Every day, people are exposed to large amounts of social information that can only be meaningfully interpreted relative to similar occurrences. A target person can appear faster in the company of another fast standard or can appear slow in comparison to the same standard. The direction of this outcome is determined by the general similarity of target and comparison standard, which has been postulated to follow from standard extremity (Mussweiler, 2003). In the present research, we propose a standardized method to test this proposition along a variety of dimensions in the domain of face perception (i.e., Dominance, Trustworthiness, Competence, Extraversion, and Likeability). We will thus empirically revisit the question, whether extreme comparison standards always evoke contrastive judgments and whether moderate standards always produce assimilative judgments.

A jogger who passes you by might seem fast compared to yourself, but he will seem slow if a second runner sprints past him. The jogger's speed has perhaps not objectively changed, but the standard to which you compare him influences your judgment all the same. In this manner, we constantly use comparison standards to calibrate our judgments regarding the traits and abilities of others and ourselves (Dunning & Hayes, 1996; Festinger, 1954), up to the point whereby different authors postulate that virtually all judgments are to some extent comparative in nature (Kahneman & Miller, 1986; Mussweiler, 2003). Returning to the example of the jogger, the assessment of his speed seemed to increase when compared to your low speed, but decrease compared to the high speed of the other runner. These are both examples of calibrating a judgment in contrast to a standard, known as a contrast effect, where one's estimate moves away from the chosen standard (Upshaw, 1978). The opposite outcome of assimilation can also occur, with the standard acting as an anchor towards which judgments shift closer, as has also been found in countless studies (e.g. Mussweiler, & Strack, 2000; Brown, Novick, Lord, & Richards, 1992; Schwarz, & Bless, 1992). In fact, this assimilative pathway has been suggested as the default direction a comparison will take by some theories (Mussweiler, 2003), although a recent meta-analysis has proposed an opposite tendency (Gerber, Wheeler & Suls, 2018).

Which of the two comparison outcomes will occur (assimilation or contrast) is contingent on various explorable variables, such as the extremity of the standard on the relevant evaluative dimension (e.g., Herr, 1986; Herr, Sherman & Fazio, 1983). The selective accessibility model

**Assimilation and Contrast in Spontaneous Comparisons: Heterogeneous Effects of Standard Extremity in Facial Evaluations**

Paul Barker*, Ron Dotsch† and Roland Imhoff*‡

Judgments we make about others often depend on the standards we use as comparisons. Investigations into the outcomes of these comparisons and potential moderators have often been limited to single dimensions and preselected standards. The current work instead uses multiple evaluative facial dimensions and a multitude of comparisons. A series of 4 experiments \((N = 665)\) attempted to detect contrast from extreme (Study 1) and assimilation to moderate standards in within (Studies 2 and 3) and between-subjects designs (Study 4). Results showed inconsistent evidence for both comparison effects and significant heterogeneity across the evaluative dimensions that were sampled. An additional 5 studies \((N = 861)\) and a single-paper meta-analysis \((K = 7)\) revealed judgment dimension specific dynamics. Facial Extraversion produced both assimilation and contrast effects as expected; Dominance and Competence displayed only contrast; Trustworthiness showed only assimilation effects; and Likability presented no signs of either. The resulting implications for theory and measurement are discussed.

**Keywords:** Social Comparison; Assimilation; Contrast; Facial Evaluation
of exposure (Willis & Todorov, 2006; Ballew & Todorov, 2007) leading to real world outcomes (Todorov et al., 2005). Thus, facial images afford an ecologically valid way to expose participants to complex social traits in the blink of an eye.

We generated and tested digital facial stimuli and related items in an initial pilot-test (Pilot Study), followed by an attempt to measure consistent contrast from extreme (Study 1) and assimilation to moderate standards in both repeated (Studies 2 and 3) and between-subjects designs (Study 4). All findings were combined in a single paper meta-analysis (together with additional studies reported in detail in an online supplement) to assess the overall consistency of the patterns.

**Pilot Study: Stimulus Development**

To develop adequate research materials, we created computer-generated faces using custom scripts building on the FaceGen Software Development Kit, which allows for the generation and manipulation of 3D facial images. This was done along five of the psychological dimensions (Competence, Dominance, Extraversion, Likability, and Trustworthiness). These dimensions were previously found to be used spontaneously by respondents when describing novel faces (Oosterhof & Todorov, 2008), suggesting they are highly valid judgment dimensions for facial stimuli. Unique average neutral IDs were created and modified, producing versions of the same faces at several positions on the corresponding psychological dimension ranging from extremely low (–4SD) to extremely high (+4SD), with less extreme values at +/-1SD and 2SD, see Appendix A for examples. For the main studies, one set of stimuli was created this way for each dimension separately. The neutral faces of each set would be judged on the corresponding dimension while the non-neutral faces would be presented alongside them as the comparison standards. All items required open-ended absolute judgments, because closed scales themselves can enforce relative thinking (Mussweiler & Strack, 1999). We developed one open-ended question related to each of the underlying dimensions (see Table 1). In a pilot study, 82 participants aged between 18 and 67 years old (M = 34.63, SD = 11.10) and 41.5% female were recruited via MTurk, and gave open-ended estimates for each of these questions for faces at –4SD, –1SD, the midpoint, +1SD and +4SD of the respective dimension. Speaking to the overall validity of the assumed correspondence between the facial dimension and open-ended questions, there was an

![Table 1: Mean differences and standard errors by Extremity level.](image)

<table>
<thead>
<tr>
<th>Extremity</th>
<th>−1SD ΔM</th>
<th>SE</th>
<th>0SD ΔM</th>
<th>SE</th>
<th>1SD ΔM</th>
<th>SE</th>
<th>4SD ΔM</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>−4SD</td>
<td>−0.358</td>
<td>0.069</td>
<td>−0.756</td>
<td>0.068</td>
<td>−0.914</td>
<td>0.083</td>
<td>−1.143</td>
<td>0.081</td>
</tr>
<tr>
<td>−1SD</td>
<td>−0.399</td>
<td>0.056</td>
<td>−0.556</td>
<td>0.068</td>
<td>−0.785</td>
<td>0.068</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0SD</td>
<td>−0.157</td>
<td>0.059</td>
<td>−0.386</td>
<td>0.055</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1SD</td>
<td></td>
<td></td>
<td>−0.229</td>
<td>0.062</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note: All differences were significant, p < .01.*
overall linear trend from \(-4SD\) to \(+4SD\), \(F(4, 320) = 90.93, p < 0.001, \eta^2 = 0.532, 90\% \text{ CI}[0.468, 0.577]\), see Figure 1 and Table 1 for all pairwise comparisons.

