

**A New Standard for Comparison Research:
Uncovering the Dynamic Interactive Pattern of Comparative
Judgments and its Implications.**



Inauguraldissertation

zur

Erlangung des Doktorgrades

der Humanwissenschaftlichen Fakultät

der Universität zu Köln

nach der Promotionsordnung vom 18.12.2018

vorgelegt von

Paul Barker

aus

Reading (Vereinigtes Königreich)

Tag der Abgabe: 23. Juli 2020

Diese Dissertation wurde von der Humanwissenschaftlichen Fakultät der Universität zu Köln im
December 2020 angenommen.

Erstgutachter: Prof Dr. Roland Imhoff

Zweitgutachter: Prof Dr. Christian Unkelbach

Tag der Disputation: 17.12.2020

Statement of Published Work

Chapter 2 is based on the following published manuscript:

Barker, P., Dotsch, R., & Imhoff, R. (2020). Assimilation and Contrast in Spontaneous Comparisons: Heterogeneous Effects of Standard Extremity in Facial Evaluations. *International Review of Social Psychology*, 33, 11. <http://doi.org/10.5334/irsp.402>.

For this work, the other authors and I developed the conceptual idea together. I then set-up the studies, collected and analysed the data, and wrote the manuscript with valuable comments from my fellow authors to improve the work.

Chapter 3 is based on a submitted manuscript:

Barker, P., & Imhoff, R. (2020). *Comparing Curves: Describing the Dynamic Interaction of Assimilation and Contrast in Facial Evaluations*. Manuscript submitted for publication.

Here I extended the idea developed in the manuscript from Chapter 2, again set-up the studies, collected and analysed the data, and wrote the manuscript. Critical feedback was given throughout by the second author.

Chapter 4 is a pre-print of an unfinished manuscript:

Barker, P., & Imhoff, R. (2020) *Moving the line: Experimental shifts in the dynamic interactive pattern of assimilation and contrast for facial stimuli*. Manuscript in preparation.

The idea was further developed by myself and my co-author, after which I took charge of the study design, collected and analysed of the data, and writing the manuscript.

Paul Barker, Cologne, July 2020

Abstract

Social judgments of the self or others are often made in comparison to some standard in the environment. How these standards influence our judgments depends heavily on their relative standing on the evaluative dimension of interest compared to the target of the judgment. Despite this consequential role, researchers have often selected items and comparison standards somewhat arbitrarily, either ignoring or simplifying their influence substantially. The current dissertation will argue that this poses serious issues for the generalisability and validity of such findings, preventing strong tests of theory. Instead, it will offer a new more holistic approach to the investigation of the comparison process which takes this key variable into account. In the first chapter, a brief overview of the comparative process and the influence of comparison standards will be given to highlight these potential issues. Chapter 2 will then show that standards with the same relative distance to the target can potentially lead to opposing comparison effects simply due to item selection alone. Chapter 3 confirms this heterogeneity at the item level, and uncovers the dynamic interactive pattern of assimilation and contrast, showing that the dichotomisation of the relative distance between target and standard into ‘Moderate’ or ‘Extreme’ standards can be problematic. Chapter 4 will show how this dynamic pattern shifts in response to other moderating variables, like a comparative focus on similarities or differences. Thereby, this chapter also offers a new paradigm that can robustly test theoretical predictions, while avoiding the aforementioned pitfalls. Finally, the last chapter will offer some concluding thoughts about the implications for the literature, limitations of the current work, and offer recommendations for future research.

Keywords: Social comparison, Assimilation, Contrast, Validity, Facial evaluations.

Acknowledgments

No man is an island, and equally so, no dissertation can be written without support from others. I would like to take this opportunity to thank some of those that helped make this possible:

Firstly, thank you, Roland, for giving me this opportunity. I truly appreciate the feedback and support you've offered, and the confidence you showed in my ability to work independently. I think, if nothing else, we've proven that long distance relationships can work.

I would also like to thank you, Christian, for accepting the role as my second supervisor as well as for leading the research unit and SoCCCo. It has always felt like a great environment for scientific exchange to me.

Although some of you have moved on to greener pastures; thank you to all my colleagues and friends from over the years. In particular, Matt, for all the laughs and banter in the office. Boris, thank you for all the nights out and coffee breaks; I am glad to have made such a good friend along the way. Thank you, Alexandra, for always knowing the answers to all my questions and helping me with my struggles in German. I also want to thank my closest friends, Daan, Steven and Chris, for the good times and for staying in touch from so far away.

Robert, our arguments as children have, without a doubt, been instrumental to my critical thinking. So thank you for keeping me sharp and, of course, for being my brother through it all.

Thank you to my Parents, for everything. You have always been there and have supported me throughout my life and all of my misguided interests. I could not have wished for more loving parents, and I could not have achieved any of this without you both.

And of course, my Annie, thank you for all your support and patience. It has meant so much to have you by my side through this. So thank you for being my partner and not rejecting me like so many reviewers have.

Table of Contents

Chapter 1: Introduction	1
1.1. Comparative Thinking	3
1.2. Theories of Comparative Outcomes	5
1.3. The Curvilinear Effect of the Standard.....	10
1.4. The Current Aim.....	14
Chapter 2: Heterogeneous Effects across Judgments	18
2.1. Pilot Study: Stimulus Development	22
2.2. Study 1	24
2.2.1. Method.....	24
2.2.2. Results.....	27
2.2.3. Discussion.....	28
2.3. Study 2	29
2.3.1. Method.....	29
2.3.2. Results.....	30
2.3.3. Discussion.....	32
2.4. Study 3	32
2.4.1. Method.....	32
2.4.2. Results.....	33
2.4.3. Discussion.....	34
2.5. Study 4	35
2.5.1. Method.....	36
2.5.2. Results.....	37
2.5.3. Discussion.....	38
2.6. Pooled analysis.....	39
2.6.1. Method.....	39
2.6.2. Results.....	40
2.6.3. Discussion.....	41
2.7. Meta-analysis.....	42
2.7.1. Moderate standards	42
2.7.2. Extreme standards	43
2.7.3. Difference scores	44

2.7.4.	Discussion.....	45
2.8.	General discussion	45
Chapter 3: The Dynamic Interactive Pattern of Assimilation and Contrast		51
3.0.1.	What constitutes an extreme standard?	54
3.0.2.	The present research	58
3.1.	Study 1	61
3.1.1.	Method.....	61
3.1.2.	Results.....	64
3.1.3.	Discussion.....	67
3.2.	Study 2	69
3.2.1.	Method.....	69
3.2.2.	Results.....	71
3.2.3.	Discussion.....	73
3.3.	Study 3	73
3.3.1.	Method.....	74
3.3.2.	Results.....	75
3.3.3.	Discussion.....	77
3.4.	Study 4	77
3.4.1.	Method.....	78
3.4.2.	Results.....	81
3.4.3.	Discussion.....	84
3.5.	Study 5	85
3.5.1.	Method.....	85
3.5.2.	Results.....	86
3.5.3.	Discussion.....	87
3.6.	General discussion	88
Chapter 4: Shifting Standards of Comparison		94
4.0.1.	The dynamic interactive pattern of assimilation and contrast.....	97
4.0.2.	The present research	102
4.1.	Study 1	103
4.1.1.	Methods.....	104
4.1.2.	Results.....	107
4.1.3.	Discussion.....	110

4.2. Study 2	110
4.2.1. Method.....	111
4.2.2. Results.....	111
4.2.3. Discussion.....	113
4.3. General discussion	114
Chapter 5: General Discussion	120
5.1. Implications.....	122
5.2. Limitations & Future Research.....	126
5.3. Conclusion.....	132
References.....	133
Appendices.....	151
Appendix A.....	151
Appendix B.....	152
Appendix C.....	153
Appendix D.....	154
Appendix E	156

Chapter 1: Introduction

In many ways, our thoughts and experiences are relative in nature. The most visceral examples of this are experiential in nature. Think, for instance, of walking into a 20 °C degree room after being in a snowstorm versus coming back from a hot summer's day. Although the temperature of the room has not changed, our evaluation of its temperature can vary widely in comparison to our recent experiences. This relative way of thinking has been suggested to extend to almost all types of stimuli (Kahneman & Miller, 1986), with the cognitive ability to compare already present at a very early age (Gentner & Rattermann, 1991).

The fundamentally relative nature of human cognition can also be seen in the social judgments we make about ourselves and others. When we want to know how attractive we look or how smart our colleagues are, we will likely think of some standard to use as a comparison to help us make an accurate assessment. This phenomenon of social comparison, first described in a formal theory by Festinger (1954), has grown into a large and complex body of research with a substantial portion focused on how these comparisons affect our judgments (see Mussweiler, 2003). Perhaps unsurprisingly, one of the most influential characteristics of a standard is its standing on the comparative dimension with respect to the target of the judgment. Whether a standard is above or below the target on the evaluative dimension will, for instance, influence the direction in which the judgments will move (Mussweiler & Strack, 2000), while the exact distance between the two is known to increase the strength of this effect (Bless, Schwarz, & Wänke, 2003). For example, think of comparing a cat to a dog versus comparing it to an elephant. However, the relative standing of the standard does not simply have a linear effect, rather it can also itself moderate the direction in which the judgment is moved. Where moderate

standards pull the judgments closer to them (i.e. an assimilative effect), more extreme ones will push them further away (i.e. a contrast effect; Herr, 1986).

Despite this fundamental and complex role, the extremity of comparison standards has often not been taken into sufficient account as many researchers still use single judgments and somewhat arbitrarily selected standards. When its role is acknowledged, it is often treated as dichotomous while including only vague boundaries surrounding what might theoretically constitute a moderate or an extreme standard. This vagueness presents not just an issue for the investigation of the extremity variable itself, but also prevents strong tests of theory as this ever present confounding variable can be raised as an alternative explanation for unexpected results.

The current dissertation will, therefore, argue that the generalisability of previous findings is hampered by the narrow selections of items, dimensions, and standards. It will show that these issues form significant barriers for true tests of the underlying theory, and will endeavour to remedy these issues by providing tangible design recommendations and proposing a new standard for comparison research. To achieve this, the current work will initially highlight the large heterogeneity in comparison effects across different judgments using a newly developed Comparative Judgment Task (CJT) in Chapter 2. Chapter 3 will then demonstrate the issues caused by the ambiguous definition of standard extremity while uncovering the full dynamic interactive pattern of assimilation and contrast using a curve fitting approach, thereby providing the first concrete definition of what may constitute an extreme standard. The implications of this dynamic pattern will become more apparent in Chapter 4, where the efficacy of the CJT-paradigm to test other moderating variables will be shown offering one potential way of conducting stronger tests of theory in future research. Finally, Chapter 5 will discuss the

implications of the findings for the literature and future research, while also discussing the limitations of the current work and providing some final guidelines for study design.

1.1. Comparative Thinking

The relative nature of human cognition has been acknowledged in psychology almost since its founding (Wundt, 1897/ 1980), with more recent evidence suggesting that it is a fundamental constraint on information processing at the neuronal level (Carandini & Heeger, 2012). Indeed, many people will be familiar with visual manipulations, such as the Ebbinghaus illusion (Figure 1.1), which make the relative nature of our experiences immediately clear. The effects of relative thinking are by no means limited to such visual trickery, but extend broadly into areas as diverse as categorisation (Nosofsky, 1986), intergroup bias (Alves, Koch & Unkelbach, 2018), stereotyping (Biernat, 2003), decision making (Kahneman & Miller, 1986) and affect (Higgins, 1987), with children as young as five months using similarities to infer things about the world (Baillargeon, 1991).

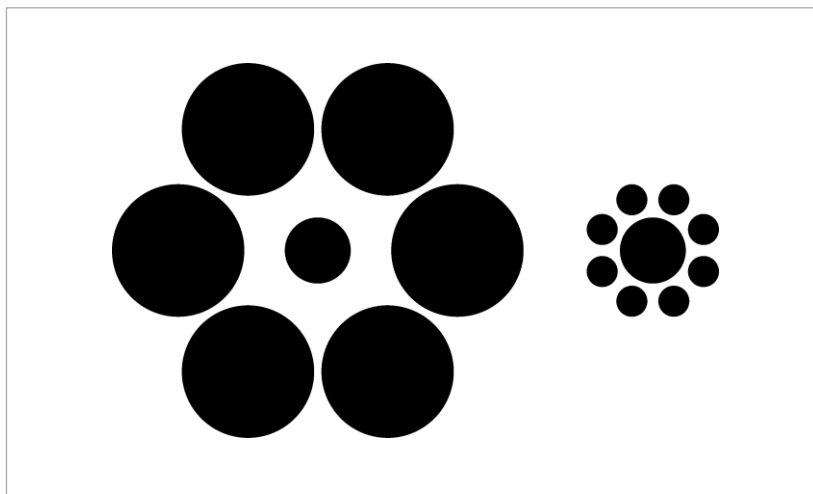


Figure 1.1. Ebbinghaus illusion: the central circle is objectively the same size in both compositions, but appears larger in the right than in the left one (i.e. a contrast effect).

Considering how fundamental relative thinking seems to be for the processing of information, it is not surprising that judgments about ourselves and others are also profoundly influenced by the standards that we use as comparisons. These social comparisons have been found to happen spontaneously when evaluating others (Dunning & Hayes, 1996) or the self (Festinger, 1954; Mussweiler & Rüter, 2003). This may be due to the fact that they seem to offer a distinct processing advantage in many situations (Mussweiler & Epstude, 2009) and take place very early on in the processing of information (Ohmann, Stahl, Mussweiler & Kedia, 2016). In fact, the use of comparison standards is so deeply-rooted that they are often used even when known to be irrelevant (Tversky & Kahneman, 1974) or when presented subliminally (Mussweiler, Rüter, & Epstude, 2004a). This has led to the suggestion that almost all judgments are to some degree comparative in nature (Kahneman & Miller, 1986; Mussweiler, 2003).

The process of comparing two objects is thought to require the initial establishment of some form of structural alignment between the two object, and thereby a broad evaluation of both the target and the standard along a shared relational structure (Markman & Gentner, 1993; Gentner & Markman, 1994; Mussweiler & Epstude, 2009). If the two are deemed sufficiently alignable, the standard can affect a judgment of the target in one of two ways: a judgment can either move closer towards the standards position on the evaluative dimension, known as an assimilation effect (e.g., Mussweiler, & Strack, 2000), or it can move further away from it, i.e. a contrast effect (e.g., Upshaw, 1978; Herr, 1986). Which direction will be most likely to occur is a highly researched topic that has led to the identification of a plethora of moderating variables such as the ambiguity of the target (Herr, Sherman & Fazio, 1983), category membership (Brewer & Weber, 1994), psychological closeness (Brown, Novick, Lord, & Richards, 1992), processing style (Förster, Lieberman, & Kuschel, 2008), comparative focus (Mussweiler, 2001a),

feelings of fluency (Häfner & Schubert, 2008), and even the temperature in the room (Steinmetz & Mussweiler, 2011), as well as a large number of theoretical accounts that aim to explain such findings. In the next section we will briefly go over the two most influential theories and how they relate to comparative outcomes specifically.

1.2. Theories of Comparative Outcomes

Although there are many models that attempt to provide frameworks which can help explain this complex relationship between assimilation and contrast effects, such as the Self-Evaluation Maintenance Model (SEM; Tesser & Campbell, 1982), the Global/Local Processing Style Model (GLOMO, Förster et al., 2008), the Reflective Evaluation Model (REM, Markman & McMullen, 2003), or the Identification-Contrast Model (Buunk & Ybema, 1997) to name a few, this section will limit itself to describing the two most prominent ones with the broadest applications for comparative outcomes.

The first is the Selective Accessibility Model (SAM, Mussweiler & Strack 1999; Mussweiler, 2003). The SAM is based on the presumption that knowledge of the target which is accessible at the time of the judgment will be more readily incorporated into the final judgment. As such, the standards influence over whether assimilation or contrast occurs is the result of the specific type of knowledge that it activates. The comparison process, as laid out in the SAM, starts with an initial holistic assessment of the similarities between the target and the standard. If the two are considered relatively similar, a biased search and retrieval of information that is consistent with this hypothesis of similarities will take place. As a result, knowledge of the target that is similar to the standard on the judgment dimension itself will also become more accessible. Hence, this similar information will be more readily integrated in the final judgment, leading to an assimilation effect. On the other hand, the biased hypothesis testing can result in more

accessible knowledge supporting the idea that the two are dissimilar if the initial holistic assessment results in a conclusion of dissimilarity. Dissimilar information with respect to the evaluative dimension will then be more accessible, which will lead to contrast effects (Figure 1.2). As an example, if one were to consider how funny they are while comparing themselves to a stand-up comedian, one might initially consider themselves similar (vs. dissimilar) to this standard which will make the memory of them making a fantastic joke more accessible (vs. the time where their jokes fell flat) and the conclusions that they are indeed funny (vs. not so funny) more likely.

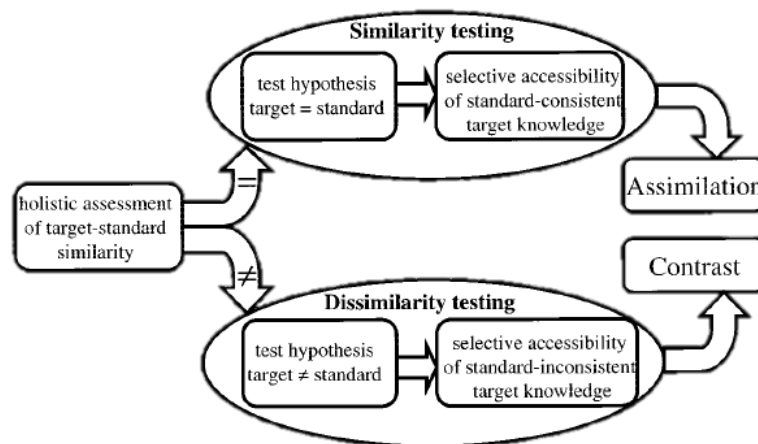


Figure 1.2. A schematic representation the comparison process according to the SAM.

Due to the fact that any comparison is thought to start with the seeking of alignable and conceptually similar features that can be compared (Markman & Gentner, 1993; Gentner & Markman, 1994) and that shared features play a prominent role in this process (Srull & Gaelick, 1983; Tversky, 1977), the default focus is thought to be one of similarities and, thus, the default outcome should be assimilative (Mussweiler & Epstude, 2009). Indeed, this focus on similarities is thought to give a distinct processing advantage over a focus on dissimilarities due to this fact

(Corcoran, Epstude, Damisch, & Mussweiler, 2011). However, any variable that affects the result of the initial judgment of similarities or differences will consequently affect the probable outcome of the comparison. In this way, even seemingly minor variables, like the sharing of a birthday (Brown et al., 1992), can influence the likelihood of assimilation or contrast to occur.

The second prominent model is the Inclusion/Exclusion Model (IEM; Schwarz & Bless, 1992a, 2007; Bless & Schwarz, 2010) and it agrees with the importance of accessible information. However, this model suggests that assimilation and contrast effects result from the construction of mental representations of both the target and the standard at the time of judgment. Any accessible information might be either included or excluded from the formation of these mental representations. In the context of comparative judgments, when similar accessible information is included in both representations, an assimilation effect will occur. Comparison-based contrast effects, occur in the opposite way. When information is included only in the standard after which it is used as a point of comparison to which the target is compared, which always results in a contrast effect. Both these comparison effects increase in strength proportional to the amount and extremity of the information that is included or excluded from the representations. Which information will be included into a representation depends on a number of factors, for instance, if the information is seen as being brought on by irrelevant means or if the information is deemed inappropriate for use due to conversational norms the information will not be included. More interestingly for the comparison process, information must also be seen as representative of the stimulus being considered if it is to be included. In this way, assessable knowledge brought on by the standard that is seen as somehow representative for the target, due to shared category membership for instance (Bless & Schwarz, 1998), assimilation is likely to occur. Conversely, non-representative information will be excluded from

the target representation, leading to contrast effects. Any variables that increases the perceived similarity between stimuli will, therefore, logically also increase the number of similar features that are included in both the representation of the target and the standard. As a result, these representations will have more overlapping features, which makes it more likely that the standard will be deemed representative of the target resulting in assimilation effects.

