexts (e.g., in a London café, while traveling on the New York subway, in a Berlin court room, etc). The present findings suggest that a given observer may sometimes form different first impressions of the same unfamiliar face in different contexts. For example, in Western cultures, cultural messages about the appearance of 'geeks' and 'jocks' may cause some faces to be judged competent in a library context but incompetent on a basketball court, and vice versa. At present, the prevailing approach in first impressions research is to present facial images with little or no contextual information. Participants are shown a facial image (typically, with the background and body occluded or removed) and asked to judge the likely traits of the person depicted without any knowledge of how or where the stranger was encountered. It is important that further work is conducted to better understand the impact of context on these judgements.

Second, our results imply that face-trait stereotypes may be resistant to change. Previously, the findings from renewal paradigms have been used in a range of applied contexts; for example, to understand the challenges faced by people recovering from drug addiction and those who seek to support their recovery (e.g., Bouton, 1994; Bouton, 2002). Historically, interventions and programs designed to stop people drinking and taking drugs have achieved disappointing results. It has been argued that one of the reasons for this ineffectiveness might be that individuals with drug addiction experience AB-A renewal effects when they leave clinical settings – Context B – and return to the environments where they originally learned that drinking and/or substance abuse is rewarding – Context A (e.g., Bouton, 2002). In other words, the positive messages and behaviours acquired in clinical settings to not over-write the self-destructive behaviour learned previously.

In a similar vein, the present results suggest that once acquired, face-trait mappings may be surprisingly resistant to counter-stereotypical

experience. There is increasing interest in the possibility that periods of lab-based training can "unteach" deeply engrained face-trait mappings acquired outside the lab (Jaeger, Todorov, Evans, & van Beest, 2020). However, recent attempts to reduce the effects of first impressions through training interventions have yielded mixed results (Chua & Freeman, 2021; Jaeger et al., 2020). Our findings suggest that this kind of intervention may have limited effectiveness. Even where training interventions appear to be successful in the lab (e.g., Chua & Freeman, 2021), they may exert little or no effect when participants return to the environments in which they originally acquired their face-trait mappings. It is important that future attempts to assess the effectiveness of training interventions consider generalisation to different contexts. A more effective strategy may be to prevent the formation of deleterious face-trait stereotypes in the first place; for example, by modifying the nature of the correlated face-trait experience our children receive. All too often, inaccurate and socially deleterious face-trait mappings are acquired early in development through exposure to social messages including those in film, TV, cartoons, and storybooks (Cook et al., 2022; Over et al., 2020a; Over & Cook, 2018). For example, facial disfigurement is frequently associated with bad character (Croley et al., 2017; Kish & Lansdown, 2000). Once these appearance-trait rules are acquired, they may be hard to unlearn in later life.

Third, these findings have implications for the TIM account of first impressions from faces. The TIM account hypothesises that first impressions of faces are the result of associative mappings that allow excitation to propagate from representations of facial appearance to representations of the trait profile of others (Cook et al., 2022; Cook & Over, 2020; Over et al., 2020b; Over & Cook, 2018). In order to account for the findings described here – that an individual with the same appearance may be attributed different trait profiles in different contexts – it must be possible for face-trait mappings to exhibit modulation by context. For example, in a particular context, certain mappings may be excited or inhibited (Fig. 6). As described above, this kind of modulation is consistent with a domain-general associative process (Bouton, 1994; Bouton & King, 1983; Nelson et al., 2011; Peck & Bouton, 1990).

13.2. The Greeble paradigm

The focus of the current article is the origin of appearance-based impressions; i.e., judgements informed by semi-permanent aspects of facial appearance that change slowly over time. These are the same invariant cues that support judgements of facial identity, and include feature shape and configuration (Calder & Young, 2005; Haxby, Hoffman, & Gobbini, 2000). First impressions based on appearance cues include the inference of trustworthiness and aggression from facial width-to-height ratio (Geniole et al., 2014; Stirrat & Perrett, 2010; Summersby, Harris, Denson, & White, 2022) and the inference of naivety from round face shape (Zebrowitz McArthur & Berry, 1987). Trait attributions based solely on appearance can be thought of as hypotheses about other people formed in the absence of relevant evidence about their past behaviour.