The open-ended questions thus corresponded to facial dimensions sufficiently well overall, although there was some variation within dimension, with the Likability item capturing its dimension best and the Competence item doing least well overall; see Table 2 for linear trends per dimension and Appendix B for individual plots. In addition to these psychological dimensions, we also created facial stimuli for dimensions in the physiological domain (i.e., jaw width, mouth width, nostril width, nose length, and the distance between the eyes). As these reflect objective distances (millimeters), no pilot study was needed.

**Study 1**

In this initial investigation, we focused on the effect of extreme comparison standards on judgments of neutral targets and expected contrast effects in that the presence of an upward extreme comparison standard would yield lower judgments of accompanying neutral stimuli than for neutral stimuli judged with an extreme downward comparison standard. We tested this across five dimensions in the psychological and five dimensions in the physiological domain.

**Method**

**Participants**

An online sample of 162 U.S. based MTurk workers was recruited for a monetary compensation of $1.40 (approx. $6 p/h). This sample size allows the detection of moderately small effects within one group at two measurement instances for the main analyses, \(\eta^2 > 0.013, \rho = 0.50, \alpha = 0.05, \beta = 0.20\) (determined in G’Power; Faul, Erdfelder, Lang, & Buchner, 2007).

The final sample in this study was 44% female and was aged between 20 and 66 years (\(M = 33.81, SD = 9.93\)).

**Comparative Judgment Task (CJT)**

In each trial of the Comparative Judgment Task (CJT), participants evaluated a neutral face on one of the five dimensions in the psychological or physiological domain.

---

**Figure 1:** Means and standard errors of Z transformed responses for each extremity step measured.

**Table 2:** Separate analyses of linear trends of extremity per dimension with the related items.

<table>
<thead>
<tr>
<th></th>
<th>(Df)</th>
<th>(F)-value</th>
<th>(\text{Sig.})</th>
<th>(\eta^2) [90% \text{ CI}]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Competence (reversed):</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>“How often does the target make a mistake at work per month?”</td>
<td>1, 80</td>
<td>22.05</td>
<td>&lt;0.001</td>
<td>0.216 [0.095; 0.336]</td>
</tr>
<tr>
<td>Dominance:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>“How often a month does the target enforce his opinion?”</td>
<td>1, 80</td>
<td>95.68</td>
<td>&lt;0.001</td>
<td>0.545 [0.418; 0.631]</td>
</tr>
<tr>
<td>Extraversion:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>“How often does the target go out a month?”</td>
<td>1, 80</td>
<td>76.25</td>
<td>&lt;0.001</td>
<td>0.488 [0.095; 0.336]</td>
</tr>
<tr>
<td>Likability:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>“How often does the target offer help to a stranger a month?”</td>
<td>1, 80</td>
<td>123.57</td>
<td>&lt;0.001</td>
<td>0.607 [0.355; 0.583]</td>
</tr>
<tr>
<td>Trustworthiness (reversed):</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>“How often does the target deceive somebody every month?”</td>
<td>1, 80</td>
<td>72.43</td>
<td>&lt;0.001</td>
<td>0.475 [0.341; 0.572]</td>
</tr>
</tbody>
</table>
in an open-ended fashion. Simply asking participants to make an absolute judgment has been shown to be enough to engage individuals in comparative processing and produce both assimilation and contrast effects, even when standards were presented subliminally and without explicit prompting to compare (Mussweiler, Rüter & Epstude, 2004a). Alongside the judgment target an extremely high (+4SD) or extremely low (–4SD) version of the image was presented, acting as an upward or downward comparison standard, respectively. Participants were told that they needed to correctly identify the Judgment target (by clicking on a radio button below the face). This attention check was later used to exclude non informative responses, see Appendix C for an example trial. Each participant judged four targets per comparison direction for each dimension in the physiological and psychological domain, amounting to 80 trials.

Additional Measures

In addition, the Iowa-Netherlands Comparison Orientation Scale (INCOM; Gibbons & Buunk, 1999) was administered. This scale aims to measure an individual’s disposition to engage in social comparisons and consists of 11 items (α = 0.89) that are averaged to create an INCOM score, with higher scores indicating a higher tendency for comparisons. Although the items in the scale focus mainly on self-other comparisons regarding abilities and opinions, the underlying construct could potentially relate to a broader tendency to engage in all types of comparisons. Therefore, the INCOM was included at the end of the study for exploratory analyses in order to investigate if a higher comparison orientation would also be related to the strength of comparison effects found in the current work.

Furthermore, a final item at the end of the study was included to let respondents indicate if they did or did not engage in the study in earnest, and whether their data should be used. Participants were guaranteed their response to this item would not have any effect on their reward but would help exclude frivolous responses. Responses ranged from ‘Definitely do not use my data’ (1) to ‘Definitely use my data’ (4). Demographics such as Sex, Age and Education level were also measured.

Procedure

Participants initially were informed regarding the general procedure of the study and data storage policy before giving their consent. Following this, the general demographics were recorded. The CJT was then explained in detail, including two practice trials to allow participants to get properly acquainted with the procedure before the main batch of 80 trials followed in random order. Upon completion of the CJT, participants completed the INCOM scale and data quality item. Finally, they were debriefed, thanked, and given a code for their compensation.

Data treatment

Five participants who indicated their responses should not be used were not considered in the analyses. In the remaining data, two percent of trials with a failed attention check were not considered in further analyses. The remaining responses were used to generate z-scores separately per dimension and participant to allow comparison across dimensions with different scales and control for personal differences in response ranges. Scores were averaged to form aggregate z-scores for each factor. Missing average values, indicating a participant failed to respond correctly to a single of the attention checks across the factor level, were dropped in a list wise fashion for the main analyses. In total, this was the case for seven participants.