Although distinct in the theoretical mechanism that leads to the comparative information being integrated into the final judgment, both the IEM and the SAM, therefore, agree that objectively similar comparative information can have the exact opposite outcome depending on small differences in context. Hence, seemingly minor variations in methodological choices could potentially lead to opposite conclusions. For instance, it has already been found that using subjective rather than objective response scales will lead to contrast effect (Mussweiler & Strack, 2000), while the specific domain being considered could increase the use of comparisons standards (Hsee, 1996); induce a focus on differences rather than similarities (Zhang & Markman, 2001); or even lead to different responses patterns for different types of standards (Buckingham & Alicke, 2002). Despite this, many studies are limited to single comparisons and evaluative dimensions (e.g., Ahrens, 1991; Buunk, Groothof & Siero, 2007; Cash, Cash & Butters, 1983, Häfner & Schubert, 2009; Martin, Suls & Wheeler, 2002; McFarland, Buehler & MacKay, 2001; Mendes, Blascovich, Major & Seery, 2001; Raat, Kuks, Van Hell & Cohen-Schotanus, 2013; Smith & Sachs, 1997). A recent meta-analysis spanning the last 60+ years of Social comparison literature by Gerber, Wheeler, & Suls (2017), similarly remark on the fact that very few studies include repeated judgments in any of the areas of social comparative research. Indeed, a structured vote-counting analysis of 74 separate studies investigating evaluative

judgments specifically, across the 43 published works included in this meta-analysis¹, showed that none had participants engage in multiple comparisons, while only six did not use a single fixed standard per condition. In addition, no more than ten of these studies analysed comparative effects on multiple evaluative dimension, with 20 using only a single item. As a result, individual studies on comparative thinking are particularly difficult to interpret or compare unless they investigate the same evaluative dimension with highly similar standards. This lack of diversity in judgment dimensions and standard sampling poses serious issues for the generalisability of findings anywhere beyond the contexts in which they were measured, and similarly can severely endanger the construct validity of their manipulations (Wells, & Windschitl, 1999). With this in mind, it might not be surprising that 66 years after Festinger (1954) introduced his theory of social comparison processes, the field is still divided regarding the question of what is the more common response to comparative stimuli, with many studies reporting assimilation (e.g., Lockwood & Kunda, 1999; Mussweiler, 2001b; Mussweiler & Strack, 2000; Mussweiler, Rüter & Epstude, 2004b; Steinmetz & Mussweiler, 2011; Häfner & Schubert, 2008; Johnson & Lammers, 2012; Lockwood et al., 2005) under relatively similar conditions as to those in which others find contrast effects (e.g., Cash, Cash & Butters, 1983; Mori & Mori, 2011; O'Brien et al., 2010; Smith & Sachs, 1997; Veldhuis, Konijn & Seidell, 2012; Buunk, Groothof & Siero, 2007; Martin, Suls & Wheeler, 2002). The next section will make this issue more apparent by describing the complex relationship of assimilative and contrastive outcomes that arise simply by

¹ This analyses was done using all the published literature that included ability estimates in the reaction study section of the meta-analysis (barring 3 papers for which no access could be obtained), as these are most closely related to the current subject. Although not fully exhaustive of the literature, see Gerber et al. (2017) for exclusion criteria, these numbers offer a useful broad indication of the standard practices employed within social comparative research.

changes in the exact relative standing of the comparison standard on the dimension of interest, a factor which is unavoidable in any comparative judgment.

1.3. The Curvilinear Effect of the Standard

Although many of the standards characteristics, such as category membership (Brewer & Weber, 1994), closeness (Brown et al., 1992), and perceptual fluency (Häfner & Schubert, 2008) can affect the direction the comparative outcome will take, the importance of the standards relative standing in relation to the target cannot be understated. Indeed, the concepts of assimilation and contrast themselves necessitate the acknowledgment of this basic fact, as each may signify an increase or decrease in the judgments estimate depending on whether the standard is relatively higher or lower on the comparative dimension than the target, an upward or a downward comparison respectively. Furthermore, as noted by the IEM, the strength of comparative effects is expected to be proportional to the size of the difference in standing on the comparative dimension (Bless et al., 2003).

However, this effect is not simply linear, but rather the larger the relative distance becomes, the more likely it is that contrast effect will occur. Herr et al. (1983), for instance, showed how ambiguous animals were assimilated towards moderate standards of ferociousness and size, while they contrasted away from extreme standards. Similar findings have since been reported in other studies (e.g., Herr, 1986; Mussweiler, Rüter, & Epstude, 2004b) confirming the moderating role of standard extremity. Both the SAM (Mussweiler, 2003) and IEM (Schwarz & Bless, 2007) have explicitly incorporated the role of standard extremity within their theoretical frameworks, with the former emphasising the way in which extreme standards are likely to produce judgments of dissimilarity and lead to contrast, while the latter suggests that increased feature overlap for moderate standards is the reason assimilation occurs. Regardless of the exact

mechanism, neither disputes the profound effect the relative standing of the standard can have on the outcome of the comparison.

Despite this consequential and unavoidable role on all aspects of the comparison outcome, only vague boundaries can be inferred from the literature regarding when a standard is extreme enough to produce contrast, with little in the way of standardisation. This poses a serious issue for research investigating this phenomena, as operationalisations of the concept vary widely from using well-known celebrities and imaginary characters (e.g., Adolf Hitler or Santa Claus, Herr 1986), to using hand crafted vignettes reflecting distinct levels of the judgment trait (Mussweiler, et al., 2004b) which offer little objectivity regarding the relative distance to the target they are meant to manipulate. Due to the lack of clarity surrounding the boundaries of this effect or an agreed upon way to operationalise the variable, any supposedly extreme standard that does not produce contrast may simply be considered not extreme enough. This forces researchers to instead use the most extreme exemplars in history, who are far removed from everyday interactions. However, even these exemplars may produce assimilation effects if they are construed as representing the trait dimensions they epitomise (Philippot, et al., 1991) or when they are deemed too extreme to be relevant (Lockwood & Kunda, 1997). In this way, abstract terms such as ‘Moderate’ or ‘Extreme’, which arbitrarily dichotomise continuous variables, prohibit any real test of the deeper theory and limit investigations to mere tests of auxiliary theories of operationalisation (Meehl, 1990).

This clearly poses severe issues for the advancement of theory as any result is either in line with the expectations or can be explained by variations in the operationalisation of the extremity variable. Similarly, previous findings with limited standard selection are also strictly limited to those standards and the exact context in which they were found. For example, using

Michael Jordan as an extreme example of athleticism and Nicki Lauda as a more moderate standard (Mussweiler, et al., 2004b) poses a number of issues. Firstly as already described, the relative standing of these standards on the dimension of interest is not particularly well defined. In this case, the extreme standard was not pre-tested, but selected based on it being the most extreme athletic example of its time. The moderate standard, on the other hand, was selected based on its ability to produce assimilative effects in previous work and testing among 21 student four years prior (Mussweiler & Strack 2000), where the somewhat arbitrary decision was made that 1.8 points above the middle of a 9-point scale was an acceptable score for a moderate standard with no mention of what the theoretical basis for this could be. This manner of standard selection reduces the clarity of when assimilation or contrast is expected to occur theoretically. If this moderate standard would have produced contrast, it is perfectly logical to conclude that Nicky Lauda, an Austrian Formula One driver, is seen as a somewhat extreme standard himself, and that a cut of 1 point above the middle of a 9-point scale might be more appropriate. Here one may note that the downward moderate standard used was Bill Clinton, judged only 1.1 points below the midpoint. However, a more severe issue with this form of standard selection is that the exemplars differ on many other traits that are not related to their differences on the athleticism dimension (e.g., age, race, nationality etc.), which could all be of significant influence on the comparative direction as they can directly affect perceived category membership and similarities. This not only limits the generalisability of the findings to the exact standards used, it poses a serious limitation to the construct validity of the extremity manipulation, which in this design includes all these secondary differences as well (Wells & Windschitl, 1999).

Note that this is not specific to the example described here, but is true to some extent in any design that does not include multiple judgments with broadly selected standards. In fact,

these concerns also extend to research not investigating the variable of target extremity itself. Any comparative judgment necessarily happens in the presence of either a more moderate or extreme standard, and is, thus, fundamentally influenced by the exact standard that is used. In other words, there is no “default” standard extremity, nor has one been suggested to function as such for standardisation practices in the interest of making studies more comparable. Instead, many examples in the literature not specifically interested in the effects of standard extremity do not seem to account for its inherent influence, nor do they consider the relative standing of their standards prior to their use (e.g., Buunk, Groothof & Siero, 2007; Häfner & Schubert, 2008; Lockwood et al., 2005; Bailis & Chipperfield, 2006; Faith, Leone & Allison, 1997; Vogel et al., 2014; Lin & Kulik, 2002; O’Brien et al, 2010). Without an objective measure of the standards relative standing, there is no way of knowing if the upward and downward conditions these studies employ are equally extreme, once again hindering generalisability and validity of the direction variable (Wells & Windschitl, 1999). Such practices make it difficult to meaningfully compare effects across studies that employ different standards, and they may even lead to erroneous conclusions of theoretical support, as well as apparent null results in specific cases as will be shown later. To offer some perspective on the scale of this issue, of the 74 studies used in the vote counting analyses mentioned previously¹, 48 did not define the relative standing of their standard before their use in the studies, while for 55 it was unclear if conditions were equal in their extremity, and only six actively considered the extremity variables moderating effects.

At the heart of these issues lies the simplification of the relative standing variable into an abstract dichotomous variable of moderate and extreme. This practice creates the illusion of an undefined ‘moderateness’ or ‘extremeness’ that can be used to separate these categories precisely, while in reality they refer to a wide range of values that need to be specified and

sampled from broadly in order to produce accurate inferences. This is the case even when it is not the central parameter under investigation, as all standards will necessarily reflect some level of relative distance. The lack of a clear definition for this variable and its narrow sampling does not only pose limits on research, it also raises new questions that have not previously been addressed. For example, what exactly constitutes an extreme standard, and what comparative outcome can be expected on the cusp of moderate and extreme standards?

1.4. The Current Aim

The current dissertation will argue for a more holistic approach to comparison research, in order to create strong tests of theory, and improve generalisability beyond the judgment dimensions and standards used in specific investigations. For this, it will propose to redefine standard extremity as a spectrum of values reflecting the relative distance from the standard to the target. At each point in this spectrum assimilative and contrastive tendencies will occur to varying degrees, resulting in a dynamic interactive pattern that shifts with changing contexts. Not only will this provide a more accurate depiction of the influence of the standard and define the extremity variable more clearly, it will also emphasise the fact that there is no default standard as they all have a unique influence on this equilibrium. Finally, it will provide tangible design recommendations for future research and propose a newly created paradigm as a first possible alternative to more traditional designs.

This endeavour will require a departure from some of the more common practices in comparison research. Firstly, the empirical work will focus mainly on other related comparisons, rather than comparing the self to a standard as is more common in the field of social comparison research. This is a necessity in order to ensure the informational content of both the target and the standard can be sufficiently controlled. Indeed, the inherent informational asymmetry

between the self as a target and other as a standard has been posited as an informational account for various biases found in comparative research with far reaching consequences (Chambers & Windschitl, 2004), not to speak of the possibility that self-enhancement motivation also affect self-comparisons (Kruglanski, 1996). In addition, the objective standing of the target (i.e. the participant) varies widely when self-related comparisons are considered, making the precise manipulation of the relative distance exceedingly difficult. Attempting to measure the objective standing of the target alone poses an issue in this context, as any self-reported behavioural frequencies could themselves be the result of some type of comparative process. However, this design decision does entail its own limits to the generalisability of the findings, as well as the ease with which the results can be compared to the broader literature. The described patterns and findings should be considered as reflecting only the process of comparing two other individuals, with the knowledge that self-related comparisons may produce vastly different patterns. Nevertheless, the implications for study design will hold regardless of this difference, and the more controlled environment should help offer a clearer first look at the proposed dynamic interactive pattern of comparative outcomes.

Secondly, a similar restriction to the generalisability results from the choice of focusing the studies on facial evaluations specifically. This decision was made as facial stimuli are processed quickly (Willis & Todorov, 2006; Ballew & Todorov, 2007) and are automatically judged on a number of social dimensions (Oosterhof & Todorov, 2008; Todorov, Olivola, Dotsch, & Mende-Siedlecki, 2015), which makes them a particularly useful way to present precise comparative information unobtrusively and with minimal strain on participants. Despite these benefits, and the fact that facial evaluations provide an exceptionally common everyday context in which comparisons occur, research in this area has largely been limited to comparative

evaluations of attractiveness (e.g., Wedell, Parducci, & Geiselman, 1987; Thornton & Moore, 1993). Although this means this area offers a domain in which comparison effects have been less well documented, it also means that the presented studies will be limited in their generalisability to comparative judgments in the facial domain specifically. This fact should be considered when evaluating the exact patterns that are uncovered as they may differ from those found in the broader literature. However, this limit to the generalisability of the exact comparative patterns should once again not pose a significant issue when considering the methodological shortcomings that the studies aim to highlight, as these will apply to all comparisons domains to a greater or lesser extent.

With these points in mind, the next Chapter will discuss the issues surrounding limited selection of items, dimensions, and comparison standards that are common throughout the literature. In doing so, it will show how opposing comparative effects can result from similar standards depending on the item, or dimension that is being considered. This issue will become even more apparent in Chapter 3, where the dynamic interactive pattern of assimilation and contrast will be modelled for the first time using a curve fitting approach. This will not only show the extent of the problems related to item and standard selection, but will also offer the first concrete definition of what may constitute an extreme standard, and show that standards in between ‘moderate’ and ‘extreme’ may produce apparent null effects. The implications of this dynamic pattern for comparison research will be shown in Chapter 4, where the efficacy of the CJT-paradigm to experimentally test the influence of different comparative foci will offer one potential way of conducting stronger tests of theory in future research. Finally, Chapter 5 will discuss the implications of the findings for the literature and future research, while discussing the limitations of the current work as well as offering some final guidelines for study design. The

empirical chapters will mainly be structured around the SAM framework when making predictions and interpretations for the sake of clarity, as this is arguably the most prominent model focused specifically on comparative thinking. However, it should be noted that the results could equally be described with a number of other models that acknowledge the role of standard extremity, such as the IEM.

Chapter 2: Heterogeneous Effects across Judgments²

Abstract

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.

² This chapter is based on Barker, P., Dotsch, R., & Imhoff, R. (2020). Assimilation and contrast in spontaneous comparisons: Heterogeneous effects of standard extremity in facial evaluations. *International Review of Social Psychology*, 33, 11. <http://doi.org/10.5334/irsp.402>.

This is a postprint that might differ from the authoritative final version in print. For citation, please refer to the authoritative final version in print.

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 shifts closer, as has also been found in countless studies (e.g., Mussweiler, & Strack, 2000; Brown, Novick, Lord, & Richards,

1992; Schwarz, & Bless, 1992b). 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 explicable 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 (SAM; Mussweiler, 2003) has suggested a comprehensive framework for understanding the mechanism underlying this variation in comparison outcomes. An initial holistic assessment of target-standard similarity critically determines the nature of the hypothesis that will be tested in the next step. If the target and standard are judged to be similar initially, a congruent hypothesis that target and standard are alike will be formed and tested (similarity testing) in the subsequent search for and activation of relevant knowledge. Conversely, if they are seen as different initially, a hypothesis of differences will be formed and dissimilarity testing will be the next step. The testing of these differing hypotheses is assumed to be biased towards hypothesis-consistent evidence in the active search for judgment-relevant information. This will lead to an accentuated impression of similarity (assimilation) or difference (contrast) on final estimates. The previously mentioned moderating variable of extremity fits into the SAM by crucially affecting the propensity for the initial assessment to be one of similarity or dissimilarity, as extreme standards have a higher propensity to lead to an initial assessment of dissimilarity, whereas moderate standards are more likely to be seen as similar (Mussweiler & Strack, 2000).

Despite the consequential role standard extremity plays in social comparison outcomes, its investigation has historically been limited to a single dimension at a time and with frequent use of the most extreme of standards (e.g., Hitler vs. Shirley Temple in Herr, 1986). However,

such results do not speak to the consistency of the proposed pattern in the multitude of evaluative dimensions that could be subject to comparative judgments, but are bound to the dimension under investigation and the standards presented. Without the availability of a paradigm that assesses comparison patterns across various evaluative dimensions and using a larger set of standards, generalized claims about the moderating role of target extremity cannot be fully supported.

The current work proposes a paradigm that includes the manipulation and testing of the critical extremity variable across a number of dimensions in the domain of face perception, relying on data-driven techniques to generate digital facial images that can be precisely varied in their extremity on a number of dimensions (Todorov et al., 2013). Another advantage of facial (vs. verbal) stimuli lies in their pivotal role as a source of information for the formation of judgements about a host of traits (Oosterhof & Todorov, 2008; Todorov et al., 2015) in the first 100ms 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.

2.1. 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 2.1). 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 overall linear trend from $-4SD$ to $+4SD$, $F(4, 320) = 90.93$, $p < .001$, $\eta_p^2 = .532$,

90% CI [.468, .577], see Figure 2.1 and Table 2.1 for all pairwise comparisons.

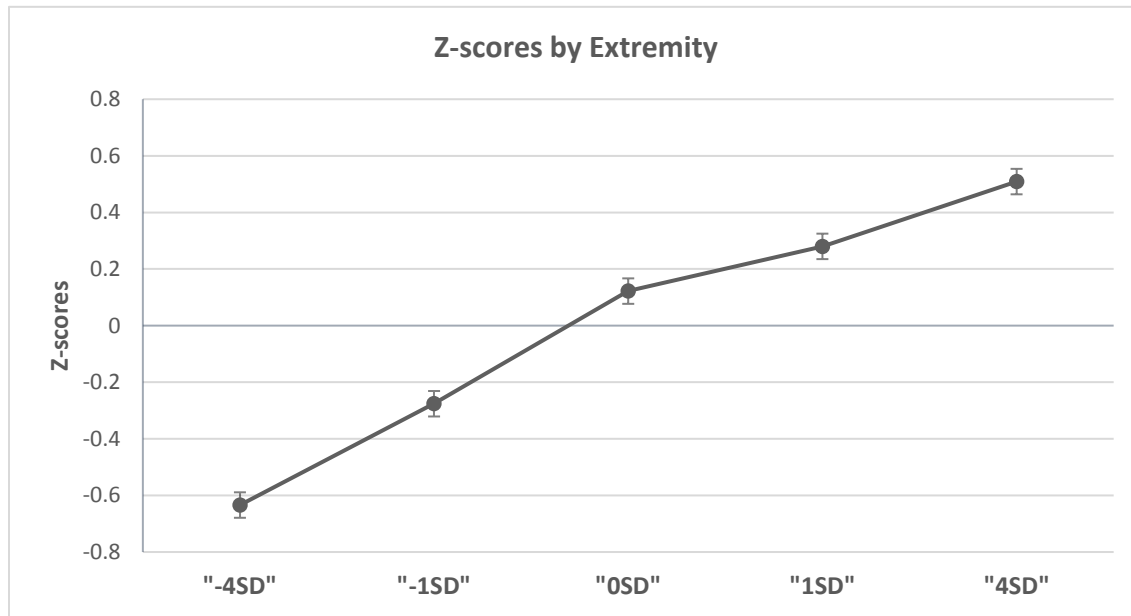


Figure 2.1. Means and standard errors of Z transformed responses for each extremity step measured.

Table 2.1.

Mean differences and standard errors by Extremity level.

	-1SD		0SD		1SD		4SD	
	ΔM	SE	ΔM	SE	ΔM	SE	ΔM	SE
-4SD	-.358	.069	-.756	.068	-.914	.083	-1.143	.081
-1SD	-	-	-.399	.056	-.556	.068	-.785	.068
0SD			-	-	-.157	.059	-.386	.055
1SD					-	-	-.229	.062

Note. All differences were significant, $p < .01$

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 2.2 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 (millimetres), no pilot study was needed.

