Face gender and stereotypicality influence facial trait evaluation: Counter-stereotypical female faces are negatively evaluated

Clare A. M. Sutherland1*, Andrew W. Young1, Carmen A. Mootz1 and Julian A. Oldmeadow2

1Department of Psychology, University of York, Heslington, UK
2Psychological Science and Statistics, Swinburne University of Technology, Melbourne, Victoria, Australia

The facial first impressions literature has focused on trait dimensions, with less research on how social categories (like gender) may influence first impressions of faces. Yet, social psychological studies have shown the importance of categories like gender in the evaluation of behaviour. We investigated whether face gender affects the positive or negative evaluation of faces in terms of first impressions. In Study 1, we manipulated facial gender stereotypicality, and in Study 2, facial trustworthiness or dominance, and examined the valence of resulting spontaneous descriptions of male and female faces. For both male and female participants, counter-stereotypical (masculine or dominant looking), female faces were perceived more negatively than facially stereotypical male or female faces. In Study 3, we examined how facial dominance and trustworthiness affected rated valence across 1,000 male and female ambient face images, and replicated the finding that dominance is more negatively evaluated for female faces. In Study 4, the same effect was found with short stimulus presentations. These findings integrate the facial first impressions literature with evaluative differences based on social categories.

First impressions from faces

A wealth of literature has documented that we rapidly and accurately determine social categories from faces, including age, sex and ethnicity (Bruce & Young, 2012). The physical cues underlying such pervasive categories are now relatively well understood (Bruce & Young, 2012), but perceivers also rapidly and consistently infer more abstract social traits from faces such as trustworthiness and dominance (Oosterhof & Todorov, 2008; Willis & Todorov, 2006). These facial first impressions have been shown to affect decisions as important as voting choices (Todorov, Mandisodza, Goren, & Hall, 2005) and court judgements (Zebrowitz & McDonald, 1991), but less is known about how they are formed.

Recently, however, a substantial advance in facial first impressions research has been to ask participants to rate faces on a range of social traits and then factor-analyse these ratings to uncover underlying dimensions (Oosterhof & Todorov, 2008; Sutherland et al.,...
This approach has revealed three key dimensions of facial first impressions: trustworthiness, dominance and youthful-attractiveness (Oosterhof & Todorov, 2008; Sutherland et al., 2013). The first two dimensions seem to relate to the appraisal of threat, while youthful-attractiveness is important in sexual selection. Interestingly, the trustworthiness and dominance facial dimensions may have parallels to warmth (cf. trustworthiness) and competence (cf. dominance) dimensions which underlie the stereotyping of category groups (e.g., Cuddy, Fiske, & Glick, 2008), as well as communal and agentic dimensions found in judgements of the self and others (e.g., Abele & Wojciszke, 2007), suggesting the potential for integration between studies of facial first impressions and wider issues in social psychology.

While the dimensional approach has been very useful, the facial first impressions literature has not yet fully investigated the effect of social categories (see Quinn & Macrae, 2011, for a similar point). In particular, this facial literature (e.g., Oosterhof & Todorov, 2008; Sutherland et al., 2013; Walker & Vetter, 2009) has not looked systematically at how perceptions of these traits or dimensions might be affected by the stereotyping or evaluation of category groups. For example, there are currently no studies of how gender stereotypes affect perceptions of trustworthiness or dominance from faces. Yet, a long history of social psychological research has shown the importance of social categories and associated stereotypes when evaluating social attributes and behaviour, with gender being an especially salient example (e.g., Allport, 1954; Fiske, 2012; Moss-Racusin, Dovidio, Brescoll, Graham, & Handelsman, 2012).

By ‘stereotypes’, we mean qualities perceived to be associated with particular groups or categories of people (cf. Schneider, 2004). We and others have shown previously that facial representations of groups are stereotypes in the sense that perceivers associate certain facial characteristics with certain groups, and these visual characteristics are closely linked to linguistic stereotypes of those groups (Imhoff, Woelki, Hanke, & Dotsch, 2013; Oldmeadow, Sutherland, & Young, 2013). This process is similar to the mechanism of ‘categorization’ described by Secord (e.g., Secord, 1958). In categorization, facial first impressions are generated by first assigning a face to a category, and then using associated category knowledge (i.e., stereotypes) to evaluate that face.

This social psychological literature on gender stereotyping, along with the similarities between the trustworthiness/warmth and dominance/competence dimensions found in the social and face perception literatures, suggests that these two dimensions of trustworthiness and dominance may also be evaluated differently in male and female faces, since traits along the first communal dimension are particularly important for females, and traits along the second, agentic dimension are particularly important for males (Cuddy et al., 2008; Prentice & Carranza, 2002; Wiggins, 1979; Wojciszke, 1994). In particular, from social psychological studies showing negative evaluations of targets with counter-stereotypical roles or behaviour (e.g., Flannigan, Miles, Quaidfiegl, & Macrae, 2013; Heilman, 2001; Rudman, 1998), we might expect negative evaluations of targets with counter-stereotypical facial traits.

Instead, the facial first impressions literature links the first, trustworthiness dimension to overall valence, and the second, dominance dimension to perceived femininity-masculinity (e.g., Oosterhof & Todorov, 2008), rather than linking traits on the first dimension to evaluations of femininity and traits on the second dimension to evaluations of masculinity, as in the social literature (e.g., Abele & Wojciszke, 2007; Bem, 1974; Prentice & Carranza, 2002). By hypothesizing this separation between valence and facial cues involved in gender stereotypicality (femininity-masculinity), these facial first impressions models have not yet examined how cues to gender
stereotypicality might interact with the gender of the face to influence the valence of facial impressions.

**Overview of current studies**

An interesting question, therefore, is whether first impressions of faces are also subject to gender stereotyping evaluations. This question brings together social psychological perspectives on gender stereotyping and group-based evaluation with the dimensional facial trait perception approach. Since the social psychological literature has shown that target gender stereotypicality has an impact on perceiver evaluations, here we manipulated facial gender stereotypicality along with facial gender and examined the valence of resulting impressions. We predicted that different sub-groups of faces that differ in gender but which are positioned in the same locations in a dimensional trait face space would be evaluated differently depending on their gender stereotypicality. Specifically, based on previous findings of negative evaluation of counter-stereotypical targets, we investigated whether facially counter-stereotypical women and men would be perceived less positively than facially stereotypical men and women.