The Greeble model has great potential to reveal insights about the learning mechanisms responsible for appearance-based trait inferences. The study of face-trait learning *de novo* is complicated by the fact that participants, even young children, already have extensive experience with faces before they enter the lab. This renders the study of first-learned associations from faces intractable. The Greeble model offers researchers a means to overcome this problem. The variation between Greeble exemplars is designed to replicate that seen between individual faces. Thus, Greebles can be categorised into two 'genders' (glips and ploks) and different 'families' (Samar, Osmit, Galli, Radok, Tasio) based on their parts and configuration. The presence of this inter-exemplar structure makes them an ideal proxy for faces in studies of appearance-trait learning.

However, there are some of kinds of trait inference that are beyond the scope of the Greeble paradigm and the TIM framework more broadly. When asked to evaluate the traits of people depicted in stimulus images, participants can base their judgements on differences in facial expression, as well as on invariant appearance cues. For example, faces are judged to be more or less trustworthy when participants detect smiles or frowns (Montepare & Dobish, 2003; Oosterhof & Todorov, 2008; Said, Sebe, & Todorov, 2009; Sutherland et al., 2013; Sutherland et al., 2018). Greebles cannot smile or frown, or engage in any other expressive behaviours. As such, there is no opportunity to model trait inferences based on expression cues.

We do not believe this is a limitation of our approach per se; rather, this reflects the fact that trait inferences from facial appearance and facial expression are qualitatively different phenomena (for detailed

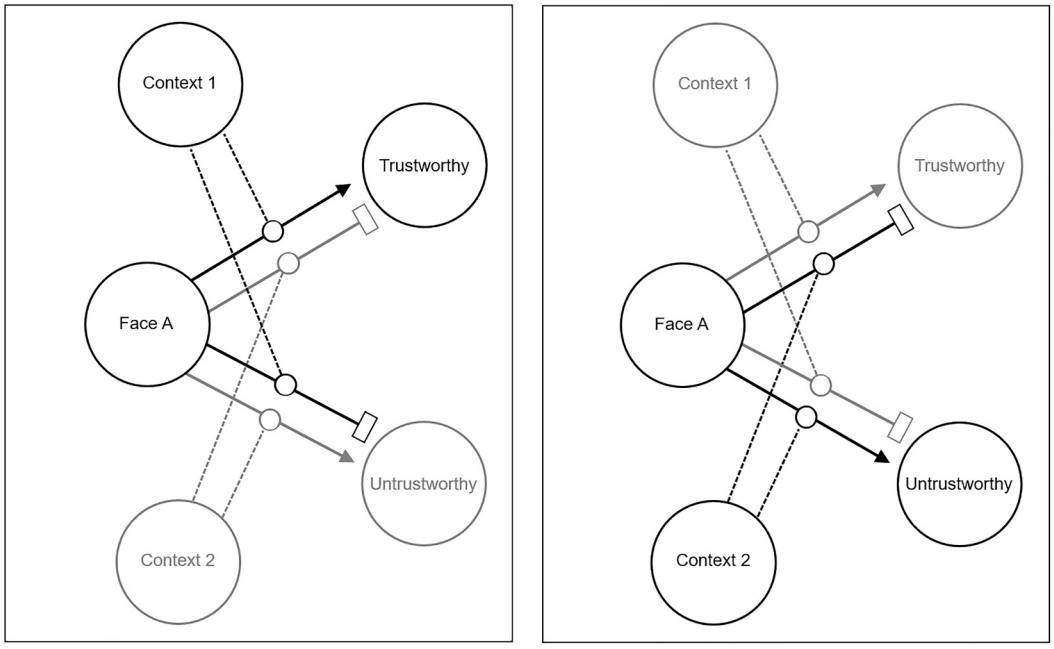

Fig. 6. Illustration of how face-trait mappings may be subject to excitation or inhibition in two different contexts. Arrows indicate excitatory associations. Blocked arrows indicate inhibitory associations. Adapted from Bouton (1994).

discussion, see: Cook et al., 2022). The expression cues present in a facial photograph can be thought of as a 'thin slice' of behaviour (Ambady & Rosenthal, 1992). That someone shown scowling is judged less trustworthy than someone shown smiling is conceptually similar to the inference that someone shouting while wielding a gun is less trustworthy than someone singing while holding a coffee mug. In both cases, the likely traits are inferred from the person's behaviour rather than from their appearance. Whereas appearance-based impressions may be likened to judging a book by its cover, behaviour-based impressions are akin to judging a book by its content (albeit just the first page). Trait inferences based on appearance and behaviour are likely mediated by different cognitive-perceptual mechanisms and may have different developmental origins (Cook et al., 2022).