Table 2.2.
Separate analyses of linear trends of extremity per dimension with the related items.

	<i>df</i>	<i>F-value</i>	<i>Sig.</i>	η_p^2 [90% CI]
Competence (reversed): “How often does the target make a mistake at work per month?”	1, 80	22.05	<.001	.216 [.095; .336]
Dominance: “How often a month does the target enforce his opinion?”	1, 80	95.68	<.001	.545 [.418; .631]
Extraversion: “How often does the target go out a month?”	1, 80	76.25	<.001	.488 [.095; .336]
Likability: “How often does the target offer help to a stranger a month?”	1, 80	123.57	<.001	.607 [.355; .583]
Trustworthiness (reversed): “How often does the target deceive somebody every month?”	1, 80	72.43	<.001	.475 [.341; .572]

2.2. 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.

2.2.1. 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, $\eta_p^2 > .013$, $\rho = .50$, $\alpha = .05$, $\beta = .20$ (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 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 ($\alpha = 0.89$) 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 participant 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 ranges³. 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 analysis. In total, this was the case for seven participants.

³ 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, $r = .813$, yielding similar results.

2.2.2. 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) \times 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, $F(1, 151) = 10.52$, $p = .001$, $\eta_p^2 = .065$, 90% CI [.016, .136]; see Figure 2.2. The interaction effect of the direction and judgment dimension also reached significance, $F(4, 604) = 3.04$, $p = .017$, $\eta_p^2 = .020$, 90% CI [.002, .036], suggesting the presence of significant heterogeneity in effect sizes across the different evaluative dimensions. See the supplemental materials for dimension specific graphs.

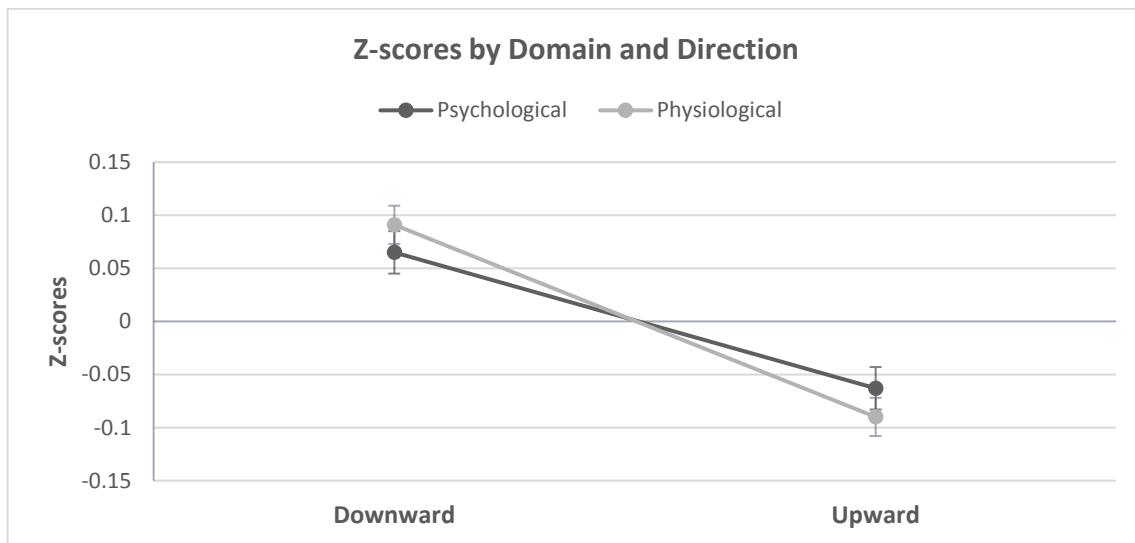


Figure 2.2. Means and standard errors of Z transformed responses separate for the two domains and the direction of the comparison.

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, $F(1, 151) = 24.82, p < .001, \eta_p^2 = .141, 90\% \text{ CI } [.065, .226]$; see Figure 2.2. Again, the data presents with significant heterogeneity across dimensions, $F(4, 604) = 4.55, p = .001, \eta_p^2 = .029, 90\% \text{ CI } [.007, .049]$. 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, $r = .165, p = .039$, 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 $r < .1$.

2.2.3. 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.

2.3. 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 comparably 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.

2.3.1. 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) \times 2 (extremity levels) \times 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.

2.3.2. Results

A 2 (Comparison direction) \times 2 (Extremity) \times 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, $F(1, 151) = 0.26, p = .610, \eta_p^2 = .002, 90\% \text{ CI } [.000, .028]$. A main effect of direction, $F(1, 151) = 13.44, p < .001, \eta_p^2 = .082, 90\% \text{ CI } [.025, .157]$, 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 ($\Delta M = -0.083, SE = 0.034, p = .017, 95\% \text{ CI } [-0.150, -0.015]$) and extreme condition ($\Delta M = -0.109, SE = 0.039, p = .006, 95\% \text{ CI } [-0.186, -0.032]$); see Figure 2.3. The analysis also yielded significant effects of the interaction between direction and dimension, $F(4, 604) = 7.43, p < .001, \eta_p^2 = .047, 90\% \text{ CI } [.019, .071]$, extremity by dimension, $F(4, 604) = 4.25, p = .002, \eta_p^2 = .027, 90\% \text{ CI } [.006, .047]$, and the three-way interaction, $F(4, 604) = 6.67, p < .001, \eta_p^2 = .042, 90\% \text{ CI } [.016, .066]$, indicating a complex picture of heterogeneous effects across the evaluative dimensions, see supplemental materials.

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, $r = .004$, $p = .962$. 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, $r = .280$, $p < .001$, but only marginally with extreme difference scores, $r = .144$, $p = .072$. The positive correlation here indicates participants with higher INCOM scores actually showed less, not more contrast.



Figure 2.3. Means and standard errors of Z transformed responses for all dimensions separate for the direction of the comparison and its extremity (+/-2SD, or +/-4SD).

2.3.3. 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.

2.4. 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 ones. These hypotheses should be reflected in a crossover interaction between direction and extremity.

2.4.1. 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 $N = 148$).

2.4.2. Results

A 2 (Comparison direction) \times 2 (Extreme) \times 5 (Dimension) factorial repeated measures ANOVA showed the expected interaction between extremity and direction, $F(1, 147) = 6.66, p = .011, \eta_p^2 = .043, 90\% \text{ CI } [.006, .108]$, with pairwise comparisons showing significant assimilation for moderate standards ($\Delta M = 0.071, SE = 0.034, p = .041, 95\% \text{ CI } [0.003, 0.139]$), even though extreme standards showed contrast short of significance ($\Delta M = -0.058, SE = 0.038, p = .132, 95\% \text{ CI } [-0.133, 0.018]$). Furthermore, it seems upward standards were more influential than downward ones in this sample (see Figure 2.4), also reflecting the main effect of extremity, $F(1, 147) = 5.97, p = .016, \eta_p^2 = .039, 90\% \text{ CI } [.004, .101]$.

The interaction terms of dimension with direction, $F(4, 588) = 8.42, p < .001, \eta_p^2 = .054, 90\% \text{ CI } [.024, .081]$, extremity, $F(4, 588) = 3.95, p = .004, \eta_p^2 = .026, 90\% \text{ CI } [.005, .045]$, as well as the interaction of all 3, $F(4, 588) = 4.750, p = .001, \eta_p^2 = .031, 90\% \text{ CI } [.082, .052]$, indicated significant variations in the effects across dimensions as in the previous study; see supplemental materials.

Correlational analyses. The correlation between moderate difference scores and extreme difference scores was also not significant in this sample, $r = .106, p = .19$. 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, $r < 0.07$.

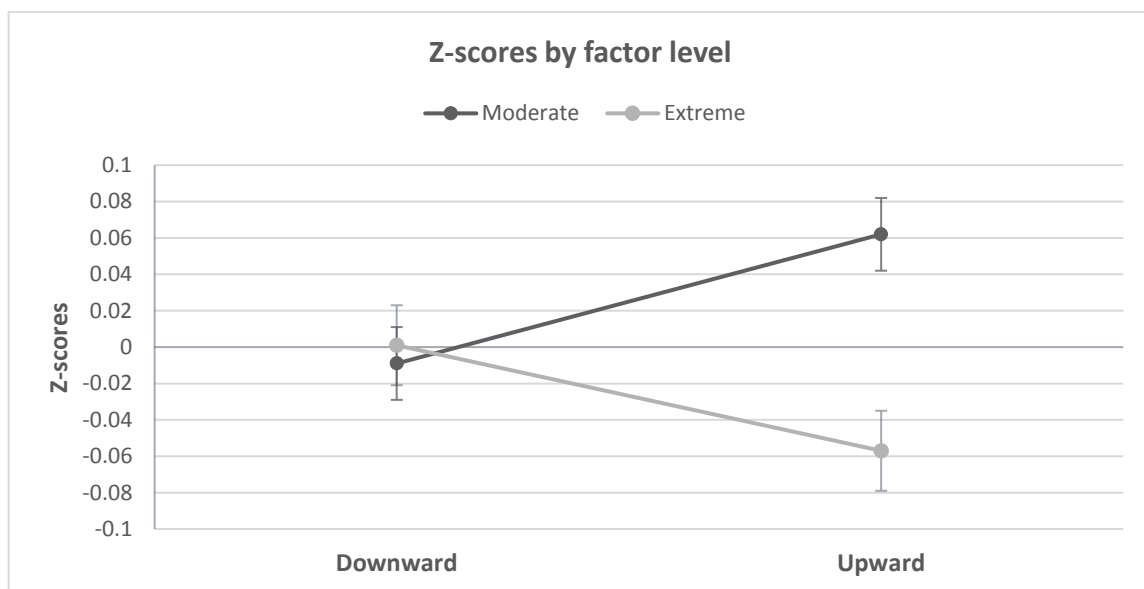


Figure 2.4. Means and standard errors of Z transformed responses for all dimensions separate for the direction of the comparison and its extremity ($\pm 1SD$, or $\pm 4SD$).

2.4.3. 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 whole sample, as reflected in the absence of consistent negative correlation in response patterns. This could in part be due to large variations in judgments and too low a sensitivity to measure individual response patterns reliably. Furthermore, the repeated measurement in the current study might be a potential methodological issue exacerbating this problem. There are ample studies proposing the notion that procedurally priming a focus on similarities (or differences) can induce assimilation (or contrast) effects to occur in unrelated subsequent comparative judgments (Mussweiler, 2001a; Mussweiler, Rüter & Epstude, 2004; Mussweiler & Epstude, 2009). In the repeated measures design of the current study, exposure to both moderate and extreme standards might bias respondents into using one of the two suggested information seeking strategies, similarity or dissimilarity testing, respectively. As a result, individual participants may only show assimilation or contrast across all trials, reflected in the descriptively positive correlation between difference scores. This would weaken the overall assimilation and contrast effects on average that could result in underestimated effect sizes, as well as suppress any correlations with the explicit measure of social comparison orientation, the INCOM. Therefore, the next study addressed this issue by changing the factor of extremity from a within to a between-subjects factor.

2.5. 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 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.

2.5.1. 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 ($N = 161$).

2.5.2. Results

A 2 (Extremity) \times 2 (Comparison direction) \times 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, $F(1, 159) = 1.75, p = .188, \eta_p^2 = .011, 90\% \text{ CI } [.000, .052]$; see Figure 2.5. Descriptively, the expected pattern did occur, but pairwise comparisons showed no significant effect of direction in the moderate, ($\Delta M = 0.023, SE = 0.044, p = .595, 95\% \text{ CI } [-0.063, 0.110]$), or extreme comparison condition, ($\Delta M = -0.058, SE = 0.044, p = .183, 95\% \text{ CI } [-0.145, 0.028]$). Finally, the interaction term of dimension with direction, $F(4, 636) = 4.61, p = .001, \eta_p^2 = .028, 90\% \text{ CI } [.007, .047]$, was again found to be significant in this dataset. See supplemental materials for graphs per dimension.

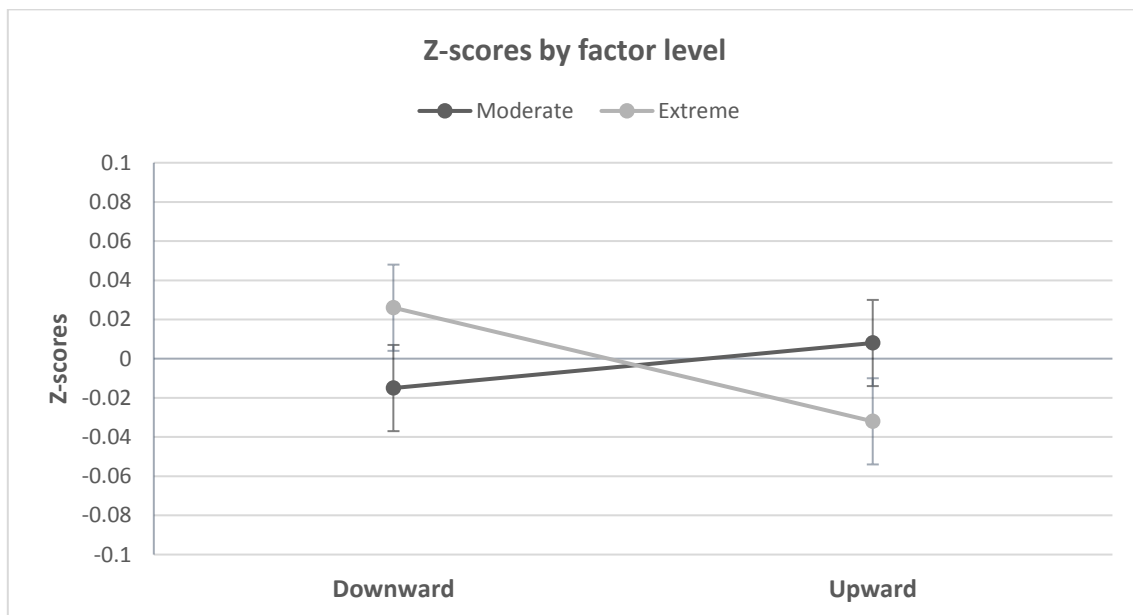


Figure 2.5. Means and standard errors of Z transformed responses for all dimensions, separately for the direction of the comparison and for each of the extremity conditions ($\pm 1SD$, or $\pm 4SD$).

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, $r < 0.11$.

2.5.3. Discussion

Contrary to the notion that the between-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.

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.

2.6. 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 materials⁴).

2.6.1. 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 8 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).

⁴ Data and analyses for all studies conducted in the research line can be found in the subliminal materials. 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.

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.

2.6.2. 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, $F(1, 91.96) = 1.529, p = .219$, as would be expected if only assimilation or only contrast effects occurred, nor was there an interaction with target extremity, $F(1, 99.90) = 0.951, p = .332$, which would indicate consistent assimilation and contrast (Figure 2.6).³ However, in line with the results from the separate studies, a significant effect was found for the interaction of dimension and direction, $F(1, 125.08) = 3.624, p = .008$, reflecting heterogeneous comparison effects across the dimensions in this pooled sample, much like in the individual studies.

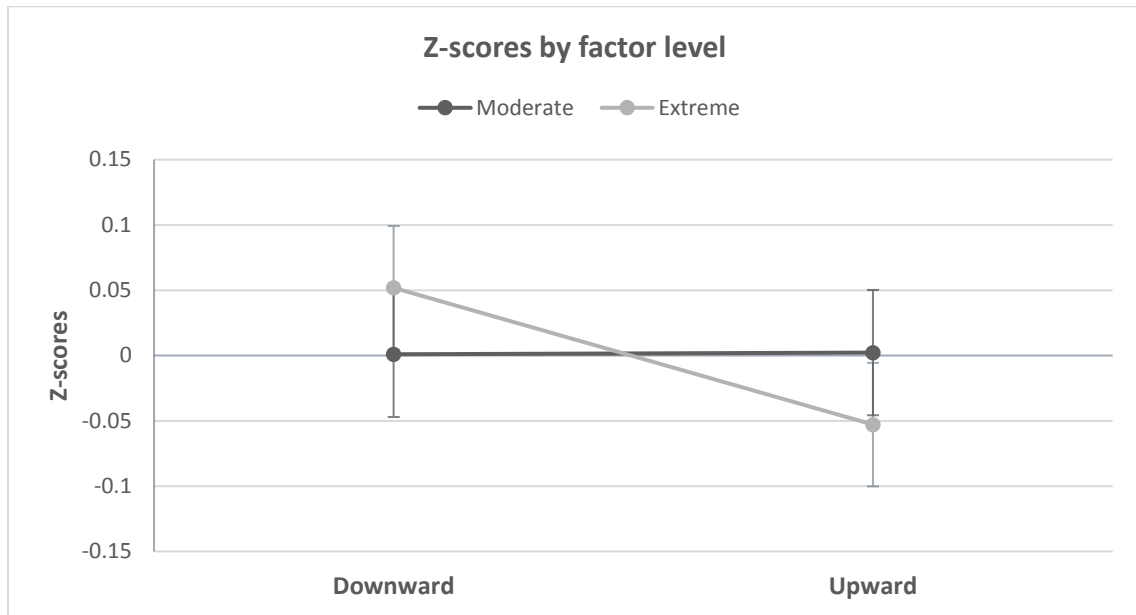


Figure 2.6. Estimated marginal means standard errors of Z transformed responses over all studies separately for the direction of the comparison and for each of the extremity conditions.

2.6.3. 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.

2.7. 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.

2.7.1. 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 dimension separately. Although moderate standards varied in Study 2 ($\pm 2SD$) compared to those in the other studies ($\pm 1SD$) 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_z) 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, with the exception of some slight heterogeneity in the Dominance dimension effects ($I^2 = 17.81\%$) that did not reach significance in this small sample, $Q(4) = 6.531, p = .163$.

The results show that consistent assimilation effects toward moderate standards were only found for the dimensions of Extraversion and Trustworthiness (Table 2.3). Conversely, effects

sizes for the Dominance and Competence dimensions were in fact contrastive in nature. The Likability dimension displayed no significant effect in either direction.⁵

Table 2.3.

Facial dimensions and their meta-analytic assimilation and contrast effect across all the studies for Moderate standards.

	<i>K</i>	<i>I</i> ²	Meta-analytic <i>d_z</i> (95% CI)	<i>Z</i>	<i>p</i>
Facial dimension					
Competence	4	0%	-.166 [-.299; -.034]	-2.456	.014
Dominance	5	17.81%	-.266 [-.405; -.127]	-3.742	<.001
Extraversion	4	0%	.168 [.035; .302]	2.469	.014
Likeability	4	0%	-.063 [-.199; .073]	-0.903	.367
Trustworthiness	6	0%	.277 [.169; .384]	5.041	<.001

2.7.2. 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_z*) 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 dimension showing any consistent effects in either direction (Table 2.4).⁵

⁵ The results of these meta-analyses are again bound to the stimuli used for each dimension. For the contrasts of the pooled analyses 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 2 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 generalise to the larger stimulus population.

Table 2.4.

Facial dimensions and their meta-analytic assimilation and contrast effect across all the studies for Extreme standards.

	<i>K</i>	<i>I</i> ²	Meta-analytic <i>d</i> _z (95% CI)	<i>Z</i>	<i>p</i>
Facial dimension					
Competence	5	0%	.006 [-.119; .130]	0.089	.929
Dominance	6	0%	-.252 [-.365; -.139]	-4.360	<.001
Extraversion	5	0%	-.217 [-.338; -.096]	-3.519	<.001
Likeability	5	0%	-.042 [-.164; .080]	-0.670	.503
Trustworthiness	7	0%	-.026 [-.131; .076]	-0.520	.603

2.7.3. 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% CI [-.035, 0.101], $Z = .955$, $p = .34$, with no significant signs of heterogeneity, $Q(6) = 8.480$, $p = .205$, and a low I^2 (14.92%). For moderate standards the meta-analytic effects were also non-significant, $z' = 0.009$, 95% CI [-.100, 0.118], $Z = .165$, $p = .869$, but showed high I^2 (60.21%) and significant heterogeneity, $Q(5) = 15.260$, $p = .009$. Further analyses indicated Study 2 was the main cause of this heterogeneity, likely due to the fact that the moderate standard for this study was at *2SD* rather than *1SD* like in the subsequent studies. Removing this study reduced heterogeneity to non-significant levels, $Q(4) = 1.041$, $p = .904$, and an I^2 of 0%, but left the conclusions unaltered, $z' = -0.050$, 95% CI [-.124, 0.025], $Z = -1.297$, $p = .195$. 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.