To investigate this, we used converging methods. For Study 1 and Study 2, we developed an approach based only on asking participants to write down their thoughts about different faces. For Study 3 and Study 4, we used a complementary but more direct approach of asking participants to rate the valence of their first impressions.

**STUDY 1**

For our first approach to the issue of how the gender-stereotypicality of the face might interact with the valence of facial impressions, we used everyday photographs of real male and female faces with typically feminine or masculine appearance. We investigated spontaneous impressions to these faces by asking participants simply to state whatever came to mind when viewing each face. To visualize these spontaneous impressions, we created word clouds depicting the frequency of attributions, and complemented this qualitative approach with quantitative assessments based on coding the participants’ descriptions on their valence (positive-negative). The key features of Study 1 are therefore that the stimuli were real faces that were selected in terms of their gender stereotypicality (feminine or masculine appearance) and that participants’ responses were unconstrained. The advantage of this method is that it does not in any way influence what attributions participants make, and therefore shows what is spontaneously perceived.

**Methods**

Ten participants (mean age, 24.3 years; 5 male) took part. Participants provided informed consent to procedures that were approved by the ethics committee of the University of York psychology department. All participants self-identified as culturally Western.

**Stimuli**

Twenty faces were chosen from a set of 1,000 highly varied ‘ambient image’ face photographs used in previous studies (Santos & Young, 2005, 2008; Sutherland et al., 2013). The concept of ambient images was introduced by Burton, Jenkins, and Schweinberger (2011; see also Jenkins, White, Van Montfort, & Burton, 2011) to
emphasize the potential importance of the variability between images of faces we see in everyday life. In order to represent this variability and thus allow us to examine natural first impressions, the ambient image database consists of 1,000 photographs of 500 male and 500 female faces of all adult ages which are deliberately allowed to vary on many cues including pose, expression, lighting and facial paraphernalia such as hairstyles and glasses (see Santos & Young, 2005, 2008; Sutherland et al., 2013 for further details). The faces were restricted to adults of Caucasian appearance, as cross-cultural or own-race biases were not the focus of this investigation.

From this database of 1,000 ambient images, we chose twenty faces for use in Study 1, selecting these on the basis of previously collected sets of ratings of each of the 1,000 images (Santos & Young, 2005, 2008; Sutherland et al., 2013). The twenty faces comprised five men and five women of gender stereotypical appearance (masculine-looking men and feminine-looking women), and five men and five women of gender counter-stereotypical appearance (feminine-looking men and masculine-looking women). Ratings of perceived femininity-masculinity (at least 6 raters per trait, taken from Santos & Young, 2005, 2008; Sutherland et al., 2013) are presented in Table 1, together with ratings of other important characteristics. Table 1 also gives separate ratings of perceived facial femininity (1 low – 7 high) for the female faces and perceived facial masculinity (1 low – 7 high) for the male faces. Note that the faces high and low in gender stereotypicality were deliberately allowed to vary on other traits, so masculine-looking faces of both genders were also rated significantly higher on stereotypically masculine traits (e.g., dominance) and significantly lower on stereotypically feminine traits (e.g., approachability) than feminine-looking faces of both genders. Importantly, the faces were selected to minimize overall trait differences between the genders, so that the male and female face sets did not differ on key gender-stereotypic traits (trustworthiness, approachability, intelligence, smiling and dominance: see Table 1). This sets up a strong test of whether the gender stereotypicality of the face changes the valence of spontaneous impressions.

Table 1. Mean trait ratings (and standard deviations) for the four face sets, at the face level

<table>
<thead>
<tr>
<th></th>
<th>Females</th>
<th>Males</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Feminine-looking</td>
<td>Masculine-looking</td>
</tr>
<tr>
<td>Femininity-masculinity</td>
<td>1.68&lt;c (0.13)</td>
<td>4.02&lt;b,c (0.26)</td>
</tr>
<tr>
<td>Femininity (female faces)</td>
<td>5.65 (0.31)</td>
<td>3.44 (1.22)</td>
</tr>
<tr>
<td>Masculinity (male faces)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Approachability</td>
<td>4.93&lt;c (1.07)</td>
<td>2.75&lt;c (0.87)</td>
</tr>
<tr>
<td>Trustworthiness</td>
<td>4.80&lt;c (0.72)</td>
<td>3.39&lt;c (0.68)</td>
</tr>
<tr>
<td>Smiling</td>
<td>4.63&lt;c (1.48)</td>
<td>2.27&lt;c (1.24)</td>
</tr>
<tr>
<td>Aggressiveness</td>
<td>2.20&lt;c (0.54)</td>
<td>3.82&lt;c (1.11)</td>
</tr>
<tr>
<td>Dominance</td>
<td>3.04&lt;c (0.13)</td>
<td>4.48&lt;c (0.30)</td>
</tr>
<tr>
<td>Intelligence</td>
<td>3.92 (0.64)</td>
<td>3.63 (0.94)</td>
</tr>
<tr>
<td>Attractiveness</td>
<td>5.47&lt;c (0.40)</td>
<td>3.33&lt;c (1.04)</td>
</tr>
</tbody>
</table>

Note. aSignificant difference across feminine males and females, p < .05. bSignificant difference between masculine males and females, p < .05. cSignificant difference between female means, p < .05. dSignificant difference between male means, p < .05. eMarginally significant difference, p < .07.
**Procedure**

Participants were tested in a quiet room on a PC running E-Prime software (version 2; Psychology Software Tools, Pittsburgh, PA, USA). Faces were shown one at a time with a blank text box. Participants were told that the study was examining first impressions and were invited to type in anything that came to mind on viewing the face, no matter how silly, judgemental or socially inappropriate. They were encouraged to be honest and reassured that their responses were anonymous and that there was no right or wrong answer. Participants were instructed to write as much as possible until they felt that they were no longer being spontaneous. Participants saw the 20 experimental faces in a random order, with two faces – which were not analysed – as a practice.

**Thematic analysis**

Our data were participants’ descriptions, split into units consisting of single words or phrases (e.g., ‘not friendly’). Word clouds (wordle.com) were then used to examine themes to see if negative evaluations occurred for counter-stereotypical faces (see Figure 1). These word clouds present higher frequency descriptions in larger font. This allows exploration of the main concepts emerging from the descriptions. Only key words and short phrases (<5 words) were included, to allow examination of the main concepts without common but uninformative words or long phrases hampering this. Words or phrases were not included where to shorten them or take them out of context would be misleading (e.g., ‘she works in a male-dominated area’ could not be fairly summarized).