Results

Since the content of the psychological and physiological dimensions did not correspond in a meaningful way, two separate 2 (Comparison direction: up vs. down) × 5 (Dimension) factorial repeated measures ANOVAs were conducted for each domain separately. Contrast effects were found for the psychological dimensions overall, with significantly lower scores for upward (M = –0.063, SE = 0.020) than for downward (M = 0.065, SE = 0.020) comparisons, F(1, 151) = 10.52, p = 0.001, η² = 0.065, 90% CI [0.016, 0.136]; see Figure 2. The interaction effect of

![Figure 2: Means and standard errors of z-transformed responses separate for the two domains and the direction of the comparison.](image-url)
the direction and judgment dimension also reached significance, $F(4, 604) = 3.04$, $p = 0.017$, $\eta^2_p = 0.020$, 90% CI [0.02, 0.036], suggesting the presence of significant heterogeneity in effect sizes across the different evaluative dimensions. See the supplemental materials for dimension specific graphs.

Similar effects were found for the physiological dimensions, where upward comparisons showed significantly lower scores ($M = −0.090$, $SE = 0.018$) than downward comparisons ($M = 0.091$, $SE = 0.018$) overall, $F(1, 151) = 24.82$, $p < 0.001$, $\eta^2_p = 0.141$, 90% CI [0.065, 0.226]; see Figure 2. Again, the data presents with significant heterogeneity across dimensions, $F(4, 604) = 4.55$, $p = 0.001$, $\eta^2_p = 0.029$, 90% CI [0.007, 0.049]. See the supplemental materials for dimension specific graphs.

Correlational analyses
An overall difference score between upward and downward comparisons was calculated by subtracting the average of all downward $z$ scores from the average of upward $z$ scores. This was further done for psychological and physiological comparisons separately. These scores should reflect the extent to which participants used the available (albeit allegedly irrelevant) comparison standard, with higher negative scores reflecting stronger contrast effects and higher positive scores indicating assimilation. Psychological difference scores were weakly correlated to physiological difference scores, $r = 0.165$, $p = 0.039$, suggesting there is slight consistency in the use of comparison standards by participants regardless of domain. However, neither overall, psychological nor physiological difference scores were significantly correlated with the INCOM score, sex or age, all $r < 0.1$.

Discussion
This initial test of the paradigm has demonstrated that it has the ability to detect the use of comparison standards and has corroborated the hypothesis that more extreme standards lead to contrast effects. To truly test the delineated predictions regarding extremity of comparison standards, the paradigm must also include moderate standards that are predicted to lead to assimilation effects. The next study, therefore, attempted to measure assimilation as well as contrast effects conditional on the extremity of the standard.

Study 2
We conducted a second study to expand our results beyond the predicted contrast effect for extreme standards and also included moderate standards, for which one would expect assimilation. We thus added trials with moderate standards and narrowed our focus to the psychological (rather than physiological dimensions) to keep the study comparable brief without losing too much measurement precision. We eliminated the physiological (rather than psychological) dimensions as physical distances may be more accurately and objectively judged by participants on their screen. The presence of such an objective basis for the decision reduces the use of comparison standards (Festinger, 1954), and will limit its functionality of uncertainty reduction (Mussweiler & Posten, 2012). As in the initial study, extreme upward standards were hypothesised to lead to lower judgments of neutral targets than extreme downward standards (contrast effect). Additionally, in accordance with the described literature, judgments of neutral targets were hypothesised to be higher when moderate upward standards were available than when moderate downward standards were (assimilation effect). These effects were expected to be reflected in a cross-over interaction effect between direction and extremity.

Method
Participants
A second online sample of 160 US-based Mturkers was recruited, with similar power considerations in mind as explained in Study 1. Each participant received a monetary compensation of $1.47 for their participation (approx $6.30 p/h). The final sample in this study was 53% female and was aged between 21 and 72 years ($M = 36.43$, $SD = 11.12$).

Measures and Procedure
Using the same procedure as described in Study 1, each participant now made 16 open-ended judgments per dimension, each time accompanied by a relevant comparison standard (either extremely low, moderately low, moderately high or extremely high, with 4 trials each). Therefore, the full design was reflected by a $2$ (comparison directions) × $2$ (extremity levels) × $5$ (dimensions) design, each measured in four trials, for a total of 80 trials. In addition, the INCOM scale, data quality item, and demographics items were again administered. The procedure was identical to the one described in Study 1.

Data treatment
Four participants indicated their data should not be used and were therefore excluded. 1.9% of trials showed a failed attention check. $z$ scores were calculated based on the remaining data in the same way as in Study 1. Four participants had missing values after aggregation and were therefore not considered in the main analysis, leaving a final sample of 152.

Results
A $2$ (Comparison direction) × $2$ (Extremity) × $5$ (Dimension) RM-ANOVA failed to show the expected interaction between extremity and direction that would mark the existence of both assimilation and contrast effects, $F(1, 151) = 0.26$, $p = 0.610$, $\eta^2_p = 0.002$, 90% CI [0.000, 0.028]. A main effect of direction, $F(1, 151) = 13.44$, $p < 0.001$, $\eta^2_p = 0.082$, 90% CI [0.025, 0.157], was found, with lower scores for upward than for downward comparisons again indicating a contrast effect. Pairwise comparisons show the contrast effect is present both in the moderate ($\Delta M = −0.083$, $SE = 0.034$, $p = 0.017$, 95% CI [−0.150, −0.015]) and extreme condition ($\Delta M = −0.109$, $SE = 0.039$, $p = 0.006$, 95% CI [−0.186, −0.032]); see Figure 3. The analysis also yielded significant effects of the interaction between direction and dimension, $F(4, 604) = 7.43$, $p < 0.001$, $\eta^2_p = 0.047$, 90% CI [0.019, 0.071], extremity by dimension, $F(4,$
Correlational analyses
Overall difference scores between upward and downward comparisons were calculated for moderate and extreme trials separately by subtracting the average of downward z-scores from the average upward scores. Moderate difference scores and extreme difference scores did not correlate, $r = 0.004, p = 0.962$. Considering the presence of a contrast effect in both conditions, it is informative that this correlation is absent. This would imply participants that show contrast to the more extreme standards do not also show contrast to moderate standards for instance and would suggest there may be personal variation in the distinctive level of standard extremity necessary to induce contrast effects. In addition, the INCOM scores did significantly correlate with moderate difference scores, $r = 0.280, p < 0.001$, but only marginally with extreme difference scores, $r = 0.144, p = 0.072$. The positive correlation here indicates participants with higher INCOM scores actually showed less, not more contrast.