2.7.4. 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.

2.8. 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 focuses 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 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 mind-set 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.

Data Accessibility Statement

All data files, analyses and additional materials can be found at:

https://osf.io/gb9k3/?view_only=98127a783c2440f09284b7b26d7c4a93

Chapter 3: The Dynamic Interactive Pattern of Assimilation and Contrast ⁶

Abstract

The extremity of a comparison standard is a central moderator for the outcomes of comparative judgments. However, what exactly constitutes a moderate or an extreme standard is ill defined in the literature. To address these issues, the current work takes a fine-grained curve fitting approach to the measurement of the comparison response patterns. A series of five experiments ($N = 2013$) measured the comparison patterns for three dimensions in the domain of facial judgments. The heterogeneity in the role of target extremity that was found in the initial studies (1-3) proved to be mainly caused by heterogeneity at the item level, rather than at the level of the judgment dimensions themselves. Dimension level dynamics (4-5) were in line with established theories of social comparison, and suggest standards in the facial domain are considered extreme enough to elicit consistent contrast effects after $3SDs$, although strong item level variation remained present.

⁶ This chapter is based on Barker, P., & Imhoff, R. (2020). *Comparing curves: Describing the dynamic interaction of assimilation and contrast in facial evaluations*. Manuscript submitted for publication. The current version of this paper is a pre-print form 15.06.2020 and has not been peer reviewed. Please do not copy or cite without author's permission.

To make more accurate social judgments we often look for standards in our environments to use as a comparison. The standing of this standard on the judgment dimension will influence how the final judgment is made. More extreme standards will lead to contrasting judgments, whereas moderate ones are expected to result in assimilative judgments (Herr, 1986). However, what exactly constitutes a moderate or an extreme standard, as pertains to the dynamics of assimilation and contrast, is not defined in the existing literature. Moreover, if what is moderate or extreme varies across judgment dimensions, this may lead to diverging patterns and replication failures. Therefore, the present research will use a fine-grained manipulation to investigate these seemingly contradictory patterns, and provide some guidance as to what may constitute an extreme standard by modelling comparison patterns for three evaluative facial dimensions (Extraversion, Trustworthiness, and Dominance) to determine the point where assimilation shifts into contrast.

Every day, people make an almost endless number of judgments about the people they meet, for instance, about how attractive they are, smart they look, or friendly they might be. Social judgments like these are often not made in absolute terms, but are relative to some standard such as other people who are close by or some internal representation (Dunning & Hayes, 1996; Festinger, 1954). Whether someone is athletic or intelligent can only be said in relation to how well others fare in these domains. In this way, most social judgments are to some degree comparative in nature (Kahneman & Miller, 1986), and who one chooses to compare to is far from inconsequential, leading to measurable shifts in the evaluations (Mussweiler, 2003).

Logically, a standard can influence the final judgment of a target in one of two ways; the standard can pull the evaluation closer to its standing, known as an assimilation effect (e.g., Mussweiler, & Strack, 2000; Schwarz, & Bless, 1992b), or repel the judgment away from it with

regards to the judgment dimension, i.e., a contrast effect (Upshaw, 1978). A large number of variables have been suggested to influence which of the two comparison effects is more likely to occur; for instance, a shared category membership (Brewer & Weber, 1994; Mussweiler & Bodenhausen, 2002), or even the temperature in the room (Steinmetz & Mussweiler, 2011). One other moderating variable of particular interest in the literature is the perceived extremity of the standard with regard to the evaluative dimension being considered. Namely, more extreme standards are increasingly likely to cause contrastive judgments, while more moderate ones should mainly produce assimilative effects (e.g., Herr, 1986; Herr, Sherman, & Fazio, 1983).

The unique aspect of target extremity as a moderator is that it relates directly to the informational content of the comparison standard itself and is, thus, necessarily considered for any comparative effect to occur at all. It is, therefore, perhaps not surprising that it has been noted as a key moderator in many of the most prominent theories of comparative judgments, such as the Selective Accessibility Model (SAM; Mussweiler, 2003), where it is theorised to influence the crucial initial judgment of similarity. The SAM posits that, at the onset of any comparative evaluation, a holistic dichotomous judgment is made regarding the perceived similarity or dissimilarity of the target and the standard. Dependant on the outcome of this initial judgment, biased hypothesis testing will follow, with congruent information becoming more accessible and influential in forming the final judgments. In this process, more extreme standards are thought to lead to a higher likelihood of forming an initial judgments of dissimilarity. This results in the increased accessibility of information that emphasises differences on the judgment dimensions, which in turn leads to contrast effects. On the other hand, moderate standards will more likely result in judgments of similarity and assimilation. Although, other models propose slightly altered mechanisms (e.g., representativeness and the feature overlap; IEM, Schwarz &

Bless, 1992a), they also posit that assimilation to moderate standards and contrast from extreme standards should occur in general.

However, recent research in this area has struggled to show this theoretically presumed pattern for all judgments, instead showing large inter-item variation ranging from the expected pattern to exclusively assimilation, only contrast, or even simply null effects (Barker, Dotsch, & Imhoff, 2020). More precisely, evaluative judgments related to the dimension of Dominance showed exclusively contrast effect to all standards regardless of their extremity. On the other hand, those related to Trustworthiness showed only the expected assimilation to moderate standards, but no effect at all for more extreme ones. In fact, only comparative judgments related to the dimension of Extraversion showed the predicted pattern of assimilation to moderate and contrast from extreme standards.

3.0.1. What constitutes an extreme standard?

A key reason for these contradictory findings may be that what constitutes a moderate or extreme standard varies wildly across judgments. This possibility has, however, never been addressed, nor has any attempt been made to concretely define the boundaries surrounding moderate and extreme standards, despite the consequential role this moderator is theorised to play in all comparative judgments.

The lack of clear guidelines has lead researchers to use substantially varying ways of operationalising a moderate or extreme standard. In the aforementioned research by Barker et al. (2020), for instance, standards were selected based on the number of standard deviations they were removed from the neutral target on the evaluative dimension, with somewhat arbitrarily chosen standards of $1SD$ chosen as moderate and $4SD$ deemed extreme. Many others have opted to purposefully craft descriptions or hand select celebrity names deemed moderate or extreme on

the judgment dimension (e.g., Bill Clinton vs. Michael Jordan in Mussweiler, Rüter, & Epstude, 2004a) with little objectivity regarding the level of their extremity. Others still have attempted to use pre-tested ratings to parse well-known people into groups of moderate and extreme standards, but in doing so only select the most radical exemplars to be extreme (e.g., Hitler and Charles Manson vs. Shirley Temple and Santa Claus, in Herr, 1986). Despite showing contrast effect in this context, these studies provide little information about such effects in response to anything but the most pivotal exemplars in history.

There are, thus, substantial differences in the definition of what a moderate or extreme standard might be. In addition, there exists a possibility that the precise level of extremity necessary to produce contrast effects might vary between evaluative judgments and participants. Taken together, one may wonder if any discrepancy in outcomes between studies could simply be ascribed to the insufficient calibration of the standard's extremity for it to produce the expected effect. Indeed, without a more systematic approach to defining the boundaries of where assimilation is suspected to become contrast for varying judgments, differences in implementation regarding the operationalisation of moderate and extreme standards can always be presented as an alternative explanation for these unexpected findings. Take for instance the pattern found in Barker et al. (2020) where evaluative judgments related to the dimension of Dominance showed contrast effects from extreme standards of $4SD$ from the mean, but also for standards as moderate as $1SD$. One may argue that for this judgment the seemingly moderate standard was simply already too extreme to produce assimilation. Similarly, if an extreme standard were to produce assimilation effects, it may have simply not been extreme enough. Thus, the vagueness of such moderating variables can limit any conclusions from contradictory

results to the validity of auxiliary theories regarding the operationalisation of the moderator rather than the underlying theory (Meehl, 1990).

Compounding this issue is the fact that assimilation and contrast are themselves ostensibly conflicting effects that move judgments in opposite directions. Although, many theoretical accounts such as the SAM (Mussweiler 2003) view the comparative process as dichotomous in its outcome on an individual judgment level (i.e. either assimilation occurs or contrast does, but not a combination of both in a single judgment), the aggregation of many similar judgments will invariably include both some assimilative and some contrastive judgments due to natural fluctuations in the initial state of the comparison process. Moderating variables such as the extremity of the standard, thus, logically only increase the likelihood of one of the two outcomes to occur for a judgment, but do not preclude individual judgments from still following the alternative route. As a consequence, these opposing comparative effects would be expected to suppress one another to varying degrees, as the likelihood of either outcome occurring increases or decreases at different levels of extremity, with either rarely become absolute. This dynamic interaction of assimilation and contrast would produce a pattern on average that is perhaps best described as roughly a lying S-shape (Figure 3.1). Moving in both directions from the neutral judgment, we see mainly the influence of assimilation to the most moderate standards, which will slowly increase in strength as the standards do as well. Once some peak is reached (indicated in the figure by the Point of Maximal Assimilation; PMA), the increasing extremity of the comparison standards will no longer only lead to stronger assimilative effects, but will instead start making contrastive judgments more likely. This increased prevalence of contrast effects will slowly start to suppress the average assimilation effect until a point is reached where both assimilative and contrastive tendencies effectively cancel each other out, indicated at the

position where the S-shaped curve crosses the x-axis in Figure 3.1. Every standard before this point is within the Window of Assimilation (WoA), while the Area of Assimilation (AoA) is represented as the blue area under the curve showing the cumulative strength of assimilation across the range of values in which it is the dominant force. Past this turning point, assimilation will become increasingly rare whereas contrast effects will become dominant, with consistent average contrast effects presenting themselves for these more extreme standards.

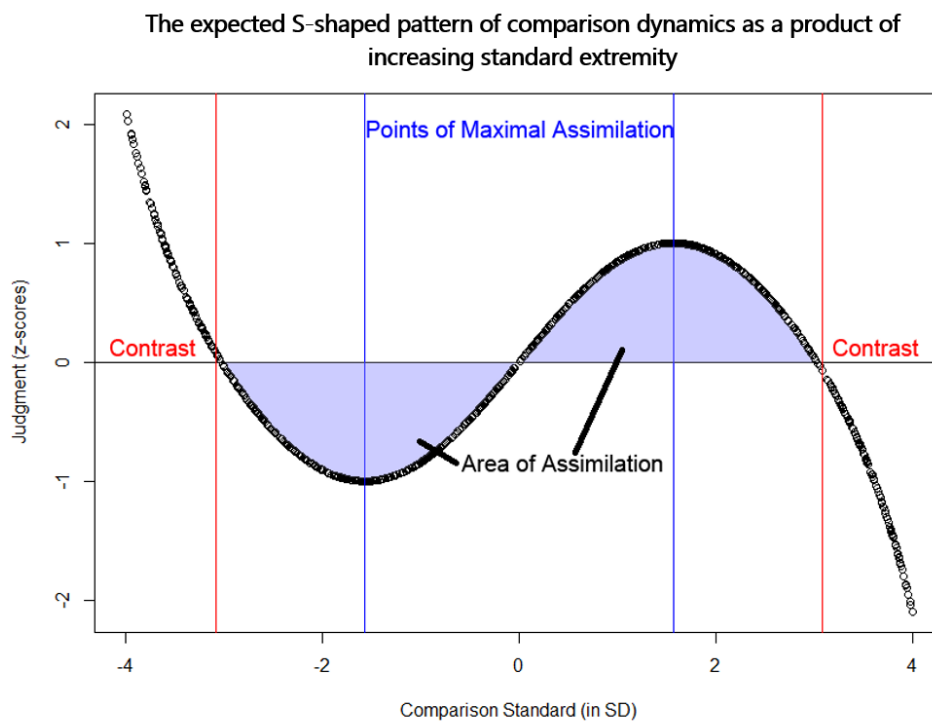


Figure 3.1. The expected S-shaped pattern of comparison dynamics as a product of increasing standard extremity. The x-axis represents the judgment that would be made with a completely neutral standard (0SD)

With this in mind, the issue of ill-defined moderators becomes even more apparent. Not only can any standard be deemed too extreme (or not extreme enough) in a post-hoc fashion when it has resulted in a contrastive (an assimilative) pattern, such critiques can likewise be presented to legitimately explain even null effects by suspecting that assimilative and contrastive

effects on average just cancelled each other out exactly at this position. Think back, for instance, to the lack of evidence for contrastive effect on the Trustworthiness dimension in previous work (Barker et al., 2020). It is perfectly possible that the assimilative and contrastive effects suppressed each other completely precisely at $4SD$ for this judgment and context, and so the broader theory of comparison outcomes remains untouched by these results. In fact, any finding from sources treating moderate and extreme standards as binary states rather than a continuous spectrum, either will confirm the predicted data pattern or may be painted as theoretically inconsequential. Any failure to confirm the predictions is easily attributed to the insufficient calibration of the extremity conditions. Therefore, the selection of single points, or narrow windows to represent moderate or extreme standards, can never suffice for strong tests of the theory itself. Moreover, the lack of clarity surrounding what an extreme standard might be is not only prohibitive to the advancement or falsification of theories of comparison, but with a reversed logic, it also strictly limits the generalisability of existing findings to the exact operationalisation of extremity standard used in the investigation.

3.0.2. The present research

To overcome the above-outlined issues surrounding the use of a dichotomised operationalisation of standard extremity, the current project will be a first attempt at modelling the hypothesised dynamic pattern of assimilation and contrast across a wide range of extremity levels. Adopting such an approach helps us move beyond the pitfalls of narrowly selected comparison standards and provide the first compelling test of the moderating role of target extremity.

For this, we will employ a curve-fitting procedure that will allow for a stronger test of the theoretically predicted pattern as well as providing a more exact definition of what constitutes an

extreme standard to guide future research. The expected S-shaped pattern can be parsimoniously described by a function with a positive linear and a negative cubic effect. If the dynamic interplay of assimilation and contrast indeed behaves as described above, we should obtain a function with these characteristics across a range of comparison standards independent of the exact locations of the turning point for a particular judgment or dimension. This approach lends itself to a critical test of the proposed effect as a failure to achieve this pattern within a reasonable range of comparison standards (e.g., from -4 to $+4$ *SDs*) now speaks against the universality of the theory itself, not against the more or less arbitrarily chosen comparison standards. In addition, the descriptive pattern that accompanies the result will provide the most fine-grained picture of the dynamic interplay between assimilation and contrast to date, which in turn will help to form a first definition of the point at which assimilation becomes contrast.

The current investigation will focus on the three facial dimensions that produced strongly varying comparison effect in previous work (Extraversion, Trustworthiness, and Dominance; Barker et al., 2020). This will not only let us further investigate the cause for these unexpected findings, but the domain of facial evaluations simultaneously presents an excellent ecologically valid way of achieving the extremely fine grained manipulation of standard extremity necessary to determine comparison patterns for multiple judgments in a standardised fashion. Indeed, within the first 100ms of an encounter (Willis & Todorov, 2006; Ballew & Todorov, 2007), a variety of complex social traits can be evaluated just from the facial features of an interaction partner (Todorov, Olivola, Dotsch & Mende-Siedlecki, 2015), and people have elaborate ideas about what others look like based on stereotypical convictions (Imhoff, Woelki, Hanke, & Dotsch, 2013). Faces, thus, allow for fast and unobtrusive presentation of comparative information. In addition, almost endless unique facial stimuli can now be easily created and

precisely manipulated on a number of social dimensions with the use of data-driven algorithms (see Todorov, Dotsch, Porter, Oosterhof & Falvello, 2013), providing the numerous comparison targets and standards needed for the current endeavour.

The initial studies focused on the judgments for Extraversion (Study 1), Trustworthiness (Study 2), and Dominance (Study 3) to see if the fine grained manipulation of extremity would uncover more consistent patterns at the item level than previously found (Barker et al., 2020). Later pre-registered studies investigated the Dominance (Study 4) and Trustworthiness dimension (Study 5) in more detail, using multiple items generated in a data-driven fashion to probe whether the patterns found in the initial studies were caused by characteristics at the item level, or if they are fundamental to the evaluative dimension itself. In doing so, they will further provided the first clear guidelines for future research regarding the operationalisation of a standards extremity.

All studies will report full descriptions of power calculations, data exclusion, manipulations, and measures used. In addition to the presented studies, one small scale pilot study was initially conducted for the full CJT procedure as well as a second one for the CJT with only 41 trials that produced patterns in line with the ones described below (see OSF page for details). Furthermore, all anonymised raw⁷ and aggregated data, and additional materials for all studies can be found on the Open Science Framework page at:

https://osf.io/6zt5w/?view_only=98eff44baf0f468c9275fead26abee28

⁷ Data about participants study area and gender are not included for lab studies nor are any timestamps for all studies to ensure the complete anonymity of participants. These data are available upon request with stricter sharing protections.

3.1. Study 1

To provide the best initial test of the extended CJT paradigm's ability to detect both comparative tendencies, this first study in the series focused on the evaluative dimension of Extraversion, which represents the only judgment dimension to show both an assimilative and contrastive effects in previous studies (Barker et al., 2020).

As described in the introduction, the pattern we expect to see for a judgment in which assimilation is more likely for moderate standards and contrast is greater for extreme ones resembles an S-shape. This pattern should, furthermore, show an initial increase of assimilation to some maximum, after which a decrease will be seen towards ever more contrastive judgments, with a clear turning point where the later becomes more prominent (Figure 3.1). In this study, this pattern will be approximated with a cubic function, where we expect at least a positive linear and a negative cubic term to represent the described dynamic. We would also expect a point at which the function crosses the intercept adjusted x-axis, to indicate the point after which consistent contrast effects can be found and, thereby, what constitutes an extreme standard for this judgment.

3.1.1. Method

Participants. To gain some indication of the sample size that would be necessary in the extended CJT paradigm, we assumed that the peak assimilative effect would be roughly the same size as the effect found for this judgment in Barker et al. (2020). Based on this, the linear effect was assumed to be around $B = 0.17$ and the cubic effect around $B = -0.017$. Simulations with varying parameters showed that a sample size of roughly 80 participants would be needed in the repeated measures design with 324 trials to ensure at least 80% power to find the cubic function even when residuals and variation in the linear slope at the participant level was relatively large.

To leave some room for drop out, a convenience sample of 85 German speakers were recruited on campus at the University of Cologne to participate in the 50 min study for a monetary reward of 10 euros. With the final sample being composed of 61.2% female participants and aged between 19 and 39 years ($M = 25.41$, $SD = 4.41$). It should be noted that the population from which this sample was taken is rather homogeneous with respects to education and cultural background, matching the populations used in many previous studies on the comparative effect (e.g., Mussweiler, & Strack, 2000; Mussweiler, et al., 2004). Although this will increase the chance of replicating previous findings, it will limit the generalisability of the patterns to the current population.

Comparative Judgment Task (CJT). An expanded version of the Comparative Judgment Task (CJT; Barker et al., 2020) was used to investigate the hypothesised comparative patterns. Across 324 trials, participants evaluated a neutral target face on a particular judgment dimension in an open-ended format as these have been previously used to elicit comparative processing without explicit prompting to compare (Mussweiler et al., 2004), and also avoid enforcing relative judgments which can be the case for closed scales (Mussweiler & Strack, 2000). At the same time, a second face filled the role of the comparison standard shown alongside the target. Unlike the target, which was neutral with respect to the evaluative dimension throughout the trials, these standards varied in their standing on the dimension, ranging from extremely high upward ($+4SD$ from the mean of all faces in face space) to extremely low downward comparisons ($-4SD$). To ensure participants understand which face needed to be judged as the target, they had to identify the face labelled as the ‘Judgment target’ prior to making their judgments. In the current set-up, this was done by pressing either ‘Z’ for the left (on the German keyboards used this was the letter ‘Y’ in the same location) or ‘C’ for the

right face. This step also formed an attention check to exclude non-informative responses from the final dataset, see Appendix C for an example trial. In addition to the measurement instances used in previous CJT paradigms, the current set-up was significantly expanded to include 4 measurement instances for all $0.1SD$ incremental extremity steps between $-4SD$ and $4SD$ away from the mean of all faces, resulting in a total of 81 steps and 324 trials.