**Content analysis**

A quantitative content analysis was carried out on all description units to complement and substantiate the qualitative observations. Since the valence of spontaneous impressions

![Figure 1](image-url)  
*Figure 1.* Word clouds depicting first impressions made to gender stereotypical (feminine female and masculine male) and gender counter-stereotypical (masculine female and feminine male) faces. Larger font size represents more frequent descriptions.
was the principal focus of interest, each individual word or phrase was blind coded by two judges for their valence. This was achieved by coding descriptions as either positive or negative if they referred to a skill or problem (e.g., ‘sporty’, or ‘has a difficulty changing her opinions’), used a positive or negative qualifier (e.g., ‘good sense of humour’ or ‘hate his moustache’) or were conventionally positive or negative (e.g., ‘happy’ or ‘angry’).

Results
The participants produced a total of 2,981 words describing the faces, which were then separated into 1027 units (words and phrases). On average, each person described each face with 5.14 units, suggesting that facial first impressions are relatively rich.

Thematic word clouds
In total, 109 units (11%) could not be included in the word clouds, either because they lacked a key word or phrase, or because shortening them would change their meaning. The remaining majority (918 units) were inspected for themes for each face group (see Figure 1).

The themes emerging for feminine women were largely positive inferences of physical attractiveness, happiness and friendliness. Masculine men were also described as friendly and approachable, with positive morality impressions (e.g., caring and kind). However, masculine women received mainly negative attributions: strict, cold, unfriendly and untrustworthy. This negativity was not as evident for feminine males, with themes describing them as funny, friendly and intelligent geeks.

Quantitative content analysis
The descriptions were blind coded by two judges for their valence to quantitatively examine the valence of attributions to the different sets of faces. Agreement was high (Kappa = .94, *p* < .001) and all disagreements were resolved before analysis. We used an index of overall valence by dividing the number of positive words (summed across the 5 faces in the relevant category) by the number of negative words at the participant level. To avoid zero values, the constant 0.5 was added to all cell counts, thus allowing division without error (Gart & Zweifel, 1967).

The data were highly non-normal because they were ratios (no negative values, with a strong rightward skew). Log-transformation of the data still left problems of kurtosis in the residuals (outliers, and high data clustering). We therefore followed Liermann, Steel, Rosing, and Guttorp (2004) in their recommendation to use non-parametric tests. Friedman’s ANOVA was used to ascertain that there was a reliable difference across the four face groups, followed with planned comparisons using Wilcoxon signed-rank *t*-tests. Figure 2 displays the overall mean ranks because this is what the Friedman’s ANOVA is based on, and because the strong skew meant that the median values were also potentially misleading.

The positivity of descriptions was significantly different across the four face categories: \(\chi^2(3) = 19.18, p < .001\) (see Figure 2). There was evidence for negative evaluations to counter-stereotypical females: the relatively masculine-looking females received significantly less positive descriptions than the masculine-looking males (\(Z = -2.40, p < .05\)) and feminine-looking females (\(Z = -2.80, p < .005\)). There was no evidence for negative
evaluations to counter-stereotypical males: the relatively masculine and feminine-looking males did not significantly differ ($Z = -1.13, p = .30$) nor did the more feminine-looking males and females ($Z = 1.78, p = .084$).

**Discussion**

Study 1 found evidence that faces were evaluated differently depending on their gender and gender stereotypicality. Each face category elicited distinctive themes, with all groups apart from the counter-stereotypical masculine-looking females receiving largely positive descriptions (see Figure 1). As predicted, relatively masculine-looking females received more negative evaluations than the stereotypical masculine males and feminine females. However, there was no evidence for negative evaluation towards the counter-stereotypical feminine-looking males. In terms of models of facial first impressions, this points to the idea that the femininity-masculinity cues theorized to underlie the dominance dimension, can interact with the valence of resulting first impressions, which is theorized to be mainly linked to trustworthiness (cf. Oosterhof & Todorov, 2008).

We should point out that to some extent, these evaluations followed the pre-ratings given to the faces initially (see Table 1). Crucially, however, the participants in Study 1 were free to describe anything they liked about the face. That these free descriptions to some extent replicated the previous ratings suggests these impressions really are salient in a natural sense; and Study 1 is therefore a naturalistic validation of the central point that male and female faces are evaluated differently. Moreover, the male and female faces at a given level of masculinity or femininity did not differ on multiple trait ratings; yet they were evaluated differently by participants.

**Figure 2.** Mean ranks from Friedman ANOVAs used to test the overall valence differences for gender stereotypical and counter-stereotypical faces in Study 1. **p < .01, *p < .05.**
Clearly, the feminine and masculine faces could (and did) differ on many perceived traits, raising the question of the potential roles of different traits in their evaluation. In particular, since prominent models of facial first impressions focus on trustworthiness and dominance dimensions (Oosterhof & Todorov, 2008), in Studies 2–4, we opted to explore the effect of manipulating facial dominance and trustworthiness on the evaluation of male and female faces.

STUDY 2
In Study 2, we examined how male and female faces varying along the two main trait dimensions identified in the facial first impressions literature (trustworthiness and dominance: Oosterhof & Todorov, 2008) were spontaneously evaluated. In order to systematically manipulate these factors and to ensure that descriptions were based only on facial cues underlying the relevant dimension, we used computer-generated images to manipulate perceived facial dominance and trustworthiness separately. Average face stimuli were created by averaging sets of male or female faces from the ambient image database rated high or low on either trustworthiness or dominance. The effect of averaging is to create face-like prototype images (referred to as ‘faces’ for simplicity) that only include those facial cues that are consistent across the individual faces used to create each average (Mattavelli, Andrews, Asghar, Towler, & Young, 2012; Oldmeadow et al., 2013; Sutherland et al., 2013).

In Study 1, we did not have enough participants to ascertain whether facial evaluations might be affected by gender-specific perceiver biases. A secondary aim of Study 2 was therefore to examine whether participant sex affected the results.

Methods
Procedure
Forty participants (mean age 20.95 years; 20 male) took part. Participants provided informed consent to procedures that were approved by the ethics committee of the University of York psychology department. The experiment was hosted online by Qualtrics (www.qualtrics.com) although participants were recruited through the University. All participants self-identified as culturally Western. Participants saw all eight average face stimuli in a random order and were asked to freely describe them (instructions as before).