Discussion
In Study 2, comparison standards were used, but failed to show the expected assimilation effect to the moderate standards. Although this may be indicative of evidence against the predictions related to assimilation, it may also be due to the fact that 2SD is still considered extreme in the eyes of participants. Thus, the next study attempted again to find assimilation effects by reducing the extremity of moderate standards further.

Study 3
Considering the possibility that a variation of 2SD is still too extreme a standard to facilitate assimilation effects for some participants, we reduced the extremity of moderate standards to only 1SD. Again, contrast effects were hypothesised for extreme targets (+/−4SD), and moderate standards (+/−1SD) were hypothesised lead to assimilation effects, with moderate upward comparisons related to higher scores than moderate downward comparisons. These hypotheses should be reflected in a crossover interaction between direction and extremity.

Method
Participants
A new group of 162 U.S. based MTurk workers was recruited with a monetary compensation of $1.47 for their participation (approx $6.30 p/h), excluding all individuals who had participated in the initial study. This sample was again chosen with similar power considerations in mind as explained in Study 1. The final sample in this study was 51% female and was aged between 19 and 68 years ($M = 33.90, SD = 10.18$).

Stimuli & Design
The procedure was kept identical to Study 2 with the exception that moderate standards of only 1SD above or below the neutral targets were used. The full design again included upward and downward comparisons with moderate and extreme standards for each of the five dimensions. Four trials were conducted at each factor level for a total of 80 trials. In addition, the INCOM data quality item and demographics items were included.

Procedure
The procedure was identical to Study 2.

Data treatment
Five participants indicated their data should not be used and were therefore excluded. 2.9% of trials failed the attention check. Z-scores were calculated in the same way as in Study 2, with nine participants failing to give sufficient correct responses to calculate these scores (final $N = 148$).

Results
A 2 (Comparison direction) × 2 (Extreme) × 5 (Dimension) factorial repeated measures ANOVA showed the expected interaction between extremity and direction, $F(1, 147) =$
SD Δ Δ

Figure 4: Means and standard errors of Z transformed responses for all dimensions separate for the direction of the comparison and its extremity (+/–1SD, or +/-4SD).

Correlational analyses

The correlation between moderate difference scores and extreme difference scores was also not significant in this sample, \( r = 0.106, p = 0.19 \). This might be a reflection of low reliability of the open response items across judgments or may indicate that social comparison effects are variable across persons and that strong assimilation to moderate standards may not be related to strong contrast from extreme ones. If anything, the descriptively positive correlation would mean individuals consistently show either assimilation or contrast to both, though not significantly so in this sample. Correlations between the difference scores and the INCOM, sex or age also were non-significant, \( rs < 0.07 \).

Discussion

Study 3 provided initial evidence for assimilation in the current paradigm, suggesting that moderate standards must be confined, for the evaluative dimensions we tested, to be as much as 1SD from average. Furthermore, there were initial indications that downward comparisons were not as influential as upward ones, in line with the idea that upward standards are preferably selected over downward comparisons, as reported in a recent meta-analysis (Gerber, Wheeler & Suls, 2018), although the robustness of this finding is questionable as no previous studies showed a similar asymmetry.

One issue that may have reduced the effects for both comparison directions is that individual participants did not consistently show assimilation to moderate and contrast from extreme standards across the entire set of trials. The hypotheses for these studies showed contrast short of significance (\( \Delta M = -0.058, SE = 0.038, p = 0.132, 95\% CI [-0.133, 0.018] \)). Furthermore, it seems upward standards were more influential than downward ones in this sample (see Figure 4), also reflecting the main effect of extremity, \( F(1, 147) = 5.97, p = 0.016, \eta^2 = 0.039, 90\% CI [0.004, 0.101] \).

The interaction terms of dimension with direction, \( F(4, 588) = 8.42, p < .001, \eta^2 = 0.054, 90\% CI [0.024, 0.081] \), extremity, \( F(4, 588) = 3.95, p = 0.004, \eta^2 = 0.026, 90\% CI [0.005, 0.045] \), as well as the interaction of all 3, \( F(4, 588) = 4.750, p = 0.001, \eta^2 = 0.031, 90\% CI [0.082, 0.052] \), indicated significant variations in the effects across dimensions as in the previous study; see supplemental materials.

Study 4

Considering the above-discussed issues with the within-subjects design, this fourth study manipulated the extremity of the comparison standards between-subjects to assure participants only engage in one type of comparative process, similarity or difference testing, across the entire set of trials. The hypotheses for these studies

\[ \text{Correlational analyses} \]

\[ \text{Discussion} \]

\[ \text{Study 4} \]
remained the same as in the previous study, with contrast effects expected to occur for the condition exposed to extreme targets, while in the moderate condition, assimilation effects should occur. Again, these hypotheses should be reflected in a cross-over interaction between direction and extremity.

**Method**

**Participants**
A new sample of 181 US-based Mturkers participated for $0.84 as compensation (approx $7.20 p.h). With similar considerations for the effect size as in study one, this sample size was also predicted to give sufficient power between two groups each with two measurement instances to detect the effects of comparison direction in both groups (again determined in G’Power; Faul et al., 2007). The final sample in this study was 47% female and was aged between 18 and 72 years ($M = 37.20, SD = 11.40$).

**Stimuli & Design**
The stimuli and procedure were identical to Study 3. However, the design was changed to include extremity as a between-subjects factor to provoke only similarity or dissimilarity testing across the entire set. Participants were, therefore, randomly allocated to a condition with either only moderate or only extreme comparison standards. This resulted in a repeated measures design with four trials for each comparison direction and each of the five dimensions, amounting to 40 trials in total per participant and extremity as a between-subjects factor.