Previous findings suggest that changes in attitudes towards *familiar* object exemplars can be limited to the context in which counter-attitudinal information was learned (reviewed by Gawronski et al., 2018). The present work complements these findings by applying similar principles to the development of appearance-based first impressions of *novel* exemplars. In our experiments participants were presented with different exemplars at training and test. By using novel exemplars at test, we were able to show that participants generalise their first-learned associations more readily to perceptually similar exemplars. It is this key feature of the Greeble paradigm that allows us to model the development of first impressions of strangers' faces (Lee et al., 2021).

14. Conclusion

The renewal effects observed here suggest that appearance-trait learning may be subject to contextual control. From early in our development, we are exposed to cultural messages about the likely appearance of good-guys and bad-guys, the competent and the incompetent; jocks and geeks. Our results suggest that once acquired, face-trait stereotypes may be resistant to counter-stereotypical experience. Training interventions to attenuate socially deleterious face-trait associations may exert little or no effect when participants return to the environments within which they originally acquired their face-trait mappings.

Credit author statement

HO and RC acquired funding and designed the study. RL, JF and TV collected and analyzed the data. HO and RC wrote the manuscript. RL, JF and TV reviewed and edited the manuscript. All authors approved the final draft.

Funding

This work was funded by a Philip Leverhulme Prize and an ERC Starting grant (ERC-STG-ERCSTG-755719) awarded to HO. RC was also supported by an award from the European Research Council (ERC-STG-715824).