Stimuli. For this initial study of the Extraversion dimension, an item based on the one used to measure this evaluative dimension previously was included to give the best chance of replicating the pattern found in this work; i.e. "How often does the target go out in 6 months?" (Barker et al., 2020).

The facial stimuli needed for the CJT were created using a custom scripts built on the FaceGen SDK according to the process used in Todorov, et al. (2013). In this process, a unique random ID is first created after which it can be manipulated precisely across a range of facial dimensions to create numerous computer-generated facial images. In the current study, the neutral IDs were manipulated along the vector of facial Extraversion to represent standards at every $0.1SD$ step from neutral in a range from $-4SD$ to $+4SD$. This was done for 4 unique IDs at each of the 81 steps to produce a total of 324 unique facial pairs, see Appendix A for examples. The neutral faces of each pair formed the judgment targets, while the non-neutral faces acted as the various levels of comparison standards.

Procedure. Participants were recruited on campus at the University of Cologne, and were fully informed regarding the general procedure of the study and data storage policy before giving their consent and taking part in the study. In the first part of the study, some basic demographics, such as age, sex, and education were recorded, after which the CJT was explained in detail and participants conducted two practice trials. If participants had no remaining questions regarding

the task they then started the main batch of 324 trials in random order. After completing the CJT trials, participants were debriefed and given their compensation.

Data treatment. Non-numeric and empty responses made up 9.4% of trials, while 8.8% showed a failed attention check leading to their exclusion in the analyses. The remaining scores were then used to calculate z-scores separately per participant to account for personal differences in response ranges. Z-scores above 3 or below -3, or instances where no z-score could be calculated were removed, which was the case for 2% of trials. Due to co-occurrences of these criteria, a total of 18.5% of the original trials were not used in the analyses. In the case of eight participants, this meant no usable trials remained, leaving the data of 77 participants to be used in the analyses.

3.1.2. Results

The main analysis was performed in R (Version 3.5.1; R Core Team, 2018) and consisted of a mixed models regression using a restricted maximum likelihood estimation (REML; using the lme4 package; Bates, Maechler, Bolker, & Walker, 2015) with fixed effects for the extremity steps up to the third polynomials term, and with similar orthogonal uncorrelated random slopes and intercepts for participants to account for participant level variation in comparison patterns. To determine the confidence intervals and p-values for the fixed effects, in this and all following studies, Satterthwaite's approximations were used to estimate the appropriate degrees of freedom (using the lmerTest package; Kuznetsova, Brockhoff, & Christensen, 2017; and the parameters package in R; Lüdtke, Ben-Shachar, & Makowski, 2020).

Results showed that in the cubic model, both the predicted positive first order and negative third order terms were present and significant, but with magnitudes far smaller than

expected (Table 3.1). The fitted values still described the expected S-shaped curve that would be expected with the presence of both assimilation and contrast that can be seen in Figure 3.2.

Table 3.1
All fixed effects and related statistics from mixed model analysis

	<i>B</i> [95% CI]	<i>SE</i>	<i>df</i>	<i>t</i>	<i>p</i>
Fixed effects:					
x	0.0191 [0.0059; 0.0324]	0.0068	127.321	2.826	.005
x ²	-0.0012 [-0.0037; 0.0012]	0.0013	91.510	-0.984	.328
x ³	-0.0020 [-0.0031; -0.0008]	0.0006	290.346	-3.350	<.001
Intercept	-0.0176 [-0.0353; 0.0000]	0.0090	162.513	-1.957	.052

Secondary analysis, using a slightly adjusted two lines test, with the Robin Hood algorithm implemented to calculate the break point (Simonsohn, 2018), were run to also provide separate evidence of assimilation (i.e. an initial linear increase of z-scores as extremity increases) followed by increased prominence of contrast effects (i.e. a negative linear relationship after some peak has been reached). This procedure estimates a separate regression line for low and high values of the predictor, leaving more flexibility in the functional form that describes the relationship. However, to account for the clustered nature of the data a Mixed GAMM with random intercepts and linear slopes for participants was preferred (using the *gamm4* package in R; Wood & Scheipl, 2017), as were clustered robust standard errors by participant (using the *sandwich* package in R; Zeileis 2004; Berger, Graham & Zeileis, 2017). Robust LMM were initially also attempted to fit the full model, but were too large a strain on the memory capacity of the available hardware. The data set was also simplified for this analysis by disregarding the direction of the comparison, with the steps now representing the absolute distance from the neutral target and z-scores always reflecting assimilation when positive and contrast when

negative. In line with the expectations, the results of this analysis showed both a significant average assimilative effect up to $0.8SD$, $B = 0.19$, $Z = 3.75$, $p < .001$, and a negative one beyond that, $B = -0.03$, $Z = -3.89$, $p < .001$.

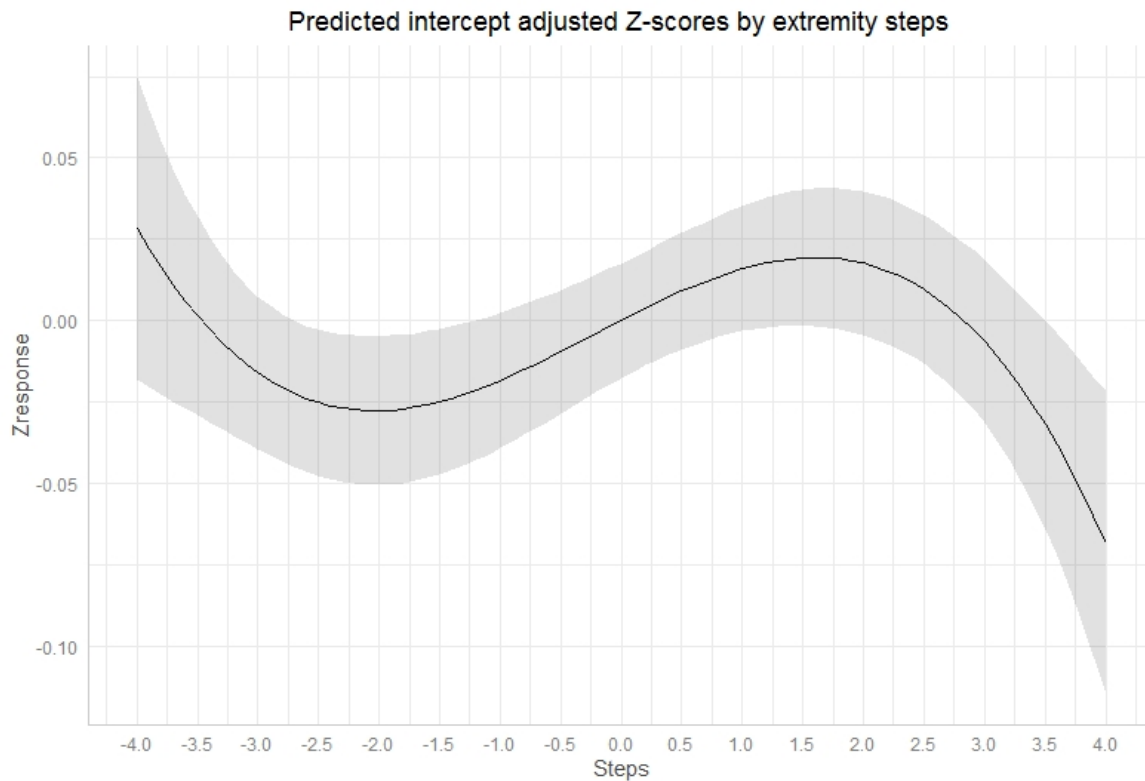


Figure 3.2. Predicted intercept adjusted marginal z-scores at each extremity step and predicted 95%CI (created with the ggeffects package, Lüdtke, 2018)

The question remains whether the pattern is one of reduced assimilation only or contrast proper (i.e. is there a point at which judgments are on average consistently below the neutral judgment). For this, we first turn back to Figure 3.2, where one can see the curve clearly crosses the x-axis (representing a judgment with a neutral standard) for both comparison directions, implying there is indeed a point at which the estimated marginal effect turns from consistent assimilation to contrast. To estimate these points and calculate their confidence intervals

precisely, a simplified model using maximum likelihood estimation with only a random linear slopes for participants was used in a bootstrapping procedure with 3000 iterations to calculate percentile CIs (utilizing the *boot* package; Canty & Ripley, 2019). For each iteration, the resulting polynomial curve was solved for zero where possible (with the *polynom* package; Venables, Hornik & Maechler, 2016). If an iteration only showed signs of assimilation across the range of measured instances the maximum measured step of $4SD$ was returned, while if only contrast effects were present the returned value was set to the minimum value of $0SD$. The returned values, thus, give a representation of the outer bounds of the window in which assimilation is the dominant tendency within the measured range. The procedure estimated these points to be as low as $-3.46SD$, $SE = 0.39$, $Bias = 0.01$, 95% CI $[-4.00, -2.71]$, and as high as $2.82SD$, $SE = 0.41$, $Bias = -0.01$, 95% CI $[1.97, 3.52]$, with the complete window of assimilation spanning $6.29SD$, $SE = 0.50$, $Bias = -0.02$, 95% CI $[5.38, 7.01]$, indicating that there were clear signs of assimilation and contrast overall, although it did not provide evidence for a turning point for downward standards separately (as -4.00 was within the 95% CI). Nevertheless, the results taken together seem to convincingly support the expected pattern of assimilation to more moderate and contrast from more extreme standards for this judgment overall, and the ability of the paradigm to detect both.

3.1.3. Discussion

This first implementation of the extended CJT successfully detected both assimilation and contrast effects in the expected S-shaped pattern for a judgment that elicited both in previous work. It, thereby, provides some evidence for the predicted moderating role that standard extremity has on the outcomes of comparative judgments, at least for the current judgment related to the evaluative dimension of facial Extraversion. Furthermore, the bootstrapping

procedure successfully provided a first insight into the scope of the window of assimilation and what constitutes an extreme standard, although confidence intervals remained quite wide with the current sample size.

Related to this issue, one should also note the rather small size of the effects, which may partially be the result of the opposing effects that are at play, likely interfering with each other to some degree at any level of comparison, since target extremity is but one of many moderators that affect the comparison outcome (Mussweiler, 2003). These issues could be exacerbated by the specifics of the current paradigm, which uses open ended responses and facial evaluations of others, which provide only minimal information for participants to base their judgments on. Additionally, comparisons are known to often be made in an egocentric manner, with trait dimensions often understood idiosyncratically and based to some extent on one's own behaviours and characteristics (Dunning & Hayes, 1996; Dunning, Meyerowitz, & Holzberg 1989). These issues could lead to larger variations in judgments overall, and a lower reliability of the measure within the context of other-related facial evaluations. Nevertheless, the predicted effect was still detected successfully, even with these caveats, highlighting the fundamental nature of comparative judgments, which can affect judgments even in the briefest of evaluative encounters.

The next study will turn towards the judgment dimension of Trustworthiness, which did not produce any signs of contrast in prior work, to gain a clearer picture of the dynamics for this judgment and if the more fine grained approach of the extended CJT may yet uncover the hypothesised contrast effects.

3.2. Study 2

Unlike the judgment for Extraversion, at least one judgment related to the facial dimension of Trustworthiness has been found to lead to only consistent assimilative tendencies (Barker et al., 2020). The fine-grained measurement of the Extended CJT can help further clarify this pattern, and if this specific judgment does indeed not produce contrast effects. For instance, a positive linear trend alone would be a sign of increasing assimilation, while if paired with a negative cubic term this would suggest a reduction in, or at least a limit to, the strength of assimilation for more extreme standards. If, counter to the expectations, consistent contrast does occur for this judgment, the fitted curve must also include a turning point that falls within the measured range where average judgments are contrastive with respect to the neutral judgments.

3.2.1. Method

Participants. Considering the small effect sizes and large CIs found for the turning point in Study 1, the sample size for this study was roughly doubled for increased measurement accuracy. Therefore, 160 German speakers were again recruited on campus at the University of Cologne to participate in the study for a monetary reward of 10 euros. One participant dropped out during the study, resulting in 159 completed cases being recorded. This final sample consisted of 62.9% females and was aged between 18 and 63 years ($M = 24.09$, $SD = 6.05$).

Stimuli. Once again, unique IDs were created using the procedure described in Study 1, but they were this time transformed along the vector of the Trustworthiness dimension, creating 324 facial pairs of neutral judgment targets and comparison standards. The judgment item for this initial test of the dimensions was the reversed item “How many times does the target deceive somebody in 6 months?”, again based on the one used in Barker et al. (2020) where only assimilative patterns were reported for this judgment.

Additional Measures. Age, sex, education, and area of study were again measured as basic demographics in this study. In addition, the Iowa-Netherlands Comparison Orientation Scale (INCOM; Gibbons & Buunk, 1999) was administered for exploratory reasons. The INCOM scale consists of 11 items ($\alpha = .77$) that are averaged to create an INCOM score, with higher scores indicating an individual has a higher disposition to engage in social comparisons in daily life. Items in the scale focus on comparisons of one's own abilities or opinions with those of others. The scale was, therefore, included at the end of the study to explore the possibility that this construct might extend to broader tendencies for making comparison of others as well. As this measure did not produce any interesting findings in any of the studies, the results will not be reported in this paper, but the relevant data are of course available in the supplemental materials.

Procedure. The procedure was identical to that of Study 1, with the exception that the INCOM scale was administered after the CJT was completed and before participants were debriefed and received their reward.

Data treatment. As was the case in the previous study, trials with null responses or only non-numeric symbols were excluded from analyses (3.7 % of trials), as well as trials in which the attention check was failed (7.1%). Remaining scores were z-transformed per participant where possible, after which the resulting z-scores were truncated above 3 and below -3 (2.6%). A total of 12.7% of all trials were excluded by these criteria. For three participants this meant none of their trials could be used, leaving 156 participants data usable in the analyses. Due to the reversed nature of the judgment item, these scores were inverted so that higher scores reflected more trustworthy behaviour to aid interpretation.

3.2.2. Results

Once again, a mixed models regression with REML estimation was implemented, with polynomials up to the third degree for the extremity steps as fixed effects and the full orthogonal and uncorrelated random slopes and intercepts for each participants. The predicted negative third order term was again significant, as was the positive linear term consistent with the possibility of both assimilation and contrast (Table 3.2). However, it seems the estimated marginal z-scores indeed decreases with more extreme standards, but never cross the axis, which would indicate consistent contrast never actually occurs within the measured range for this judgment (Figure 3.3).

Table 3.2
All fixed effects and related statistics from mixed model analysis.

	<i>B</i> [95% CI]	<i>SE</i>	<i>df</i>	<i>t</i>	<i>p</i>
Fixed effects:					
x	0.035 [0.0249; 0.0443]	0.0050	210.423	6.966	<.001
x ²	0.0012 [-0.0006; 0.0029]	0.0009	154.630	1.282	.202
x ³	-0.0018 [-0.0025; -0.0010]	0.0004	1603.829	-4.644	<.001
Intercept	0.0305 [0.0179; 0.0430]	0.0064	288.020	-4.768	<.001

To see if separate evidence can be provided for the initial increase and apparent decrease in scores, the adjusted two lines procedure was implemented once again on a simplified unidirectional dataset. A significant initial linear increase signalling assimilation was found up to 1.9SD, $B = -0.03$, $Z = 3.25$, $p = .001$. However, beyond this point the linear decrease failed to reach significance at the standard level, $B = -0.015$, $Z = -1.88$, $p = .06$, indicating the increased prevalence of contrast for more extreme standards remains unsupported for this judgment.

Despite the visual evidence against the presence of a turning point and the lack of convincing evidence for even a decrease in assimilative effects in the two lines test, the bootstrapping procedure detailed in Study 1 was conducted for this dataset as well to confirm the conclusions from the other analysis. In line with the evidence from the other analysis, all of the 3000 iterations returned no turning point for upward comparisons within the measured range of $4SD$, $SE = 0.006$, $Bias = -0.00$, $95\%CI [4, 4]$ and only a limited few for downward ones, $-4SD$, $SE = 0.11$, $Bias = -0.07$, $95\%CI [4, -3.62]$, resulting in the window of assimilation spanning the entire range of measured steps, $8SD$, $SE = 0.11$, $Bias = -0.07$, $95\%CI [7.62, 8]$. These results combined with the previous analysis suggest that only consistent assimilation occurs for this judgment.

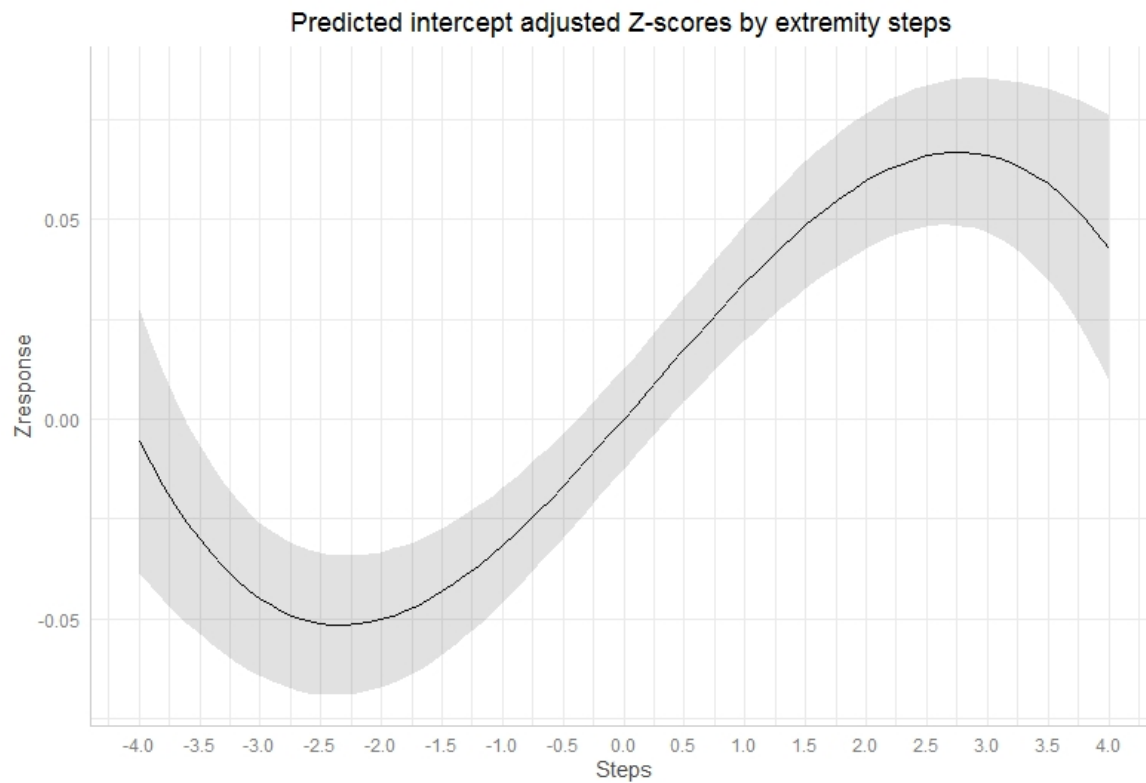


Figure 3.3. Predicted intercept adjusted marginal z-scores at each extremity step and predicted 95%CI

3.2.3. Discussion

In alignment with previous work, the current judgment showed only convincing evidence of increasing assimilation that levelled off for more extreme standards, but did not show significant signs of a reduction in strength, let alone contrast. It is of course possible that even more extreme standards outside of this range might still produce contrast effects. However, broadening the possible window of assimilation even further would mean even fewer standards would exist in the population capable of eliciting the contrast effect, making it decreasingly relevant in any real world scenario.