**Data treatment**
Five participants indicated their data should not be used and were therefore excluded. Of the trials, 4.8% failed the attention check and were not considered when the Z-scores were calculated in the same way as in Study 1. Fifteen participants failed to give sufficient correct responses to calculate z scores for all factor levels and were not included in the main analyses ($N = 161$).

**Results**
A 2 (Extremity) × 2 (Comparison direction) × 5 (Dimension) factorial mixed measures ANOVA, with standard extremity as the between-subjects factor but failed to show the expected interaction effect between direction and extremity, $F(1, 159) = 1.75, p = 0.188, \eta^2 = 0.011, 90\% CI [0.000, 0.052]$; see Figure 5. Descriptively, the expected pattern did occur, but pairwise comparisons showed no significant effect of direction in the moderate, ($\Delta M = 0.023, SE = 0.044, p = 0.595, 95\% CI [-0.063, 0.110]$), or extreme comparison condition, ($\Delta M = -0.058, SE = 0.044, p = 0.183, 95\% CI [-0.145, 0.028]$). Finally, the interaction term of dimension with direction, $F(4, 636) = 4.61, p = 0.001, \eta^2 = 0.028, 90\% CI [0.007, 0.047]$, was again found to be significant in this dataset. See supplemental materials for graphs per dimension.

**Correlational analyses**
Due to the between-subjects design no correlations between moderate and extreme difference scores could be calculated. Both difference scores did not significantly correlate with the INCOM, sex or age, $r < 0.11$.

**Discussion**
Contrary to the notion that the within-subject design would increase the consistent assimilative or contrastive use of the provided comparison standards by restricting participants to perform either similarity or difference testing, none of the previously found effects reached significance in this study, although the pattern remained consistent. This may partly be due to a slight loss of power in this between-subjects design, although the effect size estimate of the main interaction was very small and roughly in line with the previous estimates, and the simple contrasts also provided no separate evidence in either condition. Thus, the within-subjects design used previously does not perform less well in detecting the use of social comparison effects, or at minimum does not greatly underestimate effect sizes.

Figure 5: Means and standard errors of Z transformed responses for all dimensions, separately for the direction of the comparison and for each of the extremity conditions (+/−1SD, or +/−4SD).
The inconsistency of the results throughout the presented studies makes it difficult to reach strong conclusions about the mechanisms of the social comparison process. Nevertheless, a consistent interaction throughout the studies was found with the dimension factor. This result suggests there may be heterogeneity of the comparison effect across dimensions. However, this heterogeneity might also be a reflection of random fluctuations in the sampling of facial stimuli representing the target and standard for each dimension rather than the actual underlying comparison effects. In order to see if there are any overall consistent comparison effect across the studies controlling for stimulus level fluctuations, the next section will present a pooled mixed models analysis.

Pooled analysis
In a mixed-models analysis, we accounted for the random factors of participant, study and stimuli, which may have masked any evidence for consistent comparison patterns in the seemingly substantially different effects found across the studies. Given the small amount of stimuli used for each factor level, and in order to get the most accurate estimates of the fixed and random effects, it is paramount to include all relevant data that are available as part of this project. In addition to the four studies presented here, we included three additional studies conducted in this research line in the pooled dataset (all these studies are also described in detail in the supplemental materials). Both the main set and supplemental studies that are part of this pooled analysis are ordered chronologically.

Method
Participants
A total of 1099 subject made up the pooled dataset coming from seven separate studies. All participants were US-based MTurk workers ranging from 19 to 73 years of age ($M = 36.23, SD = 11.26$), of whom 50% were female.

Stimuli
A total of 144 unique facial image pairs were included overall, with 6 image pairs per dimension, except for Dominance (40 pairs) and Trustworthiness (80 pairs), as not all dimensions were investigated equally in the additional studies. It should be noted that the pooled data still contains relatively few stimulus level observations for each factor level, which could mean the ability to accurately estimate some fixed and random effects, as well as the power of statistical tests, might be lower than desirable (Bell et al., 2014).

Data treatment
Data treatment was the same as described throughout the initial studies and in the supplemental materials. However, participants were no longer removed due to missing values on the factor level due to the more flexible mixed models design.

Results
A mixed-models analysis was conducted (using the lme4 package in R; Bates, Maechler, Bolker, & Walker, 2015), including fixed effects for Comparison Direction, Extremity, and Dimension, with random slopes varying for subjects and study where possible, and random intercepts for all stimuli. An ANOVA using type 3 Sums of Squares and the Satterthwaite’s approximation for the degrees of freedom (realised with the lmerTest package in R; Kuznetsova, Brockhoff, & Christensen, 2017) showed no significant main effect of the direction of the comparison, $F(1, 91.96) = 1.529, p = 0.219$, as would be expected if only assimilation or only contrast effects occurred, nor was there an interaction with target extremity, $F(1, 99.90) = 0.951, p = 0.332$, which would indicate consistent assimilation and contrast (Figure 6). However, in line with the results from the separate studies, a significant effect was found for the interaction of dimension and direction, $F(1, 125.08) = 3.624, p = 0.008$, reflecting heterogeneous comparison effects across the dimensions in this pooled sample, much like in the individual studies.

![Figure 6: Estimated marginal means and standard errors of Z transformed responses over all studies shown separately for the direction of the comparison and for each of the extremity conditions.](image-url)
Discussion
In this pooled analysis, no evidence for consistent comparison effects were found across the studies and dimensions, in contrast to the initial expectations of this research line. However, an interaction between the direction of the comparison and the dimension in question suggested that, across all studies, any possible effects of comparison direction varied with the specific judgment that was made. This means the precise pattern of social comparison effects could be different depending on which dimension is being judged or which item is presented. We thus performed summative meta-analyses over all studies per dimension and extremity level to more clearly present the heterogeneities found in these data.

Meta-analysis
Due to the heterogeneous nature of the social comparison effect found in the discussed studies and pooled analysis, we summarised the effects for moderate and extreme standards separately and on all social dimensions for all studies used in the pooled analysis.