Stimuli
The stimuli were eight carefully controlled computer-generated face-like stimuli (referred to as ‘faces’ for convenience), created to be high and low on perceived trustworthiness or dominance for each gender (see Figure 3). These were taken from pre-existing trustworthiness and dominance stimuli sets consisting of 100 average face images that varied on either trustworthiness or dominance against perceived gender. See Mattavelli et al. (2012) for a full description of the trustworthiness matrix from which the trustworthiness stimuli used in the current study were selected. A similar procedure to that used in Mattavelli et al. (2012) was used to create the dominance matrix used to select the dominance stimuli used in the current study.

Briefly, the 15 most and 15 least dominant male and female faces were selected from the ambient image database, subject to constraints that the photographs included no
spectacles, were as close to frontal view as possible, showed no beards or moustaches, and that there were no more than two faces with hats in each set (in order to generate clear stimuli). The faces in each set of 15 photographs were then averaged using PsychoMorph software (Tiddeman, Burt, & Perrett, 2001) to create four prototype faces. Two quasi-linear continua of 10 male or female face-like images of varying dominance were created by caricaturing or anti-caricaturing each prototype at two levels to increase or decrease its distance from the opposite prototype. These continua of 10 images were then presented in random order and rated for dominance (1–7 scale, low–high) by 14 pilot raters (7 female, mean age 20 years) who did not otherwise participate in the current study. It was necessary to match continua so that the male and female prototype images were equivalently high or low in dominance. We therefore selected male and female images that were rated equally high or low in perceived dominance. The rest of the 10 × 10 stimuli matrix was generated by morphing the faces between the prototypes along the dominance and the gender dimensions and adding a caricatured image in each of the four directions, in steps of 14% [see Mattavelli et al. (2012) for a description of the trustworthiness stimuli].

Figure 3. Computer-manipulated high and low Dominance (A) and Trustworthiness (B) male and female face stimuli used in Study 2, with word clouds presenting participants’ descriptions. Gender stereotypical faces include non-dominant and trustworthy female and dominant and untrustworthy male average faces. Gender counter-stereotypical faces include dominant and untrustworthy female and non-dominant and trustworthy male average faces. Larger word cloud font size represents more frequent descriptions.
From these stimuli sets, we had unambiguously male and female images rated on their trustworthiness and dominance by new pilot raters who otherwise did not take part in the current study in order to choose the stimuli used in the current study (\(n = 10, 5\) female; mean age: 25 years). Crucially, we chose eight faces (see Figure 3; male and female, high and low on perceived dominance or trustworthiness) so that high and low dominance and trustworthiness prototype faces did not differ in rated traits across gender, and so that the dominance faces varied more on dominance than the trustworthiness set and vice versa (see Table 2).

### Results

Overall, participants used 6,843 words to describe the faces. Descriptions could be divided into 1,778 units (words and phrases; 5.56 units on average per participant and face).

All data procedures were as for Study 1. As expected, the data (and residuals) were highly non-normal due to a strong positive skew, and log-transformation was not completely satisfactory due to high clustering and outliers, so we followed Liermann et al. (2004) and used non-parametric tests. Figure 4 again plots the mean ranks from the Friedman’s ANOVAs. A second individual coded 75% of the data and agreement was high (Kappa = .74, \(p < .001\)). All disagreements were resolved before analysis. Descriptions given to trustworthiness and dominance faces were analysed separately.

### Thematic word clouds 2: Dominance faces

In total, 73 units (8%) could not be included. The majority (809 units) were entered into word clouds and examined for themes. These echoed Study 1 (see Figure 3). Specifically, the counter-stereotypical (dominant) female face primarily received negative themes of strictness or sternness. In contrast, the dominant male face was described as a patriarchal, largely benevolent leader (e.g., as a wise businessman). The non-dominant female received stereotypically feminine descriptions, focusing on her as kind, nice and motherly. Finally, the non-dominant male face was seen as friendly and outgoing.

### Table 2. Mean trait ratings (and standard deviations) for the face-like images used as stimuli in Study 2 (shown in figure 3), at the participant level

<table>
<thead>
<tr>
<th>Face set</th>
<th>Face characteristics</th>
<th>Trust.</th>
<th>Dom.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dominance</td>
<td>Low Female</td>
<td>5.1 (1.45)</td>
<td>2.4 (0.70)&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>5.2 (0.92)</td>
<td>2.0 (0.82)&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>High Female</td>
<td>4.2 (1.69)</td>
<td>5.4 (1.10)&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>4.0 (1.41)</td>
<td>5.8 (0.63)&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Trustworthiness</td>
<td>Low Female</td>
<td>3.5 (0.97)&lt;sup&gt;c&lt;/sup&gt;</td>
<td>4.8 (1.40)&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>3.5 (0.85)&lt;sup&gt;d&lt;/sup&gt;</td>
<td>4.4 (0.97)</td>
</tr>
<tr>
<td></td>
<td>High Female</td>
<td>5.2 (1.32)&lt;sup&gt;c&lt;/sup&gt;</td>
<td>2.9 (1.20)&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>5.6 (1.43)&lt;sup&gt;d&lt;/sup&gt;</td>
<td>4.1 (2.08)</td>
</tr>
</tbody>
</table>

Note. \(a\) significant difference between low and high dominance females \(p < .05\).  
\(b\) significant difference between low and high dominance males, \(p < .05\).  
\(c\) significant difference between low and high trustworthy females \(p < .05\).  
\(d\) significant difference between low and high trustworthy males \(p < .05\). All other differences \(p > .07\).
We used Mann–Whitney U tests to ascertain whether there was a difference in the valence of the descriptions made by female and male participants. No significant differences were found (all $U > 145$, $p > .1$) so the data were pooled across participants (see Figure 4).

The overall Friedman’s ANOVA was significant: $\chi^2 (3) = 20.01$, $p < .001$. We found that the dominant (counter-stereotypical) female face was described less positively than the dominant male ($Z = -2.76$, $p < .01$) and non-dominant female face ($Z = -3.18$, $p < .005$). As in Study 1, there was no evidence for negative evaluations of the counter-stereotypical male, with no significant differences in valence between the non-dominant male and female faces ($Z = 0.60$, $p = .56$) or the dominant and non-dominant male faces ($Z = 1.45$, $p = .15$). In sum, the findings for the dominant faces were highly similar to Study 1.