Therefore, it can be concluded that for the current judgment no meaningful contrast effects occur, regardless of the extremity of the standard. However, there does seem to be some upper limit to the size of the assimilation effect. Whether this pattern is indeed limited to the exact judgment presented here, or is a broader phenomenon for all judgments related to the Trustworthiness dimension, is not yet clear, and will be further investigated in Study 5. However, first we will turn to a judgment that has produced the opposite pattern, with only signs of consistent contrast effects detected, namely a judgment related to the evaluative dimension of Dominance.

3.3. Study 3

In contrast to the text book pattern found for the Extraversion judgment and the purely assimilative effect for the Trustworthiness judgment, a surprising lack of assimilative effects were found in previous work for the facial dimension of Dominance (Barker et al., 2020). One reason this might have occurred is the fact that the judgment only has an exceptionally narrow window of assimilation, rather than a complete lack of consistent assimilation effects altogether. If this is the case, measuring such an effect would be challenging without a very fine-grained

manipulation of standard extremity. The extended CJT paradigm was, therefore, implemented to investigate this possibility directly and estimate the window of assimilation if present.

Once again, a positive linear and negative cubic term would be expected if both assimilation to moderate and contrast from extreme standards occurs with a turning point that falls within the measured range. Alternatively, if only contrast effects are produced by this judgment dimension, only a negative linear or negative cubic slope would be expected, with no turning point and the window of assimilation estimated to be zero.

3.3.1. Method

Participants. With similar power considerations as in Study 2, 160 German speakers were recruited on campus at the University of Cologne to participate in the study for a monetary reward of 10 euros. Two participants dropped out before completion, leaving a final sample of 158 participants, 38.6% of which were female and who had an age between 18 and 42 years ($M = 22.52$, $SD = 4.14$).

Stimuli. The necessary 324 facial pairs were generated in the same manner as described in Study 1, but the faces that formed the comparison standards were now varied on the Dominance rather than Extraversion dimension. The judgment item was based on the one used in Barker et al. (2020) to ensure the best chance of replicating the consistent contrast effects it seemed to elicit; “How many times does the target enforce his opinion in 6 months?”.

Procedure. The procedure was identical to that of Study 2.

Data treatment. The same data treatment was administered as in the previous studies. Trials were excluded from analyses if they were non-numeric or empty (4.3% of trials), or if they were preceded by a failed attention check (11%). All other trials were used to calculate z-scores per participant where possible, after which they were truncated above 3 or below -3 as

representing extreme values (1.4%). Combined this meant that a total of 15.8 % of all trials were not included in the final analyses, with 7 participants not providing enough usable trials to be included in the analyses. The final sample used in the analyses thus consisted of 151 participants.

3.3.2. Results

The main analysis was identical to that used in study one, utilizing a mixed models regression with REML estimation with fixed effects up to the third polynomial for the extremity steps, with identical but orthogonal and uncorrelated random slopes and intercepts for each participant. The cubic model showed a small significant negative third order term, in line with the predictions, but this was not the case for the expected positive linear term, which in this dataset was negative and non-significant (Table 3.3). In Figure 3.4, one can see a lack of consistent assimilative effects for the moderate standards, but with slight contrast away for more extreme standards.

Table 3.3
All fixed effects and related statistics from mixed model analysis.

	<i>B</i> [95% CI]	<i>SE</i>	<i>df</i>	<i>t</i>	<i>p</i>
Fixed effects:					
x	-0.0030 [-0.0146; 0.0087]	0.0060	133.074	-0.497	.621
x ²	-0.0018 [-0.0041; 0.0005]	0.0012	151.463	-1.535	.127
x ³	-0.0008 [-0.0016; -0.0001]	0.0004	143.981	-2.229	.027
Intercept	-0.0230 [-0.0388; -0.0072]	0.0081	277.073	-2.848	.005

To ensure the more restrictive cubic form of the function did not mask any weak assimilative effects at very small intervals, the more flexible adjusted two lines test was implemented again on a simplified unidirectional dataset. Results showed no evidence for an initial increase, but rather a non-significant negative slope up to 0.7SD, $B = -0.03$, $Z = 0.07$, $p =$

.482, with significant signs of contrast for values beyond that point, $B = -0.02$, $Z = -5.51$, $p < .001$. These findings seem in line with the visual inspection of Figure 3.4, where no clear signs of assimilation present themselves, but where contrastive effects do increase with more extreme standards.

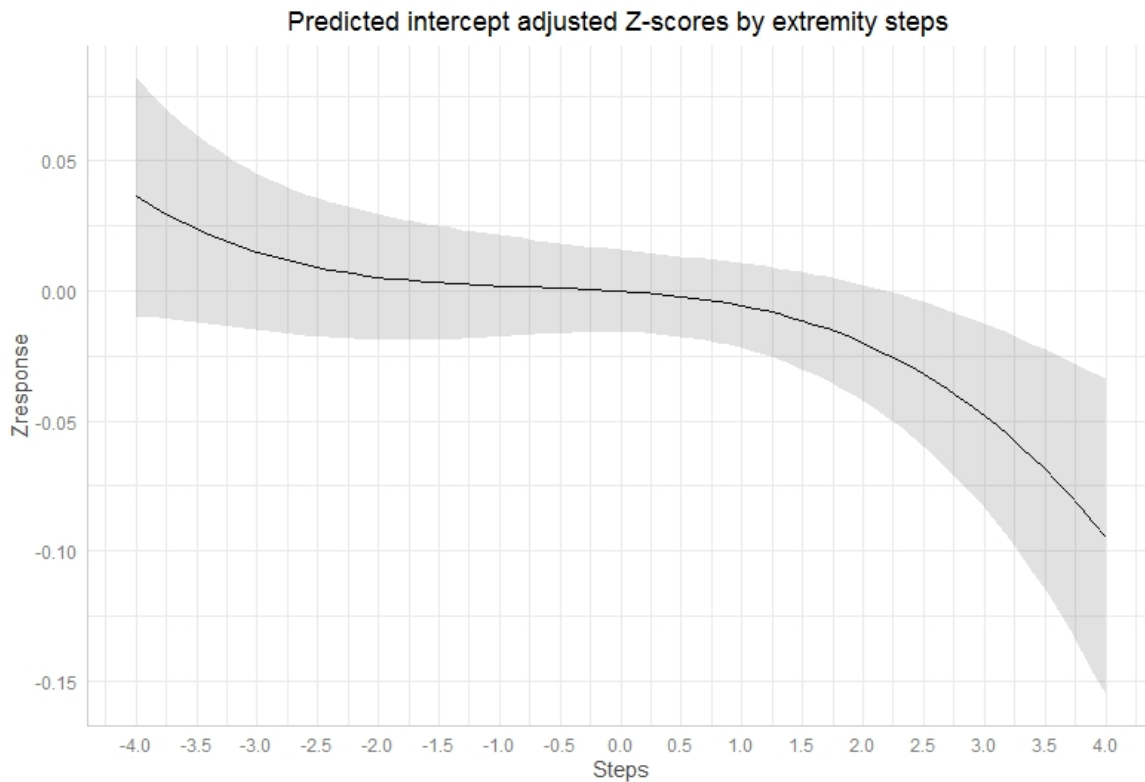


Figure 3.4. Predicted intercept adjusted marginal z-scores at each extremity step and predicted 95% CI

Despite the lack of any evidence of assimilation, even for standards very close to neutral, a similar bootstrapping procedure as outlined in Study 1 was conducted to nevertheless attempt to estimate the turning point. Unsurprisingly, the large majority of the 3000 iterations returned an assimilative window of zero, $SE = 1.25$, $Bias = 0.64$, 95% CI [0, 3.85], meaning there were only signs of contrast in most cases, for both upward, 0 , $SE = 0.38$, $Bias = 0.17$, 95% CI [0, 1.35] and downward comparisons, $0SD$, $SE = 0.90$, $Bias = -0.47$, 95% CI [0, -2.71]. In agreement with the

expectations, it seems that there is no consistent assimilative tendency at any level for this specific judgment.

3.3.3. Discussion

In line with previous findings for this judgment dimension (Barker et al., 2020), no consistent signs of assimilation were obtained with no window of assimilation for even the narrowest ranges. On the contrary, the results did show increasingly consistent contrastive judgments as standards became more extreme. These results, therefore, indicate that for some judgments no consistent assimilation effects are present, no matter how moderate the standard appears. Instead, such judgments may only show contrastive tendencies that become increasingly strong for more extreme standards.

Taken together with the results found for the Trustworthiness judgment in Study 2, it is clear that the specific judgment one is making can have direct consequences on the comparison pattern that will be observed. This is true even to the extent that two completely contradicting conclusions regarding the comparison process can be drawn for different judgments. One in which contrast does not occur even for the most extreme standards, and one where it is the sole influence on outcomes. Whether this strong variation in the comparison pattern is limited to certain items, or is caused by the dimensions of Trustworthiness and Dominance that underlie them, is still unclear. The next study will, therefore, use a shortened CJT procedure, designed to reduce strain on participants and made to be suitable for online testing, in order to investigate the comparative pattern for the Dominance dimensions across multiple items.

3.4. Study 4

With the previous two studies showing such different patterns, it is clear that the type of judgment one is making can profoundly influence the dynamic pattern of assimilation and

contrast, as well as its relation to the extremity of the comparison standard. However, what is still unclear is if this variation resides at the item level, or is the result of the underlying evaluative dimension as a whole. This is especially relevant as participants often idiosyncratically construe trait dimensions in a somewhat egocentric manner (Dunning et al., 1989; Dunning & Hayes, 1996). Therefore, this pre-registered study investigated the Dominance dimension by using multiple items to measure the comparative pattern. These items were generated by and pre-tested on other participants from the same population as those that took part in the final study. The pre-registration documentation can be found in the supplemental materials, and was followed in full unless explicitly stated otherwise.

As in all previous studies, a positive linear term would indicate assimilation, while a negative cubic term would mean contrast is present also. Finally, the turning point will be estimated to attempt to determine the window of assimilation for this dimension as a whole, where none was found in the previous study.

3.4.1. Method

Participants. Using simulations, a sample size of 800 participants was determined to be sufficient to find effects of a similar size to those found in a pilot test of the paradigm using only 41 trials and the single Extraversion judgment, with some room for drop out, 80% of the time.⁸ An online sample of 800 U.S. based MTurk workers were recruited for a monetary compensation of \$1.20. Nine participants completed the study without claiming their compensation and, thus, were not accounted for in the quota of 800, meaning the total sample of completed responses

⁸ Findings for this pilot test will not be discussed here in detail to conserve space, but the pattern found in Study 1 was replicated successfully, with a significant linear, $t(1055.44) = 2.50$, $B = 0.0156$, $p = .013$, and cubic effect, $t(748.80) = -3.96$, $B = -0.0023$, $p < .001$, and a turning point around $2.61SD$. A more detailed description can be found in the supplemental materials.

consisted of 809 participants. This final sample was 53.8% female and was aged between 20 and 96 years ($M = 37.62$, $SD = 11.73$). Although this sample is slightly more diverse than the lab based ones when it comes to education, it was created by once again sampling from an overwhelmingly western population. It thus remains to be seen if the comparative patterns and the size of the AoA would extend to populations with other cultural backgrounds.

CJT-41. The extended CJT was reduced to require only 41 trials per participants in order to decrease strain on participants, creating the CJT-41. To still conserve the same granularity with the reduced number of trials the CJT-41, the exact steps of extremity that participants were exposed to was not the same for all participants. Instead, participants were randomly assigned to one of two groups, the first of which judged 41 faces with standards that varied between $-4SD$ to $+4SD$ in intervals of $0.2SD$. The second group judged faces for 40 steps of $0.2SD$ starting at $-3.9SD$ to $3.9SD$, as well as the neutral $0SD$ comparison. Combined the two groups, thus, provided measurement instances at each of the 81 steps that were also included in the extended CJT.

In addition to the shortening of the task, participants were now randomly assigned to answer one of 6 items for all trials, in order to investigate if assimilation and contrast may occur across these items. Additionally, the attention check, which was mapped to key board buttons in the previous studies, was now completed by clicking a radio button underneath the images to identify the judgment target.

Stimuli. Eighty-one facial pairs similar to those used in the previous studies were created, with comparison standards at each of the $0.1SD$ steps between $-4SD$ and $4SD$. To develop the range of items needed for the study in a more data driven fashion, a two-phased pre-test was conducted. The first part consisted of asking an online sample of 49 participants to provide 5

examples of particularly dominant behaviours and 5 examples of submissive behaviours in an open ended fashion. The responses were then processed into a list of the most frequently mentioned examples of behaviours. The 10 most commonly given examples of dominant and 10 most common submissive behaviours were then used in the second phase, where a new sample of 180 participants made a single judgment for every item about an isolated face, which randomly varied along the dominance dimension in $1SD$ steps between $-4SD$ and $4SD$. This resulted in approximately 20 measurement instances at each step for each item. The resulting data were then included in a mixed models linear regression for each item separately. The three items that showed the strongest linear relationship for submissive behaviour were included as reversed items in the final CJT-41. For the dominant behaviours, the top two items were selected as well as the same item used in Study 3, see Table A1 in Appendix E for a list of all the selected items. A more in depth description of the pre-testing phase can be found in the supplemental materials.

Additional measures. Besides the basic demographics of age, sex, and education, an additional item at the start of the survey was included, asking participants to describe shapes presented on an image in order to make sure the images were indeed loading correctly. Furthermore, an item was included after the main CJT-41 was over, posing the same question participants had been asked throughout, but this time they were to use themselves as the judgment target. Both these items were not used in any of the analyses, but their data are included in the supplemental materials.

Finally, a new data-quality item was included at the end of the study, which allowed participants to report if their responses were made in a conscientious manner, and if the resulting data should be used or not, to increase the quality of the data in the final analyses. Responses to

this item were guaranteed not to negatively affect participants or their compensation in any way, but were merely to help clean up the data for use. Responses to this item ranged from “Definitely do not use my data” (1) to “Definitely use my data” (4). Any responses of 2 or lower were used to exclude participants from the analyses.

Data treatment. As in all previous studies, trials with non-numeric or empty responses were removed (0.2%), as were those in which the attention check was failed (4.8%). Responses were then used to calculate z-scores per participant and truncated above 3 or below -3 (2.2%). For the three reversed items, the z-scores were flipped so that higher scores always reflected more dominant behaviours. At least one of these exclusion criteria was met in 7.3% of all trials, resulting in twelve participants not providing any usable trials. Finally, the new data-quality item was used to further exclude 26 participants who indicated they did not think their data were of high enough quality to be included (a score of 2 or lower), leaving 771 participants in the dataset for use in the analyses.

3.4.2. Results

Once again, the initial analysis consisted of a mixed models regression, using a REML estimation, but considering the reduced number of measuring points a different model structure was used. As the first and third order polynomial terms for the extremity steps are of most theoretical relevance for detecting assimilation and contrast, and there were no signs of a meaningful quadratic term in the previous studies, along with recent meta-analytical evidence showing no differences between upward and downward comparisons (Gerber, Wheeler & Suls, 2017), only these two terms were pre-registered to be included as fixed effects in the model. As a result the model will no longer be able to distinguish between upward and downward comparisons and, thus, we will also not estimate the turning point for upward and downward

comparisons separately, as these will necessarily be fully symmetric under the models constraints. The full random slope structure and intercepts were fitted per participant, with only random intercepts included per item, as the data were not sufficient to support the addition of the linear and cubic effects for each item.⁹ Note, that this entails that the results related to the linear and cubic terms cannot strictly be generalised to the evaluative dimension as a whole beyond the items that were used in this investigation.

In contrast to the results found in Study 3 using the Dominance dimension, the data from this study did show evidence for assimilation, reflected in a significant first order effect, as well as for contrast, seen in the negative third order terms, see Table 3.4, which together describe the hypothesised the S-shaped curve, see Figure 3.5.

Table 3.4
All fixed effects and related statistics from mixed model analysis.

	<i>B</i> [95% CI]	<i>SE</i>	<i>df</i>	<i>t</i>	<i>p</i>
Fixed effects:					
x	0.0155 [0.0036; 0.0274]	0.0061	1097.68	2.55	.011
x ³	-0.0020 [-0.0031; -0.0009]	0.0006	774.38	-3.62	<.001
Intercept	-0.0001 [-0.0263; 0.0265]	0.0053	5.02	0.01	.994

Additional analyses were again performed to formally test whether the descriptive S-shaped curve includes both a significant initial increase and subsequent decrease of scores respectively. A simplified unidirectional dataset was, therefore, used in the adjusted two lines procedure, with random intercepts for participants. Both the linear increase up to 1.1SD, $B =$

⁹ The random effects structure used in the final model diverges from the one described in the pre-registered model, which only mentions linear random effects and intercepts for participants and no random intercept for each item. The decision to include these extra random effects was made to increase the generalizability of the model after it became clear that the data could support this more complex structure. Results using the simplified structure from the preregistered analyses are almost identical and do not change the conclusions.

0.06, $Z = 2.18$, $p = .03$, and the decrease thereafter, $B = -0.03$, $Z = -3.56$, $p < .001$, were found to be present and significant in these data.

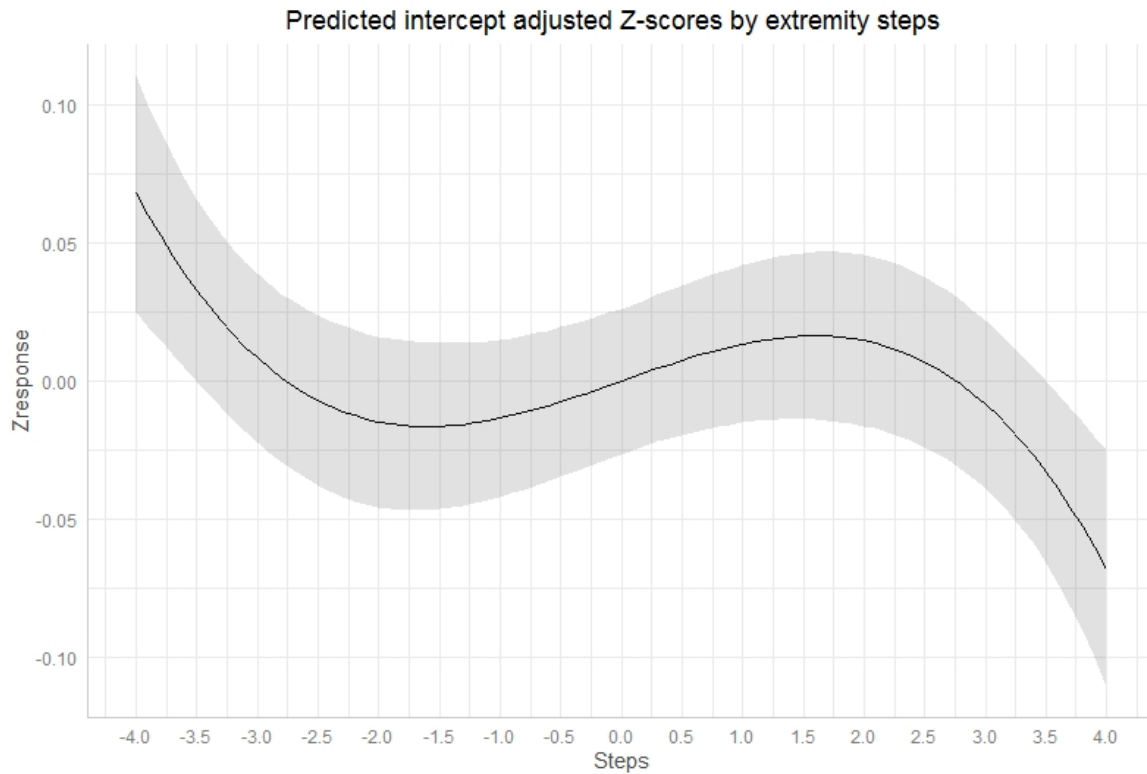


Figure 3.5. Predicted intercept adjusted marginal z-scores at each extremity step and predicted 95%CI

Finally, the bootstrapping procedure was implemented again as was done in previous studies. The results after 3000 iterations showed the average turning point for the judgments was around $2.76SD$, $SE = 0.34$, $Bias = -0.05$, 95% CI [1.99, 3.14], with the window of assimilation spanning $5.52SD$, $SE = 0.68$, $Bias = -0.10$, 95% CI [3.98, 6.27]. Taken together, the results suggest that the consistent contrast found in Study 3 is indeed a reflection of item level, rather than dimension level, heterogeneity in comparison patterns.