Moderate standards
For each of the studies that included moderate standards, the average responses to trials with moderate downward and moderate upward scores were calculated for all the judged social dimensions separately. Although moderate standards varied in Study 2 (+/−2SD) compared to those in the other studies (+/−1SD) as described in the relevant sections, these conditions were found similar enough to be included in the same analyses. The resulting scores were used in separate paired-sample t-tests to provide the within-subjects effects sizes (Cohen’s d) for use in the meta-analyses (utilizing the metafor package in R; Viechtbauer, 2010). Positive values indicate average judgments in the presence of an upward standard are higher than when a downward standard is shown (assimilation effect) and negative values indicate the opposite (contrast effect). Separate forest plots for all analyses are provided in Appendix D. All effect sizes were homogeneous across all studies for each dimension (See Table 4).

Difference scores
In addition to the overall patterns of assimilation and contrast, we investigated the meta-analytic correlational effects found between the INCOM and the difference scores for extreme and moderate standards separately in this section. All correlations with the INCOM scale were calculated and transformed into Fisher’s Z for moderate and extreme difference scores separately for use in the meta-analysis (again using the metafor package in R; Viechtbauer, 2010). For the extreme difference scores, no meta-analytic effect was found, \( z’ = 0.033, 95\% \text{ CI} \ [-0.035, 0.101], Z = 0.955, p = 0.34 \), with no significant signs of heterogeneity. For moderate standards, the meta-analytic effects were also nonsignificant, \( z’ = 0.009, 95\% \text{ CI} \ [-0.100, 0.118], Z = 0.165, p = 0.869 \), but showed high \( I^2 \) (60.21%) and significant heterogeneity, \( Q(5) = 15.260, p = 0.009 \). Further analyses indicated Study 2 was the main cause of this heterogeneity, likely due to the fact that the moderate standard was not reached significance in this small sample, \( Q(4) = 6.531, p = 0.163 \).

Extreme standards
In an identical fashion, average judgments in the presence of extreme downward and extreme upward standards were calculated for all the judged social dimensions in each study, with extreme standards separately. Separate paired-sample t-tests again provided the within-subjects effects size (Cohen’s d) used in the meta-analyses, with positive values indicating assimilation effects and negative values indicating contrast effects. The effects were homogeneous across all studies for each dimension (See Appendix D for forest plots).

Dominance and Extraversion were the only dimensions that showed consistent contrast effects to extreme standards across the studies, with none of the other dimensions showing any consistent effects in either direction (Table 3). Table 3: Facial dimensions and their meta-analytic assimilation and contrast effects across all the studies for Moderate standards.

<table>
<thead>
<tr>
<th>Facial dimension</th>
<th>( K )</th>
<th>( p )</th>
<th>Meta-analytic ( d ) (95% CI)</th>
<th>( Z )</th>
<th>( p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Competence</td>
<td>4</td>
<td>0%</td>
<td>(-0.166 [-0.299; -0.034])</td>
<td>(-2.456)</td>
<td>0.014</td>
</tr>
<tr>
<td>Dominance</td>
<td>5</td>
<td>17.81%</td>
<td>(-0.266 [-0.405; -0.127])</td>
<td>(-3.742)</td>
<td>0.001</td>
</tr>
<tr>
<td>Extraversion</td>
<td>4</td>
<td>0%</td>
<td>(0.168 [0.035; 0.302])</td>
<td>(2.469)</td>
<td>0.014</td>
</tr>
<tr>
<td>Likeability</td>
<td>4</td>
<td>0%</td>
<td>(-0.063 [-0.199; 0.073])</td>
<td>(-0.903)</td>
<td>0.367</td>
</tr>
<tr>
<td>Trustworthiness</td>
<td>6</td>
<td>0%</td>
<td>(0.277 [0.169; 0.384])</td>
<td>(5.041)</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>
Table 4: Facial dimensions and their meta-analytic assimilation and contrast effects across all the studies for Extreme standards.

<table>
<thead>
<tr>
<th>Facial dimension</th>
<th>K</th>
<th>F</th>
<th>Meta-analytic $d_{4}$ (95% CI)</th>
<th>Z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Competence</td>
<td>5</td>
<td>0%</td>
<td>0.006 [–0.119; 0.130]</td>
<td>0.089</td>
<td>0.929</td>
</tr>
<tr>
<td>Dominance</td>
<td>6</td>
<td>0%</td>
<td>–0.252 [–0.365; –0.139]</td>
<td>–4.360</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Extraversion</td>
<td>5</td>
<td>0%</td>
<td>–0.217 [–0.338; –0.096]</td>
<td>–3.519</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Likeability</td>
<td>5</td>
<td>0%</td>
<td>–0.042 [–0.164; 0.080]</td>
<td>–0.670</td>
<td>0.503</td>
</tr>
<tr>
<td>Trustworthiness</td>
<td>7</td>
<td>0%</td>
<td>–0.026 [–0.131; 0.076]</td>
<td>–0.520</td>
<td>0.603</td>
</tr>
</tbody>
</table>

for this study was at 2SD rather than 1SD like in the subsequent studies. Removing this study reduced heterogeneity to non-significant levels, $Q (4) = 1.041$, $p = 0.904$, and an $F$ of 0%, but left the conclusions unaltered, $z^2 = –0.050$, 95% CI [–0.124, 0.025], $Z = –0.1297$, $p = 0.195$. Taken together, these results offer no evidence that inter-individual differences in the disposition for comparative thinking about one’s own opinions and abilities is related to a broader tendency to spontaneously compare others consistently.

**Discussion**

The results of the separate meta-analyses describe more clearly the heterogeneity in comparison effects across different dimensions, but also show a remarkable consistency of effects within each dimension which were not apparent when evaluating the studies separately. The Likability dimension seems unaffected by comparison standards of any of the presented extremity conditions. Judgments on the Dominance and Competence dimensions show exclusively contrast to comparison standards as moderate as only one standard deviation away from the target, with only the Dominance dimension also showing contrast to more extreme standards of up to four standard deviations. Furthermore, the Trustworthiness dimension only showed assimilation effects to moderate standards, but did not show contrast from the more extreme standards. The only judgment dimension in our sample that showed the expected pattern of assimilation to moderate and contrast away from extreme standards was that of Extraversion. These results suggest the moderating effect of extremity may be at best dimension or judgment sensitive, at least to the extent that the different tested dimensions could have significantly varying thresholds for what is considered extreme or moderate.