Thematic word clouds 2: Trustworthiness faces
In total, 75 units (8%) could not be included; the majority (821 units) were entered into word clouds. Themes mainly reflected a large difference between trustworthy and untrustworthy faces, supporting Oosterhof and Todorov’s (2008) suggestion that valence is highly linked to perceived trustworthiness (see Figure 3). The trustworthy male received highly positive descriptions centring on his friendliness, kindness and age. The trustworthy female face also received similar positive descriptions, although some participants thought that her friendliness was socially polite rather than genuine. The untrustworthy male was described somewhat ambivalently; although some participants described him as friendly and intelligent, others described him as arrogant, stern and scary. Finally, the untrustworthy female was described mainly negatively, as cold and intimidating.

Content analysis 2: Trustworthiness faces
In general, male and female participants gave similar descriptions: the majority of Mann–Whitney U-test comparing participant gender was not significant ($U > 138$, $p > .09$). However, female participants described the trustworthy and untrustworthy male faces in significantly more positive terms than the male participants. The female participant mean rank for the trustworthy male face was 24.45, with the male participant mean rank as
16.55 ($U = 121, p < .05$). The female participant mean rank for the untrustworthy male face was 24.53 compared to the male participant mean rank of 16.48 ($U = 119.50, p < .05$). Consequently, we analysed the trustworthiness descriptions separately by participant gender (see Figure 4).

The overall Friedman’s ANOVAs were significant for both male, $\chi^2 (3) = 31.98, p < .001$, and female participants, $\chi^2 (3) = 31.22, p < .001$. For male participants, the untrustworthy faces were described in significantly less positive terms than the trustworthy faces (male faces: $Z = -3.92, p < .001$; female faces: $Z = -3.06, p < .005$). Similarly, for female participants, the untrustworthy faces were described in significantly less positive terms than the trustworthy faces (male faces: $Z = -3.29, p < .001$; female faces: $Z = -2.55, p < .01$). There was no evidence of differentially negative evaluations directed at the counter-stereotypical female face, as the untrustworthy female and male faces did not significantly differ for either female ($Z = 1.78, p = .078$) or male participants ($Z = 0.63, p = .54$). Interestingly, the female participants also gave significantly more positive descriptions to the high trustworthy male face than the high trustworthy female face ($Z = 3.36, p < .001$). This comparison was not significant for the male participants ($Z = 0.51, p = .63$).

**Discussion**

In Study 2, it was again clear that different subgroups of faces defined by their gender and perceived social traits were evaluated in different ways. The faces that differed in dominance, but not the faces that differed in trustworthiness, closely followed the pattern of results in Study 1. That is, the counter-stereotypical (dominant-looking) female face was described in less positive terms than the stereotypical (non-dominant) female face and the stereotypical (dominant) male face. Again, there was no evidence for differentially negative evaluations directed at the counter-stereotypical male face.

For the faces manipulated on trustworthiness, the counter-stereotypical (trustworthy) male face was actually described in more positive terms than the stereotypical (trustworthy) female face, at least by female participants. Interestingly, this is contrary to the ‘women are wonderful effect’ described in social psychology, where women are generally perceived more positively than men because they are perceived as higher in stereotypically feminine, pro-social traits (e.g., Eagly & Mladinic, 1994). Nevertheless, both the female and male faces high on perceived trustworthiness received more positive evaluations than those low on perceived trustworthiness, supporting Oosterhof and Todorov’s (2008) suggestion that the valence of first impressions is strongly linked to perceived trustworthiness.

Using computer-manipulated average faces in Study 2 offers a well-controlled test because the image manipulation techniques remove any variance between the face photographs that does not contribute to the trait of interest (e.g., extra-facial paraphernalia), while allowing consistent cues to remain. This minimizes the possibility that extra-facial or facially idiosyncratic characteristics were the cause of these findings. Of course, the face averages differ on a variety of cues. For example, inspection of Figure 3 shows that age and expression contribute to some extent to perceived dominance and trustworthiness. This is expected because Zebrowitz and colleagues have shown that facial maturity is linked to dominance (e.g., Berry & McArthur, 1986) and Todorov and colleagues have shown that facial expression is closely linked to trustworthiness (e.g., Todorov, Baron, & Oosterhof, 2008). Here, we took the approach of allowing all of the
facial cues to trustworthiness or dominance to naturally contribute to these trait judgements, rather than attempting to ‘control’ these cues. Indeed, controlling these cues would change the judgement being made, and we wanted to examine trustworthiness and dominance judgements as they are naturally made. We note that discussing these results in terms of trait (e.g., trustworthiness) judgements or in terms of cues to trait judgements (e.g., happiness expressions) does not change the interpretation of the results (see Keefe, Dzhelyova, Perrett, & Barraclough, 2013). Instead, these facial cues are intrinsic to the trait judgements being made. The important points are that the averaged faces differed systematically in the perceived traits of trustworthiness and dominance and were matched as closely as possible across face gender.

**STUDY 3**

In the previous two studies, we deliberately used a small sample of faces in order not to overburden participants with the free descriptions task and/or because they were prototypical average images. In addition, the method used in Study 1 and Study 2 involved independent raters coding the valence of participants’ spontaneous descriptions rather than asking participants to note the valence of their own impressions. The advantage of using spontaneous descriptions is that participants are not offered any suggestions as to what might be important evaluative dimensions. However, it would be useful to ascertain that our results held with a larger sample of faces and a direct judgment of the valence of the participants’ first impressions.

Participants in Study 3, therefore rated the full ambient image database set (1,000 faces) on the valence of their first impression (i.e., how negative or positive their first impression was). We then used regression to predict these valence ratings from ratings of dominance or trustworthiness previously collected from an independent sample of raters (see Sutherland et al., 2013). From the pattern of results across Studies 1 and 2, we expected to find that the valence of the impression depended on an interaction between the dominance of the face and the gender of the face, so that dominance would be evaluated more negatively for female faces compared to male faces. However, we expected to find that the trustworthiness of the face was strongly related to the valence of the first impression for both male and female faces, based on Oosterhof and Todorov’s (2008) claim that the first dimension is heavily valenced and also on our finding in Study 2 that both male and female faces high in trustworthiness were perceived more positively than those low in trustworthiness.

**Method**

Twelve participants (mean age 18.67 years; 6 female) took part. Participants provided informed consent to procedures that were approved by the ethics committee of the University of York psychology department. All participants self-identified as culturally Western.