3.4.3. Discussion

In contrast to the results from Study 3, clear assimilation and contrast effects emerged for the Dominance dimension across the different judgment items used. This suggests the strong heterogeneity in comparison effects is likely caused by item level variables and is not invariably caused by the evaluative dimension as a whole. Regretfully, the current work cannot speak towards the exact item level conditions that are responsible for the purely contrastive effects that presented themselves in Study 3, but suffice it to say that they are profoundly influential on the dynamics at play. Thus, research in this area should steer clear of using single or a small number of handpicked items until these boundary conditions are more clearly defined. Regardless of the reason, these findings do provide evidence that the theoretically predicted moderating role of target extremity can occur for some judgments related to the evaluative dimension of Dominance. The uncovered S-shaped curve has, furthermore, offered a first detailed glimpse of the dynamic pattern of comparison outcomes across a range of judgments. In doing so it has made further steps towards clearer guidelines regarding what constitute an extreme standard. Under the current conditions, we can say this is limited to standards that are more extreme than those at the turning point of $2.76SD$.

However, these guidelines might vary significantly between evaluative dimensions, which is especially likely for the dimension of Trustworthiness, which in Study 2 showed only signs of assimilative effects. The next study, therefore, investigated if the use of additional items for the Trustworthiness dimension similarly led to the theoretically predicted assimilative and contrastive effects overall, reducing the unexpected variation to the item level. If both tendencies were detected, the turning point for the Trustworthiness dimension could also be estimated and compared with that of the Dominance dimension.

3.5. Study 5

The results from Study 4 make it apparent that item level variation can strongly influence the comparison dynamic. It is still to be determined whether such item level effects also underlie the purely assimilative results found in Study 2 for the Trustworthiness dimension. The current study, therefore, investigated the comparison dynamic for the Trustworthiness dimension using multiple items in a pre-registered study. Expectations regarding the comparison pattern were the same described throughout the studies.

3.5.1. Method

Participants. As in Study 4, 800 participants were sought on the online platform MTurk for a \$1.20 reward. The recruited sample was 45.6% female and was aged between 19 and 73 years ($M = 36.01$, $SD = 10.46$).

Stimuli. As in Study 4, a dual phased pre-test was conducted to generate and select six items that related to the facial Trustworthiness dimension most clearly. The procedure was identical to the one used in Study 4, resulting in three items related to trustworthy behaviours and three items to untrustworthy ones, see supplemental materials for more in depth descriptions and Table A2 in Appendix E for all items. Finally, 81 Facial pairs were created as was done throughout the studies.

Data treatment. Data treatment was again in line with the procedure used in all the previous studies. Trials were removed if non-numeric or empty responses were given (4.2%) and if the attention check was failed (8.1%). The remaining scores were z-transformed per participant where possible, with values above $3SD$ or below $-3SD$ being removed (3.5%) after which z-scores for the three reverse coded items were flipped. A total of 15.6% of all trials were deemed non-informative by these exclusion criteria, which meant no high quality trials remained for 25

participants. Another 26 participants were removed as they themselves indicated their data was of low quality, leaving 749 participants to be analysed.

3.5.2. Results

A mixed models regression using a REML estimation was run, with the first and third order polynomial terms for extremity once again included as fixed effects, and with similar orthogonal uncorrelated random slopes and intercepts for participants used in the main analysis. In addition, random intercepts were included for each item, see Footnote 9. As with the previous investigation of this dimension, the model showed both a significant positive first order and negative third order effects (Table 3.5). However, this time the curve did seem to cross the 0-point within the measured range, suggesting contrast may have occurred (Figure 3.6).

Table 3.5
All fixed effects and related statistics from mixed model analysis

	<i>B</i> [95% CI]	<i>SE</i>	<i>df</i>	<i>t</i>	<i>p</i>
Fixed effects:					
x	0.0376 [0.0244; 0.0508]	0.0067	1079.73	5.59	<.001
x ³	-0.0040 [-0.0052; -0.0028]	0.0006	735.32	-6.53	<.001
Intercept	-0.0021 [-0.0168; 0.0205]	0.0095	4.00	0.19	.854

The adjusted two lines procedure was again implemented as an additional analysis of the increase and decrease of the S-shaped curve separately. The assimilative linear increase was found to be significant up to 2.1SD, $B = 0.03$, $Z = 2.28$, $p = .02$, as well as the decrease for more extreme values, $B = -0.07$, $Z = -4.60$, $p < .001$, which is in support of a pattern of assimilation followed by contrast as extremity increases.

Lastly, the bootstrapping procedure was implemented again with 3000 iterations. The results this time did show a clear turning point of around $3.06SD$, $SE = 0.10$, $Bias = -0.01$, 95% CI [2.84, 3.25], with the widow of assimilation spanning $6.12SD$, $SE = 0.20$, $Bias = -0.01$ 95% CI [5.68, 6.50], which is surprisingly close to the turning point found for the Dominance dimension in Study 4. Taken together, the results seem to be firmly in line with the theoretical prediction, meaning the variation found in earlier studies was once again likely restricted to item level effects.

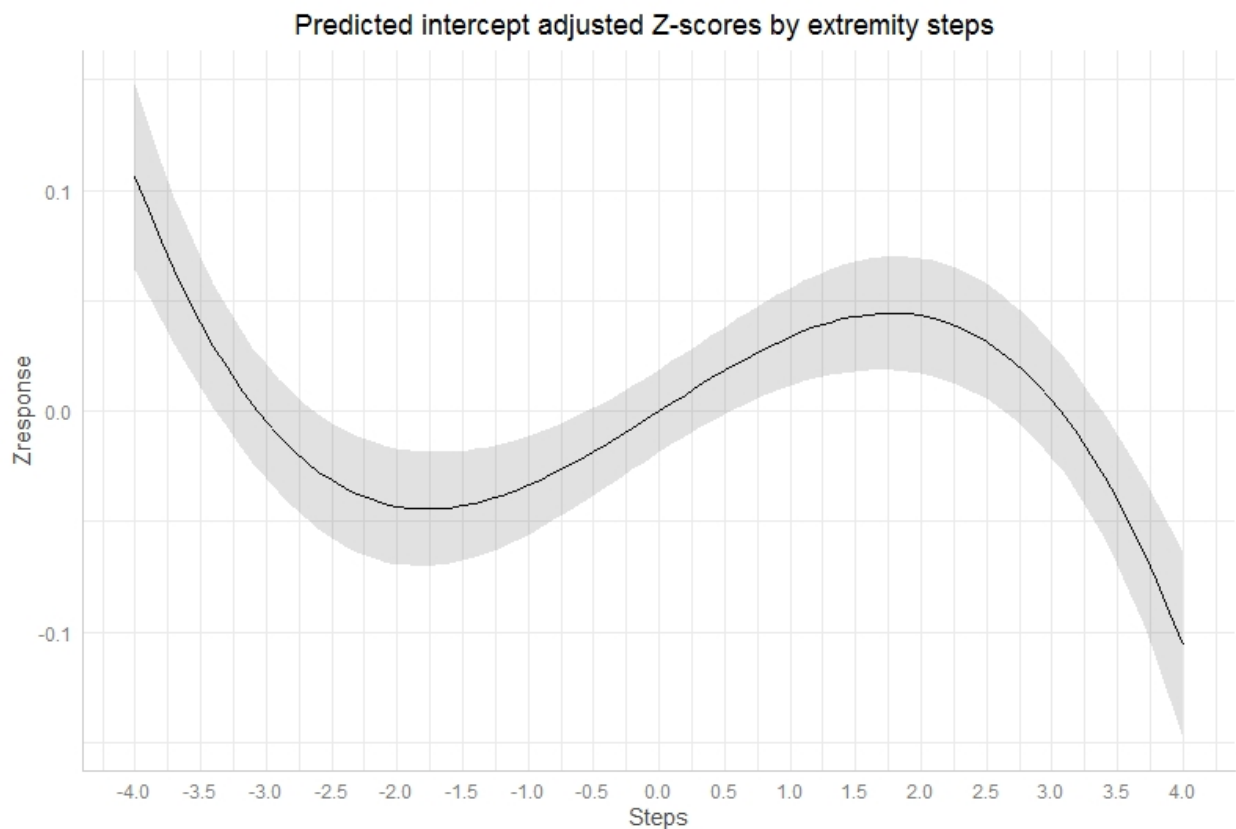


Figure 3.6. Predicted intercept adjusted marginal z-scores at each extremity step and predicted 95%CI

3.5.3. Discussion

The purely assimilative patterns found in Study 2 seem to be related to the item level, rather than that of the evaluative dimension, as was the case for the unexpected findings for the

Dominance dimensions. Across multiple items, consistent assimilation and contrast effect were found dependent on the extremity of the presented standard. These results further highlight the substantial influence item formulations can have on the dynamics of the comparison process, and the need for more clarity surrounding the conditions, which lead to items resulting in opposing outcomes.

Nevertheless, at the level of the evaluative dimension of Trustworthiness, both assimilation and contrast effects presented themselves in the theoretically predicted manner for the judgments used. The uncovered pattern, furthermore, showed a very similar boundary to the assimilative effect as was found for the Dominance dimension, with assimilative effects this time disappearing when standards around $3.06SD$ or higher were present.

3.6. General discussion

A fundamental shortcoming in comparison research to date is the way in which moderating variables such as standard extremity have been arbitrarily dichotomised. This has made strong tests of theoretical predictions almost impossible, as post-hoc judgments of their operationalisation remained an ever present obstacle for falsification and generalisation. The current work has provided the first step to remedy this issue by offering a glimpse into the predicted dynamic pattern of assimilation and contrast across a wide range of standard extremities, while avoiding the pitfalls of arbitrary standard selection. Results revealed evidence consistent with the theorised moderating role that standard extremity has on comparison outcomes, and the predicted S-shaped pattern it would produce for a number of items and evaluative dimensions.

The initial three studies confirm that the variation in comparative effects found in Barker et al. (2020) extend across the full range of standards measured here. This implies that the

unexpected heterogeneity found in this previous work is not likely a result of differences in extremity thresholds and standard operationalisation alone. Instead, it seems to be more fundamental to the judgments themselves. The last two studies demonstrated that, rather than being caused by the underlying evaluative dimension itself, the contradictory patterns found in Study 2 and 3 are most likely a result of effects that reside at the item level. Both evaluative dimensions themselves showed, on average, exactly the expected pattern of assimilation and contrast. These findings, thus, support the theoretical predictions with regards to the moderating role of extremity, even in the minimal context of facial evaluations for these judgments. Furthermore, the patterns uncovered for these aggregated judgments provide clear guidelines concerning the operationalisation of standard extremity for future research, although it should be noted that the substantial item variation found in the initial studies might considerably shift these estimates from item to item. Regardless, across the multiple items and evaluative dimensions used in the final two studies, a consistent pattern was found in which the assimilative peak could be found around approximately $1SD$ to $2SD$, with a turning point from assimilation to contrast close to $3SD$ from the target. Therefore, if one were to define extreme in the context of comparative judgments, anything above $3SD$ might be an appropriate rule of thumb based on the current data, with greater values more likely to produce consistent contrast effects.

An important question that remains unanswered is which conditions underlie the strong item level fluctuations that were found in the initial studies. Without additional research to better understand how different items can produce such contradictory findings, researchers should be cautious when designing items to sample from, as well as when interpreting findings from the existing literature when conducted with only a few items. Even within the current work, which sampled multiple items, the uncovered comparative patterns strictly only extend to the

dimensions, items, stimuli, samples, and contexts in which they were tested, and cannot be generalised without making strict assumptions. This issues compound when we consider that the initial items and dimensions used in the studies were purposely selected from previous work based on the fact that they indeed showed some consistent comparison effects, where other items appeared to elicit none (e.g., Likability; Barker et al., 2020). Whether these comparative patterns and the estimated turning points would be comparable for other facial dimensions, or even extend beyond the facial domain itself, will need to be the subject of future research. However, the relative consistency of results found here at the dimensional level seems encouraging.

Notwithstanding the limits to the generalisability of the conclusions, the current work is a first step towards producing clearer boundaries for moderating variables, and is the first to provide a detailed description of the underlying dynamic pattern of assimilation and contrast in any context. By doing so, it highlights an important issue within the social comparison literature and beyond, where moderating variables such as the standard extremity are routinely abstractly defined and arbitrarily dichotomised (e.g., Warm vs. Cold, Steinmetz & Mussweiler, 2011; or Certain vs. Uncertain, Pelham & Wachsmuth, 1995). Often this is done with little guidelines regarding the appropriate way to operationalise these new variables. This vagueness leaves ample room for corroborating findings to be presented as evidence for their efficacy, while unexpected results are attributed to issues of operationalisation, leaving real tests of the moderating variables, let alone their underlying theories, untenable. Although more onus to clarify and guide future research should be placed on the initial claims of moderation and theories of comparison as a whole, for now researchers may mitigate the issue somewhat by refraining from treating continuous variables in a dichotomous fashion. This is not to say that every investigation must sample the entire range of the moderator as was attempted here, or even

fully move away from simplifying terms like moderate or extreme, but rather that they must avoid only including single comparisons or very narrow bands of the moderating variable. For instance, one might choose to define moderate as between $1SD$ and $2SD$ and sample standards equally within this range. This will not only improve generalisability beyond the exact standard extremity used, but will simultaneously bindingly define the boundaries within which researchers expect a certain outcome, making findings more easily verifiable by others.

The importance of bindingly defining moderating variables like this should be apparent from the detailed depiction of the dynamic pattern presented. In the reported studies, the interaction of the opposing comparison effects lead to the apparent overall suppression of both around the turning point at roughly $2.5SD$ to $3SD$. Measuring comparison effects in this bandwidth would likely produce the conclusion that no comparison occurred, as the assimilative and contrastive judgments on average cancel out each other. Furthermore, the comparative patterns and related turning points might shift with the experimental context, or under manipulation of other moderators such as category membership (Brewer & Weber, 1994) or similarity focus (Mussweiler, 2001a). For instance, rather than assuming a similarity focus would invariably produce assimilation, it should rather be considered as also simply affecting the likelihood of assimilative effects to arise, thus, increasing the window of assimilation and shifting the turning point. This dynamic could lead to contrast effect from a standard under a dissimilarity focus, but an apparent null effect under a similarity focus as the turning point where both effects suppress each other has now shifted upwards to the position of the presented standards' extremity. To emphasise this possibility, we point to a recent meta-analysis which exactly reported this pattern (Gerber, Wheeler & Suls, 2017), describing no evidence overall for the assimilative effect under a similarity focus, but contrastive patterns for a dissimilarity focus.

This lead to the conclusion that a similarity focus does not produce assimilation in the strictest sense. A true test of the moderating role of a similarity focus, however, would necessarily need to sample standards from a wide range of extremities for this conclusion to be valid. In this way, the CJT paradigm itself may also provide an additional tool to test similar theoretical predictions not strictly related to the extremity variable, as long as they are expected to alter the underlying comparative dynamics of assimilation and contrast in some manner.

Another important observation throughout our studies was the small effect sizes, which could partially explain why previous research has occasionally struggled to find consistent effects. As mentioned, this may be the result of the dynamic process itself, in which, on aggregate, the opposing effects cancel each other out to some degree. However, although the current paradigm offers substantial control over the information that is presented to the participants, it does so to a minimal degree in the form of facial features of standards and targets, of which one has no further knowledge. This may pose a further limitation on these studies. Prominent models like the SAM (Mussweiler, 2003) have emphasized the role of the activation of standard-consistent knowledge and the resulting confirmatory processing of information about the target. Arguably, the limited knowledge and information available from faces might result in minimal comparison effects.

Although our paradigm demonstrates the fundamental nature of comparative judgments in these most minimal conditions, and provides the most unadulterated patterns of the comparison dynamics possible, it may well underestimate the size of comparison effects in other contexts. Indeed, the bulk of comparison research has focused on self-related judgments, where one has an incomparably intimate knowledge of the target (the self). At the same time they also present well known individuals or detailed descriptions of standards (Mussweiler, Rüter, & Epstude, 2004a),

which provide ample additional information that can be drawn from¹⁰. Expanded versions of such paradigms may, therefore, produce stronger assimilative and contrastive effects, which could result in different levels of extremity being optimal for producing the respective outcomes. Future research investigating domains other than facial evaluations, as well as judgments related to self-evaluations, will thus be needed to test the universality of the uncovered patterns, and provide guidelines in other comparative contexts.

Nevertheless, the current work extends the theoretical prediction of assimilation to moderate, and contrast from extreme standards, to the common everyday judgments of facial evaluations, while providing the most detailed picture of the dynamic interaction of these two opposing effects to date. It not only provides evidence for the fundamental role comparative processing can play even in minimal contexts, but offers the first guidelines concerning the operationalisation of moderate or extreme standards. In doing so, it highlights some widespread shortcomings in the field of comparison research, while offering a new general paradigm as a tool to more robustly test theoretical predictions.

Supplemental materials

All data files, analyses and additional materials can be found at:

https://osf.io/6zt5w/?view_only=98eff44baf0f468c9275fead26abee28

¹⁰ Problematically, this information is likely to additionally contain many other theoretically relevant moderating variables, which can only be accounted for if a wide representative sample of similar standards is included.

Chapter 4: Shifting Standards of Comparison¹¹

Abstract

Previous investigations widely treat dichotomous moderating variables as invariably leading to either assimilation or contrast without accounting for the inherent moderating role of standard extremity which is present in any comparative judgment limiting findings to the exact standards used. The current work argues that such moderating variables are better understood as simply affecting the likelihood of assimilation or contrast to occur, thereby shifting the inherent dynamic interactive pattern of assimilation and contrast by changing the equilibrium between the two outcomes at each point in the spectrum of extremity values. Across two studies ($N = 1905$) this shift is shown on various metrics by modelling the comparative pattern under a similarity versus a dissimilarity focus. Results confirm the predicted influence of this fundamental moderator of comparative tendencies and highlight the need for broader standard selection in comparative research of moderating variables.

¹¹ This chapter includes studies that are part of an ongoing project; Barker, P., & Imhoff, R. (2020) *Moving the line: Experimental shifts in the dynamic interactive pattern of assimilation and contrast for facial stimuli*. Manuscript in Preparation. Therefore, the related data and materials are not yet publicly available online, but will be released in full at a later date on the authors OSF page. Please do not copy or cite any of this work without the author's permission.

The previous chapters have highlighted the problematic practice of using single judgment dimensions and standards in comparison research, while uncovering the dynamic interactive nature of assimilation and contrast effects. Existing research has often overlooked this relationship between opposing effects, leading them to treat moderating variables such as different comparative foci as either invariably leading to assimilation or to contrast based on single or narrow bands of comparison standards. The current chapter will show that this oversimplified conceptualisation of the comparison process, which does not take the standards relative standing into account, can lead to mistaken claims of the absence of effects or conclusions regarding which comparative outcome is more dominant. In doing so, it will offer an alternative approach to the investigation of theoretical predictions in a way which does acknowledge the fundamental role that the relative standing of the standard plays in the comparison process.

Whenever we encounter someone new, we make almost instant judgments of them on a large number of dimension based on very little information. For instance, stereotypical notions can create deeply ingrained ideas of what other should look like (Imhoff, Woelki, Hanke, & Dotsch, 2013), with judgments made on a variety of traits based solely their facial features (Todorov, Olivola, Dotsch & Mende-Siedlecki, 2015) within the first 100ms of an encounter (Willis & Todorov, 2006; Ballew & Todorov, 2007). However, social judgments are often not made in isolation, but are based at least partially on some standard present in the environment or one which is internally held (Kahneman & Miller, 1986; Mussweiler, 2003). Recent research has shown that this is no different even for these split second facial evaluation (Barker & Imhoff, 2020) emphasising the fundamentally comparative nature of human judgment (Dunning & Hayes, 1996; Festinger, 1954).