Data availability

The data is available on OSF

References

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin, 111*, 256–274.
Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods, 52*(1), 388–407.
Bouton, M. E. (1994). Context, ambiguity, and classical conditioning. *Current Directions in Psychological Science, 3*(2), 49–53.
Bouton, M. E. (2002). Context, ambiguity, and unlearning: Sources of relapse after behavioral extinction. *Biological Psychiatry, 52*, 976–986.
Bouton, M. E., & King, D. A. (1983). Contextual control of the extinction of conditioned fear: Tests for the associative value of the context. *Journal of Experimental Psychology: Animal Behavior Processes, 9*, 248–265.
Calder, A. J., & Young, A. W. (2005). Understanding the recognition of facial identity and facial expression. *Nature Reviews Neuroscience, 6*(8), 641–651.
Chen, F. F., Jing, Y., Lee, J. M., & Bai, L. (2016). Culture matters: The looks of a leader are not all the same. *Social Psychological and Personality Science, 7*(6), 570–578.
Chua, K. W., & Freeman, J. B. (2021). Facial stereotype bias is mitigated by training. *Social Psychological and Personality Science, 12*(7), 1335–1344.
Chua, K. W., & Freeman, J. B. (2022). Learning to judge a book by its cover: Rapid acquisition of facial stereotypes. *Journal of Experimental Social Psychology, 98*, Article e104225.
Cogsdill, E. J., Todorov, A. T., Spelke, E. S., & Banaji, M. R. (2014). Inferring character from faces: A developmental study. *Psychological Science, 25*(5), 1132–1139.
Cone, J., Flaharty, K., & Ferguson, M. J. (2021). The long-term effects of new evidence on implicit impressions of other people. *Psychological Science, 32*(2), 173–188.
Cook, R., Eggleston, A., & Over, H. (2022). The cultural learning account of first impressions. *Trends in Cognitive Sciences, 26*(8), 656–668.
Cook, R., & Over, H. (2020). A learning model can explain both shared and idiosyncratic first impressions from faces. *Proceedings of the National Academy of Sciences of the USA, 117*(28), 16112–16113.
Croley, J. A., Reese, V., & Wagner, R. F. (2017). Dermatologic features of classic movie villains: The face of evil. *JAMA Dermatology, 153*(6), 559–564.
Dion, K., Berscheid, E., & Walster, E. (1972). What is beautiful is good. *Journal of Personality and Social Psychology, 24*(3), 285–290.
Eagly, A. H., Ashmore, R. D., Makhijani, M. G., & Longo, L. C. (1991). What is beautiful is good, but…: A meta-analytic review of research on the physical attractiveness stereotype. *Psychological Bulletin, 110*(1), 109–128.
Eggleston, A., Flavell, J. C., Tipper, S. P., Cook, R., & Over, H. (2021). Culturally learned first impressions occur rapidly and automatically and emerge early in development. *Developmental Science, 24*(2), Article e13021.
Ewing, L., Sutherland, C. A., & Willis, M. L. (2019). Children show adult-like facial appearance biases when trusting others. *Developmental Psychology, 55*(8), 1694–1701. https://doi.org/10.1037/dev0000747
Falvello, V. B., Vinson, M., Ferrari, C., & Todorov, A. (2015). The robustness of learning about the trustworthiness of other people. *Social Cognition, 33*, 368–386.
FeldmanHall, O., Dunsmoor, J. E., Tompary, A., Hunter, L. E., Todorov, A., & Phelps, E. A. (2018). Stimulus generalization as a mechanism for learning to trust. *Proceedings of the National Academy of Sciences of the United States of America, 115*(7), E1690–E1697.
Fleischmann, A., Lammers, J., Stoker, J. I., & Garretsen, H. (2019). You can leave your glasses on: Glasses can increase electoral success. *Social Psychology, 50*(1), 38–52.
Gauthier, I., Tarr, M., Anderson, A., Skudlarski, P., & Gore, J. (1999). Activation of the middle fusiform 'face area' increases with expertise in recognizing novel objects. *Nature Neuroscience, 2*, 568–573.
Gauthier, I., & Tarr, M. J. (1997). Becoming a "Greeble" expert: Exploring mechanisms for face recognition. *Vision Research, 37*, 1673–1681.
Gauthier, I., Williams, P., Tarr, M. J., & Tanaka, J. (1998). Training 'greeble' experts: A framework for studying expert object recognition processes. *Vision Research, 38*, 2401–2428.
Geniole, S. N., Molnar, D. S., Carré, J. M., & McCormick, C. M. (2014). The facial width-to-height ratio shares stronger links with judgments of aggression than with judgments of trustworthiness. *Journal of Experimental Psychology: Human Perception and Performance, 40*(4), 1526–1541.
Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences, 4*, 223–233.
Hooper, J. J., Sutherland, C. A. M., Ewing, L., Langdon, R., Caruana, N., Connaughton, E., … Rhodes, G. (2018). Should I trust you? Autistic traits predict reduced appearance-based trust decisions. *British Journal of Psychology*. https://doi.org/10.1111/bjop.12357
Ilardi, V. (2007). *Renaissance vision from spectacles to telescopes*. Philadelphia: American Philosophical Society.
Jaeger, B., Todorov, A. T., Evans, A. M., & van Beest, I. (2020). Can we reduce facial biases? Persistent effects of facial trustworthiness on sentencing decisions. *Journal of Experimental Social Psychology, 90*, Article e104004.
Jones, B., DeBruine, L., Flake, J., Aczel, B., Adamkovic, M., Alaei, R., … Vásquez-Amézquita, M. (2021). To which world regions does the valence-dominance model of social perception apply? *Nature Human Behaviour, 5*(1), 159–169.
Keen, S. (1991). *Faces of the enemy: Reflections of the hostile imagination: Harper San Francisco*.
Kish, V., & Lansdown, R. (2000). Meeting the psychosocialimpact of facial disfigurement: Developing a clinical service forchildren and families. *Clinical Child Psychology and Psychiatry, 5*, 497–512.
Lakshmi, A., Wittenbrink, B., Correll, J., & Ma, D. S. (2021). The India face set: International and cultural boundaries impact face impressions and perceptions of category membership. *Frontiers in Psychology, 12*, Article e627678.
Lee, R., Flavell, J. C., Tipper, S. P., Cook, R., & Over, H. (2021). Spontaneous first impressions emerge from brief training. *Scientific Reports, 11*(1), Article e15024.
Montepare, J. M., & Dobish, H. (2003). The contribution of emotion perceptions and their overgeneralizations to trait impressions. *Journal of Nonverbal Behavior, 27*(4), 237–254.
Nelson, J. B., Sanjuan, M. D. C., Vadillo-Ruiz, S., Pérez, J., & León, S. P. (2011). Experimental renewal in human participants. *Journal of Experimental Psychology: Animal Behavior Processes, 37*, 58–70.
Olivola, C. Y., Funk, F., & Todorov, A. (2014). Social attributions from faces bias human choices. *Trends in Cognitive Sciences, 18*, 566–570.