**Stimuli**

The participants viewed all 1,000 faces (500 male) from the ambient image database used to construct the stimuli in Studies 1 and 2 (see Methods 1 and 2). The female faces in this database \( k = 500 \) are rated as significantly more trustworthy than the male faces, \( t \)
(981.70) = 13.69, p < .001, with a mean trustworthiness rating of 4.37 (0.66 SD) compared to the male faces at 3.75 (0.75 SD). The female faces are also rated as significantly less dominant than the male faces, t(998) = −35.72, p < .001, with a mean dominance rating of 3.31 (0.61 SD) compared to the male faces at 4.73 (0.65 SD). The facial ratings of trustworthiness and dominance therefore parallel extra-facial gender stereotypes (e.g., Prentice & Carranza, 2002).

Participants were tested in a quiet room on a PC running E-Prime software (version 2; Psychology Software Tools). Faces appeared in random order with a valence rating scale underneath. Participants were instructed to rate the faces on a scale of 1–7 for how positive their first impression of the person was, with 1 being labelled as a very negative impression, and 7 as a very positive impression. They pressed the number key that corresponded with their rating and the next face photograph then appeared after a blank interval of approximately 750 ms. Face photographs were 150 pixels in height (approximately 5 cm on screen) with varying width to preserve aspect ratio. Given that 1,000 faces had to be rated, participants were allowed to take a break whenever they wished. They were given as much time as they wanted to look at each face, but were encouraged to go with their ‘gut instinct’ (Todorov et al., 2005). They first saw 10 faces, randomly drawn from the database, as a practice.

Results

The reliability of the valence ratings was good, with Cronbach’s alpha = .85 (p < .001). These valence ratings were correlated to previously collected ratings of dominance and trustworthiness, for male and female faces separately (Santos & Young, 2005, 2008; Sutherland et al., 2013; see Figure 5). At least six raters had previously rated each trait and all trait reliabilities (alphas) were above .7 (see Santos & Young, 2005, 2008; Sutherland et al., 2013 for further details).

We employed hierarchical regression analyses to assess the interactions between the trait ratings and gender in predicting the valence of first impressions (i.e., moderated regression with a categorical predictor using dummy coding: Aiken & West, 1991; Warner, 2013). Here, the interaction is calculated as the product of the moderator and independent variable, and directly entered as a term in the regression (Aiken & West, 1991; Warner, 2013). For dominance, we first used the face gender (0,1) and dominance rating (centred) to predict the valence ratings, in order to test these main effects. Figure 5 shows the scatterplot of dominance ratings against valence ratings. In a second block we included the interaction of dominance and face gender (0,1) to predict the valence ratings in order to directly test the interaction. The dominance and gender main effects were significant (see Table 3), demonstrating that female faces and faces low in dominance were linked to more positive evaluations. Importantly, the interaction between dominance and gender was also significant and explained a significant proportion of the variance over and above the main effects (see Table 3). To examine this significant interaction, we ran separate regressions predicting the valence of first impressions from dominance for male and female faces separately. For female faces, dominance (centred) significantly negatively predicted the valence of impressions: beta = −.36, t(498) = −8.50, p < .001, and also explained a significant proportion of the variance of the valence of impressions: $R^2 = .13$, $F(1, 498) = 72.21$, $p < .001$. However, for male faces, dominance (centred) significantly positively predicted the valence of impressions: beta = .12, t(498) = 2.58, p = .01, and explained a small but statistically significant
Figure 5. Correlations between dominance and trustworthiness ratings for 500 male and 500 female ambient image faces with the valence of impressions made under (A) unconstrained viewing conditions (Study 3) and (B) 500 ms viewing time (Study 4). **p < .001, *p < .05.

Table 3. $R^2$ and standardized beta weights for individual predictors in predicting valence. Traits were centred. Results are shown for Study 3 (unlimited presentation) and Study 4 (500 ms presentation)

<table>
<thead>
<tr>
<th></th>
<th>Valence</th>
<th></th>
<th>Valence 500 ms</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dominance</td>
<td>Trust.</td>
<td>Dominance</td>
<td>Trust.</td>
</tr>
<tr>
<td>Step 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trait (dominance/trustworthiness)</td>
<td>-.141*</td>
<td>.864**</td>
<td>-.257**</td>
<td>.852**</td>
</tr>
<tr>
<td>Gender</td>
<td>-.249**</td>
<td>-.011</td>
<td>-.141*</td>
<td>.005</td>
</tr>
<tr>
<td>Step 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trait × Gender</td>
<td>.397**</td>
<td>.029</td>
<td>.271**</td>
<td>-.037</td>
</tr>
<tr>
<td>$R^2$ change</td>
<td>.046**</td>
<td>.000</td>
<td>.022**</td>
<td>.001</td>
</tr>
<tr>
<td>Total $R^2$</td>
<td>.181**</td>
<td>.754**</td>
<td>.162**</td>
<td>.723**</td>
</tr>
</tbody>
</table>

Note. **p < .001, *p < .05.
amount of the variance in the valence of impressions: $R^2 = .01, F(1, 498) = 6.65, p = .01$. Our prediction that dominance would be more negatively evaluated for female faces compared to male faces was therefore supported (see Figure 5).

We then used the face gender (0,1) and trustworthiness ratings (centred) to predict the valence ratings, in order to test these main effects. Figure 5 shows the scatterplot of trustworthiness ratings against valence ratings. In a second block we included the interaction of trustworthiness and face gender (0,1) to predict the valence ratings in order to directly test the interaction. The main effect of trustworthiness was positive and highly significant, but the main effect of face gender and the interaction between trustworthiness and gender were not significant and the interaction did not explain any more variance over and above the trustworthiness rating (see Table 3). This shows that the valence of first impressions is highly positively linked to the perceived trustworthiness of both male and female faces (see Figure 5), supporting Oosterhof and Todorov (2008).

**Discussion**

As predicted, there was an interaction between facial dominance and facial gender which significantly predicted the valence of the first impression, so that dominance was evaluated more negatively for female faces compared to male faces. Dominance was also negatively correlated to the valence of the first impression for female faces, whereas for male faces the correlation between dominance and valence was slightly positive. This supports the findings of Study 2 using a large database of 1,000 highly variable images and a direct measure of perceived valence. We also found that trustworthiness correlated strongly with valence for both male and female ambient image faces, as would be predicted by Oosterhof and Todorov (2008).