**General discussion**

Despite the difficulty in finding the predicted pattern of social comparison effects moderated by target extremity in the separate studies looking across all dimensions, the novel face-judging paradigm did manage to successfully detect both consistent assimilation and contrast effects on a number of specific facial dimensions. Although the dimension of Extraversion showed assimilation to moderate and contrast from extreme standards, the majority of dimensions showed only assimilation (Trustworthiness), only contrast (Dominance, and Competence), or no effect at all (Likability). These results suggest the moderating role of extremity could be fundamentally influenced by the dimension of interest. This unexpected variation in the comparison patterns across the tested judgment dimensions raises new questions about the cause of these dimension-specific comparison dynamics.

As a first possibility, one could note the inherent entangled nature of some facial dimensions and related social categories. For instance, with increasing trustworthiness faces become more feminine, whereas with increasing dominance they look more masculine. More extreme positions on these dimensions may therefore affect not only the extremity of the dimension itself, but could also affect the perceived category membership, another important moderator for the comparison direction (Brewer & Weber, 1994; Mussweiler & Bodenhausen, 2002). Although this might be an issue with the more extreme faces in general and remains an issue with the use of facial dimensions, as noted by Oosterhof and Todorov (2008) themselves, this cannot fully explain the current findings. In fact, the influence of this proposed effect would likely be in line with the effects predicted for target extremity, not counter to them. For instance, more extreme standards that include opposing category membership information should increase the likelihood of contrast effect as initial dissimilarity judgments become more common, while moderate standards with arguably the same category membership information should be unaffected and still produce assimilation. Indeed, many studies that successfully showed the moderating effects of standard extremity have used stimuli that include clear category information (e.g. Shirley Temple vs. Hitler in Herr, 1986; or Michael Jordan vs. Bill Clinton in Mussweiler, Rüter & Epstude, 2004b). With this in mind, it may be even more surprising that the current research did not find this pattern for any dimension other than for Extraversion. In fact, the results even showed contrast away from moderate standards with seemingly no potentially discrepant category membership for two of the dimensions, while showing no contrast from the more extreme and potentially most discrepant faces for four dimensions.

A second explanation is that the conceptual content of the dimensions themselves might prompt initial similarity
or difference judgments. A dimension such as Dominance could be seen as an inherently asymmetric relational construct. One can only be dominant over a more submissive other, but can be neither in isolation. Therefore, the informational value the Dominance dimension expresses might be fundamentally linked to differences and contrastive judgments. Other dimensions that might show similar inherently entail relational differences (e.g. status) could show the same pattern by making judgments of differences more likely. In contrast, one person’s trustworthiness does not necessarily imply much about another person’s, meaning dyads can logically be composed of two equally trustworthy people. Some level of interpersonal closeness with others might even be inferred for a trustworthy person, which might lead to similarity focusses and the assimilation we detected (Alves, Koch & Unkelbach, 2016; 2017). Such dynamics might be concept specific, or could point to larger underlying principles of human evaluation, as similar dimensions that map on closely to these two have been found in models of interpersonal perception (e.g. Affiliation & Dominance in Wiggins, Phillips & Trapnell, 1989) and intergroup perception (e.g. Warmth & Competence in Fiske, Cuddy & Glick, 2007; or Communion & Agency in Koch et al., 2016; for a critical discussion of their orthogonal nature see Imhoff & Koch, 2017).

However, it is important to note that this theorizing is speculative, and these varying dynamics could also be even more specific to the items used to operationalise the dimensions in these studies. The paradigm was designed to measure overall comparison effects and not to measure the unexpected variations of the effect on individual dimensions accurately. Despite the testing of the items to capture the underlying dimensions to a reasonable extent overall, some dimensions were more accurately represented by the items than others, and only a single item was used per dimension. This means the current research cannot fully disentangle dimension-specific effects from judgment-item specific effects. Although this issue indeed limits the generalisability of the results to the items that were used in this study, the strong variability in comparison effects at either level highlights the need for a broader selection of items and dimensions in comparison research, instead of relying on single measures and hand-picked standards.

Note that the variability also need not necessarily imply substantially different dynamics which are counter to the general principle of assimilation to moderate and contrast from extreme standards (as outlined in prominent models such as the SAM), but that the threshold of what constitutes an extreme (or dissimilar) standard might drastically differ per dimension, item, participant and measurement time. If one standard is judged as extreme by one participant, but moderate by another, their aggregated effects could cancel out and appear as a lack of social comparison effects overall. This leads to another important critique of the paradigm, in which the selection of moderate standards and extreme standards across all dimensions assumes the threshold is uniform among all persons and all dimensions. However, if there are dimension- or item-specific thresholds for assimilation and contrast, such a one-size-fits-all approach cannot test the predictions completely, as there may always be some small area left untested where the effects could occur.

With a similar logic, this highlights a lack of boundary conditions in the current literature as a whole. By relying on pre-selected exemplars as the moderate or extreme standards in previous work, the limits of what exactly constitutes a moderate or an extreme standard have not been defined, nor is there agreement on how these parameters might be estimated. This lack of clarity poses a problem for comparison research, as it often leaves the comparison outcomes themselves as the only way to judge if a standard is moderate or extreme enough for the predicted comparison direction to occur. Post-hoc judgments of the standards extremity, therefore, remain an almost unavoidable possible explanation for unexpected findings. For example, given a predicted contrast effect does not appear, it could be said that the standard was simply not extreme enough or too extreme to be relevant. Without a systematic approach to bindingly define moderate and extreme in some manner or by having some standard for estimating these parameters, a failure to show the predicted effect can always be seen as indicative of a problem in the operationalization of ‘moderate’ and ‘extreme’ instead of evidence against the theory, thus speaking to the validity of auxiliary theories about operationalisation rather than theory proper (Meehl, 1990).