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences of the USA, 105*, 11087–11092.
Over, H., & Cook, R. (2018). Where do spontaneous first impressions of faces come from? *Cognition, 170*, 190–200.
Over, H., Eggleston, A., & Cook, R. (2020a). Ritual and the origins of first impressions. In *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*.
Over, H., Eggleston, A., & Cook, R. (2020b). Ritual and the origins of first impressions. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 375*(1805), Article e20190435.
Peck, C. A., & Bouton, M. E. (1990). Context and performance in aversive-to-appetitive and appetitive-to-aversive transfer. *Learning and Motivation, 21*, 1–31.
Rule, N. O., Krendl, A. C., Ivcevic, Z., & Ambady, N. (2013). Accuracy and consensus in judgments of trustworthiness from faces: Behavioral and neural correlates. *Journal of Personality and Social Psychology, 104*, 409–426.
Said, C. P., Sebe, N., & Todorov, A. (2009). Structural resemblance to emotional expressions predicts evaluation of emotionally neutral faces. *Emotion, 9*(2), 260–264.
Siddique, S., Sutherland, C. A., Palermo, R., Foo, Y. Z., Swe, D. C., & Jeffery, L. (2022). Development of face-based trustworthiness impressions in childhood: A systematic review and metaanalysis. *Cognitive Development, 61*, Article e101131.
Sofer, C., Dotsch, R., Oikawa, M., Oikawa, H., Wigboldus, D. H., & Todorov, A. (2017). For your local eyes only: Culture-specific face typicality influences perceptions of trustworthiness. *Perception, 46*(8), 914–928.
Stirrat, M., & Perrett, D. I. (2010). Valid facial cues to cooperation and trust: Male facial width and trustworthiness. *Psychological Science, 21*(3), 349–354.
Summersby, S., Harris, B., Denson, T. F., & White, D. (2022). Tracking sexual dimorphism of facial width-to-height ratio across the lifespan: Implications for perceived aggressiveness. *Royal Society Open Science, 9*(5), Article e211500.
Sutherland, C. A., Burton, N. S., Wilmer, J. B., Blokland, G. A. M., Germine, L., Palermo, R., … Rhodes, G. (2020). Individual differences in trust evaluations are shaped mostly by environments, not genes. *Proceedings of the National Academy of Sciences of the United States of America, 117*(19), 10218–10224.
Sutherland, C. A., Collova, J. R., Palermo, R., Germine, L., Rhodes, G., Blokland, G. A., … Wilmer, J. B. (2020). Reply to cook and over: Social learning and evolutionary mechanisms are not mutually exclusive. *Proceedings of the National Academy of Sciences of the USA, 117*(28), 16114–16115.
Sutherland, C. A., Liu, X., Zhang, L., Chu, Y., Oldmeadow, J. A., & Young, A. W. (2018). Facial first impressions across culture: Data-driven modeling of Chinese and British perceivers' unconstrained facial impressions. *Personality and Social Psychology Bulletin, 44*, 521–537.
Sutherland, C. A., Oldmeadow, J. A., Santos, I. M., Towler, J., Burt, D. M., & Young, A. W. (2013). Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition, 127*, 105–118.
Thierry, S. M., & Mondloch, C. J. (2021). First impressions of child faces: facial trustworthiness influences adults' interpretations of children's behavior in ambiguous situations. *Journal of Experimental Child Psychology, 208*, 105153. https://doi.org/10.1016/j.jecp.2021.105153
Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science, 308*(5728), 1623–1626.
Todorov, A., Olivola, C., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual Review of Psychology, 66*, 519–545.
Todorov, A., & Uleman, J. S. (2002). Spontaneous trait inferences are bound to actors' faces: Evidence from a false recognition paradigm. *Journal of Personality and Social Psychology, 83*(5), 1051–1065.
Verosky, S. C., & Todorov, A. (2010). Generalization of affective learning about faces to perceptually similar faces. *Psychological Science, 21*(6), 779–785.
Walker, M., Jiang, F., Vetter, T., & Sczesny, S. (2011). Universals and cultural differences in forming personality trait judgments from faces. *Social Psychological and Personality Science, 2*, 609–617.
Wilson, J. P., & Rule, N. O. (2015). Facial trustworthiness predicts extreme criminal-sentencing outcomes. *Psychological Science, 26*(8), 1325–1331.
Zebrowitz, L. A. (2017). First impressions from faces. *Current Directions in Psychological Science, 26*, 237–242.
Zebrowitz, L. A., & Montepare, J. M. (1992). Impressions of babyfaced individuals across the life span. *Developmental Psychology, 28*(6), 1143–1152.
Zebrowitz, L. A., Wang, R., Bronstad, P. M., Eisenberg, D., Undurraga, E., Reyes-García, V., & Godoy, R. (2012). First impressions from faces among US and culturally isolated Tsimane'people in the Bolivian rainforest. *Journal of Cross-Cultural Psychology, 43*(1), 119–134.
Zebrowitz McArthur, L., & Berry, D. S. (1987). Cross-cultural agreement in perceptions of babyfaced adults. *Journal of Cross-Cultural Psychology, 18*(2), 165–192.