**STUDY 4**

In Studies 1–3, we did not control the length of time people could view the faces and respond. This was deliberately done to ensure that we recorded all initial impressions and to keep the task as natural as possible. Moreover, this follows the procedure of many studies in the facial first impressions literature (e.g., Oosterhof & Todorov, 2008; Todorov et al., 2005). However, one could argue that if the differential evaluations we noted in Studies 1–3 truly reflect first impressions, then the overall pattern of findings should remain even with short stimulus presentation, since studies have shown that facial first impressions made to faces shown for less than a second largely correspond to unconstrained facial first impressions (Todorov, Pakrashi, & Oosterhof, 2009; Willis & Todorov, 2006). Therefore, in Study 4, we repeated Study 3 but restricting the presentation time for each stimulus to 500 ms and following this with a disruptive pattern mask. We chose 500 ms since this lies in the middle of the range of presentation times used by previous investigations of facial first impressions (cf. Willis & Todorov, 2006). We aimed to replicate the finding that gender and dominance interacted to predict the valence of the first impression to faces seen for less than a second.

**Methods**

Twelve participants (mean age, 19.83 years; 6 male) took part. Participants provided informed consent to procedures that were approved by the ethics committee of the University of York psychology department. All participants self-identified as culturally
Western. The stimuli and procedures were identical to Study 3, except now the faces appeared for only 500 ms before being replaced by a visual mask. The masking stimulus was a mosaic scrambled face, constructed by jumbling segments of the average face taken over the whole database, following Todorov et al. (2009). The faces were presented at 150 pixels in height (approximately 5 cm) with varying width to preserve aspect ratio and the visual mask was the size of the largest image (150 by 160 pixels).

Participants rated all 1,000 faces on the valence of their first impression (negative-positive). As in Study 3, faces appeared in random order, but trials were divided into five blocks with breaks to rest in between, since the way the Study 4 had to be set up meant that (unlike Study 3) participants could not take a break at any point. Participants were allowed to rate each face after the onset of the visual mask if necessary (following Todorov et al., 2009) but were encouraged to go with their ‘gut instinct’ (Todorov et al., 2005) and to try to respond while the face was still visible. All other aspects of presentation were as Study 3.

Results
The reliability of the valence ratings at 500 ms face presentation was good, with Cronbach’s alpha = .88 (p < .001). Moreover, the 500 ms valence ratings correlated highly with the unconstrained valence ratings from Study 3: r\text{female faces} = .84 (p < .001), r\text{male faces} = .85 (p < .001).

Again, we used previously collected ratings of dominance and trustworthiness (as for Study 3) to predict the valence ratings (now from a 500 ms viewing of the face) for male and female faces separately. Figure 5 shows the scatterplot of dominance and trustworthiness ratings against valence ratings.

To test the interactions between trait impressions and the gender of the face directly, we repeated the hierarchical regression procedure used in Study 3. We first used the perceived facial dominance ratings (centred) and face gender (0,1) to predict the valence of impressions to time-limited faces, and then entered the interaction between dominance and gender as a predictor in a second block. As in Study 3, the main effects of face gender and dominance were both significant, so that having a female face and looking low in perceived dominance were linked to more positive evaluations (see Table 3). Importantly, as in Study 3, the interaction between facial dominance and gender was significant and predicted a significant proportion of the variance in valence impressions over the main effects (see Table 3). To examine this significant interaction, we ran separate regressions predicting the valence of first impressions from dominance for male and female faces separately. For female faces, dominance (centred) significantly negatively predicted the valence of impressions: beta = -.34, t(498) = -8.19, p < .001, and explained a significant proportion of the variance of the valence of impressions: R^2 = .12, F(1, 498) = 67.03, p < .001. However, for male faces, dominance (centred) did not significantly predict the valence of impressions: beta = -.03, t(498) = -0.75, p = .454; or the variance in the valence of impressions: R^2 = .001, F(1, 498) = 0.56, p = .454. Our finding from Study 3 that dominance would be more negatively evaluated for female faces compared to male faces was therefore replicated in Study 4.

We also used the perceived facial trustworthiness ratings (centred) and face gender (0,1) to predict the valence of impressions to time-limited faces, and then entered the interaction between trustworthiness and gender as a predictor in a second block. As in Study 3, the main effect of trustworthiness was positive and highly significant, but the
main effect of face gender and their interaction were not significant (see Table 3). Again, this shows that the valence of first impressions is highly positively linked to the perceived trustworthiness of both male and female faces (see also Figure 5).

**Sex of participant**
Since the ratings were highly correlated between Study 3 and 4, we pooled the participants to enable an investigation of participant sex differences. We did this in order to demonstrate that our effects were not driven by one sex (e.g., they were not restricted to male participants’ judgements, for example). Here, our question was whether perceived dominance or trustworthiness would affect the valence of first impressions made by male \( (n = 12; \text{alpha} = .80) \) or female participants \( (n = 12; \text{alpha} = .92) \) differently. The main findings remained the same for female and male participants’ valence ratings when analysed separately. The interaction between facial dominance and gender was significant and predicted a significant proportion of the variance in the valence of first impressions over the main effects, so that dominance was evaluated more negatively in female faces for both male and female participants (both \( p < .001 \)). For both male and female participants, dominance (centred) significantly negatively predicted the valence of impressions made to female faces (both \( p < .001 \)) but not to male faces (both \( p > .26 \)). For both male and female participants, as expected, the interaction between trustworthiness and gender was not significant in predicting the valence of first impressions (both \( p > .7 \)).

**Discussion**
In Study 4, we again found that perceived dominance was more negatively valenced for female faces than male faces, even when the valence impression was made with only 500 ms exposure to the faces. This confirms that social categories can influence the first impression of facial trait dimensions. Moreover, we also found that trustworthiness was again highly correlated with valence at 500 ms exposure for both male and female ambient image faces, as predicted by Oosterhof and Todorov (2008). These results held across the pooled data from Studies 3 and 4 for male and female participants analysed separately. Finally, we found that unconstrained and 500 ms constrained ratings of the valence of overall impressions were highly correlated, adding to the literature showing that first impressions can be formed reliably from time-limited presentation of face stimuli (e.g., Willis & Todorov, 2006).