If no boundary conditions are specified for either direction, this leaves little room in the way of falsifying any theories that predict effects conditional on target extremity. This demonstrates that result-contingent categorization of standards as sufficiently extreme or moderate creates a problem for tests of comparison theories across multiple dimensions, and highlights the need for more in-depth and rigorous testing of these basic comparison findings. Having noted this critique, the scope of the current research cannot resolve this theoretical issue, but does demonstrate that result-contingent categorization of standards as sufficiently extreme or moderate creates a problem for any real test of comparison theories across multiple dimensions, leaving their predicative power limited.

Notwithstanding these issues, the current work has shown that the threshold to come to an assimilative compared to contrastive mindset need not be consistent across different judgments and potentially conceptual dimensions, leading to markedly different data patterns. In doing so, it should at least nuance generalized claims of comparison patterns, as they seem not to easily translate to every evaluative judgment. This highlights the need for more in depth, critical and rigorous tests of basic findings in this research area, taking into account the possible heterogeneous nature of effect sizes, as many findings may not extend beyond the items or judgment dimensions that were tested.
Appendix A

Appendix B

Appendix C
Appendix D

- Competence: Moderate
  - Study 1: 0.19 (0.45, 0.87)
  - Study 2: 0.31 (0.26, 0.67)
  - Study 3: 0.20 (0.30, 0.32)
  - Study 4: 0.25 (0.20, 0.33)
  - Study 52: 0.18 (0.04, 0.39)

- Competence: Extreme
  - Study 1: 0.09 (0.09, 0.27)
  - Study 2: 0.04 (0.01, 0.22)
  - Study 3: 0.04 (0.03, 0.37)
  - Study 4: 0.02 (0.04, 0.41)
  - Study 52: 0.05 (0.01, 0.20)

- Dominance: Moderate
  - Study 1: 0.19 (0.12, 0.38)
  - Study 2: 0.20 (0.15, 0.36)
  - Study 3: 0.21 (0.20, 0.30)
  - Study 4: 0.14 (0.13, 0.20)
  - Study 52: 0.12 (0.11, 0.20)

- Dominance: Extreme
  - Study 1: 0.27 (0.93, 0.19)
  - Study 2: 0.25 (0.52, 0.90)
  - Study 3: 0.23 (0.46, 0.83)
  - Study 4: 0.20 (0.45, 0.90)
  - Study 52: 0.13 (0.81, 0.82)

- Extraversion: Moderate
  - Study 1: 0.17 (0.23, 0.32)
  - Study 2: 0.12 (0.12, 0.38)
  - Study 3: 0.23 (0.23, 0.36)
  - Study 4: 0.14 (0.13, 0.34)
  - Study 52: 0.11 (0.11, 0.24)

- Extraversion: Extreme
  - Study 1: 0.17 (0.48, 0.19)
  - Study 2: 0.27 (0.52, 0.62)
  - Study 3: 0.13 (0.30, 0.12)
  - Study 4: 0.09 (0.02, 0.42)
  - Study 52: 0.32 (0.54, 0.12)

- Likability: Moderate
  - Study 1: 0.09 (0.13, 0.11)
  - Study 2: 0.09 (0.20, 0.38)
  - Study 3: 0.11 (0.43, 0.42)
  - Study 4: 0.15 (0.36, 0.90)
  - Study 52: 0.06 (0.20, 0.07)

- Likability: Extreme
  - Study 1: 0.13 (0.48, 0.44)
  - Study 2: 0.83 (0.23, 0.33)
  - Study 3: 0.82 (0.24, 0.27)
  - Study 4: 0.15 (0.53, 0.33)
  - Study 52: 0.99 (0.01, 0.12)

- Trustworthiness: Moderate
  - Study 1: 0.22 (0.04, 0.48)
  - Study 2: 0.41 (0.15, 0.82)
  - Study 3: 0.20 (0.14, 0.26)
  - Study 4: 0.27 (0.30, 0.40)
  - Study 52: 0.13 (0.17, 0.34)

- Trustworthiness: Extreme
  - Study 1: 0.07 (0.04, 0.10)
  - Study 2: 0.07 (0.03, 0.10)
  - Study 3: 0.13 (0.29, 0.32)
  - Study 4: 0.12 (0.06, 0.12)
  - Study 52: 0.04 (0.18, 0.20)
  - Study 53: 0.05 (0.36, 0.30)
Data Accessibility Statement
All data files, analyses and additional materials can be found at: https://osf.io/gb9k3.

Notes
1 The effect sizes obtained from all simple comparisons per dimension using these z-scores and for ones using the raw scores were highly correlated throughout the studies, r = 0.813, yielding similar results.
2 Data and analyses for all studies conducted in the research line can be found in the supplemental materials on OSF. Some studies reported there also included a baseline measure for targets presented without a comparison standard. For the pooled- and meta-analysis, however, we did not consider this in order to compare similar effects across all studies. An additional study (Study S4) deviated from all other studies in not using open-ended measures, but a non-linear Likert scale as a dependent measure and was thus not included in these analyses either.
3 There is no straightforward, agreed-upon way of calculating effect size measures like $\eta_p^2$ for mixed models due to issues surrounding the inclusion and decomposition of the variance from the random effects. Therefore, none are shown in this section.
4 The results of this meta-analysis are again bound to the stimuli used for each dimension. For the contrasts of the pooled analysis one can refer to the supplemental materials. Assimilation to moderate standards for Trustworthiness, and contrast from extreme ones for Dominance remained significant with no other effects reaching significance. However, only two stimulus observations can be used to estimate the random effects for each factor level for most dimensions, leading to very large standard errors. Nevertheless, we could not exclude the possibility that these effects would not generalize to the larger stimulus population.

Funding Information
This project was supported by a DFG Grant (IM147/3-1) awarded to Roland Imhoff as part of the DFG Research Unit “Relativity in Social Cognition” (FOR 2150).

Competing Interests
The authors have no competing interests to declare.

Author Contributions
All authors contributed to the conceptualization and design of the studies. Stimuli were provided by RD. RI and PB both helped to run the studies, analyze the data, and write the manuscript, with critical feedback throughout from RD.

References