**GENERAL DISCUSSION**
We examined how gender stereotypes affect the valence of first impressions along fundamental dimensions of social face perception. Across four studies, we found that qualitative descriptions and quantitative ratings of faces differed in valence depending on the gender category and perceived social traits of the face. In particular, facially more masculine-looking and more dominant-looking female faces were described and rated more negatively than their male counterparts. Importantly, this pattern of findings was consistent across approaches involving unconstrained responses (Study 1 and Study 2) and a more conventional rating procedure (Study 3 and Study 4). This convergence of findings across such different methods strengthens the claim that it represents an important phenomenon. It is also worth highlighting that most of our findings were
similar across male and female participants. In particular, the negative evaluations of relatively dominant-looking female faces came from women as much as from men.

While the dimensional approach in facial first impressions (e.g., Oosterhof & Todorov, 2008; Sutherland et al., 2013) has made important theoretical advances in understanding the nature of these first impressions (Bruce & Young, 2012), the current research points to the need to integrate social category groups with facial trait dimensions. Certainly, the trustworthiness dimension explained most of the variance in the valence of the first impressions, supporting Oosterhof and Todorov’s (2008) claim that the trustworthiness dimension is primarily a judgement of valence. However, we also found clear evaluative differences based on gender category groups along the dominance dimension so that counter-stereotypical female faces were perceived more negatively than stereotypical male or female faces. In contrast, we found little evidence that counter-stereotypical male faces were perceived more negatively than stereotypical male or female faces.

Having demonstrated such clear gender-based evaluative differences, we need to consider potential mechanisms underlying them. Our findings suggest a number of possible avenues for future investigation. One possibility is that these evaluations could represent a simple association between facial cues and gender stereotypes linked to evaluations, similar to the finding that more prototypically ‘black’ faces are associated with congruent racial stereotypes and negative evaluations (e.g., Livingston & Brewer, 2002). This explanation is similar to the categorization mechanism of facial first impressions described by Secord (e.g., 1958). In categorization, if perceivers can assign a face to a category, they will then use knowledge associated with that category (stereotypes) to make inferences about the face.

An alternative account relates the evaluations of the faces to potential differences in the frequencies with which they might be encountered. Since women are rated as less facially dominant and more facially trustworthy than men (see Study 3), it is possible that familiarity or perceptual or categorization fluency might play a role in the negative evaluation of dominant and masculine female faces if these counter-stereotypical women are less frequently encountered or less easily categorized than their stereotypical counterparts. However, we note that this alternative account would not explain the relatively favourable evaluations of the feminine, non-dominant or trustworthy-looking men, who under this account would presumably also be less familiar or less easily categorized. Moreover, these stereotyping and familiarity/fluency explanations need not be mutually exclusive: counter-stereotypical faces may be less familiar due to their inherent facial structural features, or due to their lesser or biased representation in our cultural lives (e.g., perhaps as result of the under-representation of female leaders in general or in the media: Cracknell, 2013). By calling these ‘stereotypes’ we are not implying that they are arbitrary.

The specificity of the negative evaluations of dominant women might also index a ‘backlash’ effect paralleling social psychological findings that women who behave counter-stereotypically (e.g., assertively or dominantly) are disliked and receive discrimination in the form of less favourable hiring decisions compared to their stereotypical male and feminine female counterparts (e.g., Heilman, 2001; Rudman, 1998; Rudman & Glick, 2001). However, Sczesny and colleagues have found that, in the context of employment decisions, rather than facing discrimination, facially masculine women are preferred to facially feminine women for a masculine-typed occupation (Sczesny, Spreemann, & Stahlberg, 2006; Sczesny & Kühnen, 2004; see also Lammers, Gordijn, & Otten, 2009). It may be that facial backlash for dominant women occurs only for overall target evaluation,
rather than hiring decisions (and we note that Amanatullah & Tinsley, 2013 found this pattern from behavioural, non-facial target judgements).

If our results represent facial backlash, then it is currently unclear if this is similarly motivated to non-facial backlash (e.g., by system justifying beliefs: Rudman, Moss-Racusin, Phelan, & Nauts, 2012). Moreover, the concept of ambivalent sexism parallels backlash by describing how sexists can be both hostile and benevolent towards women, depending on whether or not they conform to traditional roles (Glick & Fiske, 1996, 1997). It would be interesting in future to examine whether individual differences such as the strength of system-justifying beliefs or the presence of ambivalent sexism moderate the strength of these evaluations, as backlash or ambivalent sexism accounts would predict.

Finally, we focused on global dimensions of facial first impressions across all four studies; gender stereotypicality, trustworthiness and dominance. However, it is worth noting that a multitude of cues (e.g., expression, attractiveness or age) may contribute towards the perception of these traits. In the current studies, we allowed the cues within faces to vary naturally in order to create the perception of main traits of interest (such as dominance), and then examined the consequences of these cues (the evaluation of these faces). We did this in order to let the cues contributing to judgments of interest vary naturally. Indeed, controlling for a cue to a trait necessarily changes the perception of the trait being studied, in turn changing the effect in question. Rather than examine cues or traits individually, we reasoned that the most theoretically powerful approach was to focus on the main dimensions implicated in the facial first impressions literature, and examine these as naturalistically as possible. As Keefe et al. (2013) have argued, one of the consequences of studying complex judgements such as trustworthiness is that one naturally also studies the effects of the cues towards such judgements (such as happiness expressions). Future work may nonetheless wish to try to isolate the influence of particular cues.

Conclusions
Using diverse methods to examine the valence of facial first impressions, the current set of studies demonstrate that how traits are evaluated in faces depends on the gender and stereotypicality of the target face. Specifically, we found that women with counter-stereotypical (less feminine or more dominant-looking) faces were evaluated more negatively than men or women with stereotypical faces. For men, being facially counter-stereotypical in trustworthiness or dominance did not lead to such negative evaluations. These findings are theoretically important in bringing together the face perception and social psychological literatures by indicating the need to integrate social category representations with dimensional accounts of facial first impressions.

Acknowledgements
This research was supported by an ESRC studentship to the first author [ES/I900748/1]. Thanks to John Towler and Philippa Howsley for collecting pilot data that were used to choose the stimuli in Study 2. We are grateful to the editor and two anonymous reviewers for their thoughtful comments and suggestions.
References


Prentice, D. A., & Carranza, E. (2002). What women and men should be, shouldn’t be, are allowed to be, and don’t have to be: The contents of prescriptive gender stereotypes. *Psychology of Women Quarterly, 26*, 269–281. doi:10.1111/1471-6402.01-1-00066


Received 27 February 2014; revised version received 15 July 2014