How exactly these comparisons influence our final judgments has been widely researched and can depend on a large number of factors, but in broad terms the integration of comparison information can logically only affect the final judgment in one of two ways; either the judgment is moved closer towards the standard (i.e. an assimilation effect) or it is pushed away from the standard (i.e. a contrast effect). One of the most prominent models describing the process that will determine which of these outcomes will occur is the Selective Accessibility Model (SAM; Mussweiler, 2003). In this model, the most fundamental determinant of which comparative outcome will occur is an initial holistic assessment of the similarity or dissimilarity between the target and the standard. When the result of this assessment is that the two seem similar, information that is in line with this hypothesis will become more accessible and more readily used in the final judgment. This process biases the final judgment of the target to assimilate towards the standard. If, on the other hand, one sees the two as dissimilar, knowledge consistent with this hypothesis of dissimilarities will be more influential leading to contrast effects.

Although, many variables can influence the initial assessment of similarity, altering the direction of the final comparative outcome (e.g., category membership, Mussweiler & Bodenhausen, 2002; or psychological closeness, Brown et al., 1992), the most direct manipulation might be to simply inducing a stronger focus on similarities versus differences outright. Changes in the comparative focus from one of similarities to differences has been found to influence anything from self- (Mussweiler, 2001a) and social judgments (Corcoran, Hundhammer, & Mussweiler, 2009), to affective reactions (Epstude & Mussweiler, 2009), and evaluative pairings (Corneille, Yzerbyt, Pleyers, & Mussweiler, 2009). This influence is not only an integral part of the SAM, but has also been acknowledged as a particularly powerful

moderator by rivalling theories of comparative judgments (e.g., the Inclusion/Exclusion model; Schwarz & Bless 2007).

Comparative foci, therefore, seem to be an important aspect of the way in which humans make comparative judgments, with the SAM proposing that the most natural state is one in which similarities are sought and assimilation is the result (Mussweiler, 2003; Mussweiler & Epstude, 2009). This is thought to be the case due to the fact that any comparative judgment must logically start with a search for similar structurally alignable features before these features can be compared (Markman & Gentner, 1993; Gentner & Markman, 1994). In line with this idea, empirical work has shown the distinct processing advantage of focusing on similarities which might explain why this bias exists (Corcoran, Epstude, Damisch, & Mussweiler, 2011). However, in stark contrast to these claims, a recent meta-analysis investigating the effect of these different comparative foci reported a stronger overall tendency for contrast to occur (Gerber, Wheeler & Suls, 2018). Even more surprisingly, it only found limited evidence for assimilation even under a similarity focus, while contrast was clearly present under a dissimilarity focus.

4.0.1. The dynamic interactive pattern of assimilation and contrast

These meta-analytical findings call into question the robustness of the influence that a similarity focus has on comparison outcomes. However, one must consider that even though the comparison effects themselves are described as being binary in their outcome for a single judgment, aggregating over many judgments will often include both types of comparative outcomes to some degree due to ever present variability in the initial state of the comparison process. This means that most moderating variables affecting comparison outcomes should be considered as increasing the likelihood of assimilation or contrast to occur, rather than invariably leading to one or the other. Even if one would posit the exceptionally strong claim that a

moderator, such as a pure focus on similarities, must always lead to assimilation, researchers are often still bound to some instrumental manipulation of this variable since many psychological variables cannot be directly changed. These manipulations, though intended to manipulate the moderating variable as effectively as possible, are themselves merely tangentially connected to the true moderator and do not have perfect construct validity (Cronbach & Meehl, 1955; Campbell, 1957; Cook & Campbell, 1979). Therefore, these manipulations will affect the variable in the intended manner imperfectly and only for a finite number of cases. Even these strong moderators, or rather the instrumental variables used to affect them, are in practice often only able to affect the likelihood for assimilation or contrast to occur to a limited degree. As a result, which outcome will dominate on aggregate will almost always also depend on other moderating variables and the context in which the judgment is made, which includes the judgment itself (Barker, Dotsch & Imhoff, 2020; Barker & Imhoff, 2020).

With this in mind, a meta-analytic finding of contrast on an aggregated level, or a non-significant result under a similarity focus, might not reflect the true state of the comparison process across situations. Instead, it could merely be an indication of the most prevalent research contexts from which the data were taken. Indeed, studies that have investigated the influence of a similarity focus have often not sufficiently standardised, nor reported, an unavoidable variable inherent to the comparative information presented, namely the relative distance, or perceived extremity, of the comparison standard on the judgment dimension (e.g., Mussweiler, 2001a)

In general, extreme standards have been found to increase the prevalence of contrast, while moderate standards lead to more assimilation (Herr, 1986). However, recent work urges the variable to more appropriately be conceptualised as a spectrum of extremity values (Barker & Imhoff, 2020). This conceptualisation entails that the two opposing comparative effects of

assimilation and contrast suppress each other to a varying degree on aggregate, dependent on the relative distance of the comparison standard in relation to the target. The resulting dynamic interactive patterns is roughly described by a lying S-shape as depicted in Figure 4.1. The propensity for assimilation to occur, all else being equal, will vary widely based on the standard extremity that is presented, but should be more likely than contrast for the more moderate standards that fall within the Window of Assimilation (WoA) bounded by the red lines at the intersections with the x-axis. The strongest effect of assimilation on average is not at the most extreme ‘moderate’ standard, but can be seen at the Points of Maximal Assimilation (PMA) roughly in the middle of the WoA. The blue Area of Assimilation (AoA) represents the cumulative strength of assimilation across the range of values in which it is the dominant force.

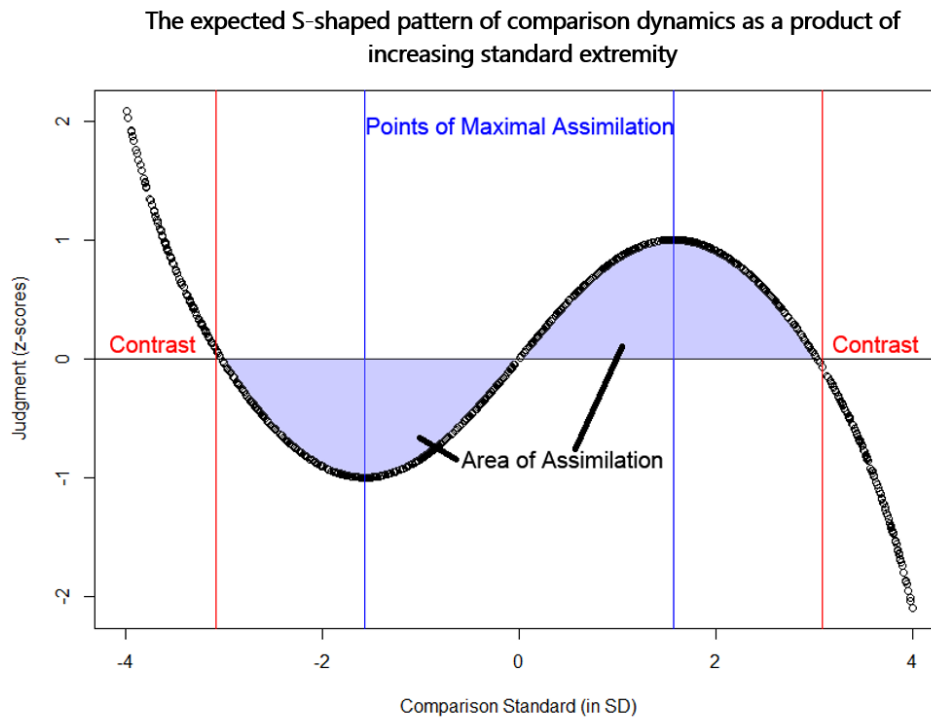


Figure 4.1. The expected S-shaped pattern of comparison dynamics as a product of increasing standard extremity. The x-axis represents the judgment that would be made with a completely neutral standard (0SD)

In light of this complex dynamic underlying the moderating role of standard extremity that is inherently present in the comparative information being considered, it should be clear that one cannot investigate comparative effects in absence of its influence. Moreover, the exact shape of the interactive pattern across the extremity spectrum itself is not static, but inextricably linked to the context in which the judgments are made. Any variable that alters the propensity for assimilation or contrast to occur will also affect the equilibrium between both comparative tendencies at each point in the extremity spectrum. This compounds the challenge of comparing comparative outcomes across different contexts even when they are in response to similar standard extremities. This was reflected clearly in the wide variation of comparative patterns that occurred when using a narrow selection of identical standard extremities across multiple judgments in Barker et al. (2020). Results simultaneously showed only assimilation, only contrast, and a combination of the two depending on the judgment being made. These effects were later found to be fundamentally consistent with theoretical predictions when fluctuations in the underlying dynamic interactive pattern of comparative outcomes were fully taken into account across a wider range of standards and multiple judgments (Barker & Imhoff, 2020).

Although these investigations focused specifically on the effects that different judgment dimensions can have on the comparative pattern, similar dynamics should occur in response to many other moderating variables. The exact influence a moderating variable will have on the underlying dynamics depends on the specific way in which it influences both comparative effects respectively. For instance, some variables might increase the strength and prevalence of both comparison outcomes, such as inducing a more comparative mind-set overall (Mussweiler & Epstude, 2009). This type of effect would not be expected to influence the WoA, but should increase the AoA by strengthening both comparative outcomes equally. However, other

moderating variables are expected to only increase the propensity for one of the two outcomes to occur, leading to a unique shift in the equilibrium of assimilative and contrastive tendencies. For instance, a focus on differences rather than similarities should increase the prevalence of contrast effects at most points on the spectrum. Thereby, this variable will shallow the curve for more moderate standards, decrease the AoA, and narrow the WoA all at once.

Note that this need not entail an absence of assimilation at every point on the scale, nor would inducing a similarity focus necessarily lead to assimilation at every point. In fact, an important characteristic of the interactive pattern is that the cumulative comparative effect is zero (i.e. assimilation and contrast cancel each other out) at the boundaries of the WoA (i.e. the points where the curve dissects the x-axis). This means that comparative judgments made in reaction to standards at this point on the extremity spectrum would likely lead to apparent null effects under either focus. However, the standard extremities that produce these null effects would not be the same for either mind-set as they shift along with the comparative pattern itself. Therefore, weak or absent assimilative effects in the literature under a similarity focus, while contrasts occur for a dissimilarity focus (Gerber et al., 2018), could legitimately be explained by the fact that the comparison standards regularly used in the literature happen to be at this turning point under a similarity focus, while they are firmly beyond this point for the dissimilarity focus.

Consequently, this proposed shift in the dynamic pattern of comparison outcomes under different comparative foci can only be detected when studies include a large range of standard extremities. Hence, in order to provide a strong test of the influence of a similarity focus and assess its robustness, a more holistic approach is necessary in which the inherent moderating role of standard extremity is sufficiently acknowledged.

4.0.2. The present research

The current work will be the first to attempt to provide such a strong test for the effect of a similarity focus, thereby uncovering the proposed shift in the dynamic interactive pattern of assimilation and contrast. In doing so, it will not only investigate the robustness of the effect of a similarity focus across a wide range of standards, but will also provide an example of how similar moderators can be investigated in the future.

Although the typical procedure for inducing different comparative foci has been to include a procedural priming task in which a focus is given to finding similarities or differences prior to the focal judgment (Mussweiler, 2001a), these subtle priming effects might not remain influential across the many trials that are necessary to model the comparative pattern accurately. Instead, the current work makes use of explicit instructions in every trial to focus on similarities or differences prior to making a judgment. Findings will therefore not be directly comparable to most previous investigations nor the recent meta-analytical findings (Gerber et al., 2018), which refer strictly to the efficacy of the priming procedure to elicit assimilation or contrast ostensibly mediated by a change in the comparative focus. Instead, the current results will only pertain to the influence of explicitly directing participants to change their focus on the frequency of assimilative and contrastive outcomes, which is similarly assumed to be mediated by a change in comparative focus. Notwithstanding this distinction, the same principles and issues that will be highlighted in the current work with regards to standard selection will apply to investigations using different manipulations of comparative focus, and potentially also different moderators all together.

The full comparative pattern under both a similarity and dissimilarity focus will be investigated using the CJT-41 paradigm (developed in Barker & Imhoff, 2020), capable of

modelling these patterns across a large range of standards for various facial dimensions. Within the current investigation, this will be done for the judgment dimension of facial Trustworthiness, which has shown both consistent assimilation and contrast effects previously (Barker & Imhoff, 2020). Although the exact results will, thus, also be limited to the facial domain and the Trustworthiness dimension specifically, this procedure offers a particularly suitable context in which the shift in comparison patterns should be easily detectable if present.

Based on the described literature, this procedure is expected to produce two distinct fitted curves, with the one produced under the similarity focus showing more signs of assimilation, a larger area of assimilation, and a wider window of assimilation, than the curve fitted for the dissimilarity condition. Inconclusive evidence was found for this pattern as a whole in an initial pre-registered study (Study 1), but was confirmed on all accounts in a later replication using an identical procedure on a different online platform (Study 2).

4.1. Study 1

To uncover the theorised shift in the dynamic interactive pattern of assimilation and contrast, brought on by a focus on differences rather than similarities, we will employ the CJT-41 to model this pattern under each focus separately in a pre-registered study¹². In general, we would expect a positive linear effect as a sign of assimilation with a negative cubic term indicating contrast to more extreme standards. Hence, we expect the focus condition to influence these two estimates in a way in which assimilative judgments are reduced under a dissimilarity focus compared to under a similarity focus. In the broadest terms this should be reflected in a significant increase in fit when including the condition variable and its interactions in the model.

¹² Pre-registration documentation can be found at <http://aspredicted.org/blind.php?x=693hq8>

To test the more specific expectations that the condition will reduce the overall strength of assimilation and lead to consistent contrast effects for more moderate standards, the area of assimilation and the window of assimilation will be estimated using a bootstrapping procedure. A smaller AoA and narrower WoA are expected in the dissimilarity compared to the similarity focus condition.

4.1.1. Methods

Participants. Simulations revealed that based on the parameters found for the Trustworthiness dimension in Barker & Imhoff (2020), 950 participants would be necessary to find a difference between the two conditions at least 80% of the time if the linear effect changes by $B = .018$. Therefore, a total of 950 US-based MTurk participants were recruited to complete the study for a reward of \$1.20. The final sample in this study was 51.4% female and was aged between 18 and 76 years ($M = 37.14$, $SD = 11.84$).

CJT-41. The CJT-41 developed in Barker & Imhoff (2020) was implemented. This paradigm required participants to make 41 judgements regarding the behaviour of a target person based on their face with a second face presented alongside them acting as a comparison standard. Participants were randomly assigned to respond to a single item out of six possible ones, which all required the absolute estimated frequency of a certain behaviour in 6 months related to the Trustworthiness dimension, see Table A2 in Appendix E. In each trial, participants first needed to identify the clearly marked judgment target from the two facial images as an attention check. The standard varied in each trials in regards to its extremity on the Trustworthiness dimension. One randomly assigned group of participants was presented with standards at each $0.2SD$ step between $-4SD$ to $+4SD$, while a second group was presented with standards between $-3.9SD$ to

3.9SD in 0.2SD intervals as well as the neutral 0SD standard. This division thus results in measurement instances at each 0.1SD interval between 4SD and -4SD and 81 extremity steps across participants.

Stimuli. To create the 81 facial pairs that were needed, a process similar to the one used in Todorov, et al. (2013) was implemented with a custom scripts built on the FaceGen SDK. This allowed for the generation of a large number of neutral facial IDs, which could then be precisely manipulated along the vector of facial Trustworthiness to create comparison standards at each of the required extremity steps, see Appendix C. All 81 neutral targets and related comparison standards were created in this manner.

The six items, three of which were reverse coded, were taken directly for Barker & Imhoff (2020) where they were pre-tested and selected for their relation to the facial dimension of Trustworthiness, see Table A2 in Appendix E for a list of these items.

Additional measures. Age, sex, and education were recorded as basic demographics. At the start of the survey, an item asked participants to describe what was presented on an image depicting shapes and colours, to ensure images were displayed correctly in the majority of cases. In addition, an item after the main task required participants to indicate whether they were asked to focus on similarities or differences throughout the trials, which acted as a simple manipulation check and exclusion criterion if answered incorrectly. Finally, a self-reported data-quality item at the end of the study asked participants if their data should or should not be used based on their effort throughout the task. Participants were told their response to this item would not affect them, but would help researchers to clean up the data in order to analyse the results. Responses

ranged from “Definitely do not use my data” (1) to “Definitely use my data” (4). Participants with a response of 2 or lower were excluded from the analyses.

Procedure. Participants were recruited using the online platform MTurk and fully informed about the data storage policy, procedure, and their rights before they were asked for their consent to take part in the study. Participants first responded to the image identification question, followed by the basic demographics. Participants were then randomly allocated to a similarity condition, where they were instructed to look at both images and focus on similarities or differences before making their final judgment. The CJT-41 was then explained in detail and two practice trials were presented before the main batch of 41 trials started. Throughout the trials participants were reminded in the middle of each page to focus on similarities or differences respectively. Upon completion of the main task, participants were presented with the manipulation check question followed by the data-quality item. Finally they were debriefed, thanked and given their compensation.

Data treatment. Initially, all trials with non-numeric or empty responses were removed (4.3%) as well as any responses marked by a failed attention check (6.7%). The remaining responses were z-transformed per participant where possible to ensure similar response scales across items and respondents. The resulting scores were then truncated above 3 or below -3 (3.8%) and flipped for reverse coded items so that higher scores reflect more Trustworthy behaviours. Combined these criteria resulted in 14.7% of all trials being removed. Lastly, a large group of 162 participants failed to correctly identify the condition they were assigned to, while 41 indicated their data should not be used. These pre-registered criteria combined lead to the exclusion of 216 participants, which was a far larger amount than expected. The final sample that could be used in the analyses, thus, only included 734 participants.

4.1.2. Results

The pre-registered analyses consisted of running two mixed models regressions using maximum likelihood (ML) estimation with the lme4 package (Bates, Maechler, Bolker, & Walker, 2015) to allow for model comparison in R (Version 3.5.1; R Core Team, 2018). Both models included linear and cubic fixed effects for the extremity steps, and had a similar structure of orthogonal uncorrelated random slopes and intercepts for participants and items¹³. The full model additionally included fixed effects for the condition variable and all interactions, estimating the marginal effect of the focus condition on the shape of the curve. Finally, an analysis of deviance was used to compare the full model to the model without any effect of condition on the comparison pattern. The results showed that, while the S-shaped curve in the difference focus condition was descriptively more shallow in line with the expectations (Figure 4.2), this difference failed to reach the standard level of significance in these data, $\chi^2(3) = 7.17, p = .067$, providing no convincing evidence for the moderating influence of the comparative foci on the comparison pattern overall.

Although the focal test failed to reach significance, separate models per condition were fit using restricted maximum likelihoods (REML) to gain some indication of their separate comparative patterns and estimate the differences in area of assimilation and the window of assimilation¹⁴. These analyses showed that in the similarity focus condition, both the linear and

¹³ This random effects structure is more extensive than the one described in the pre-registration. Upon reviewing the data, it was found that this more extensive random effects structure was able to be modeled, which is preferable since it allows for broader generalizability of the results. The simplified structure from the preregistered analyses only increase the strength of the uncovered associations.

¹⁴ Satterthwaite's approximations were used to determine the appropriate degrees of freedom in order to calculate confidence intervals and p-values for the fixed effects (using the lmerTest package; Kuznetsova et al., 2017; and the parameters package in R; Lüdtke et al., 2020).

cubic effects were significant and in the direction expected for a pattern of assimilation to moderate and contrast from extreme standards (Table 4.1). However, in the difference focus condition, the positive linear effect was not detected, but the negative cubic effect was still present (Table 4.2) providing only evidence for contrast under a dissimilarity focus.

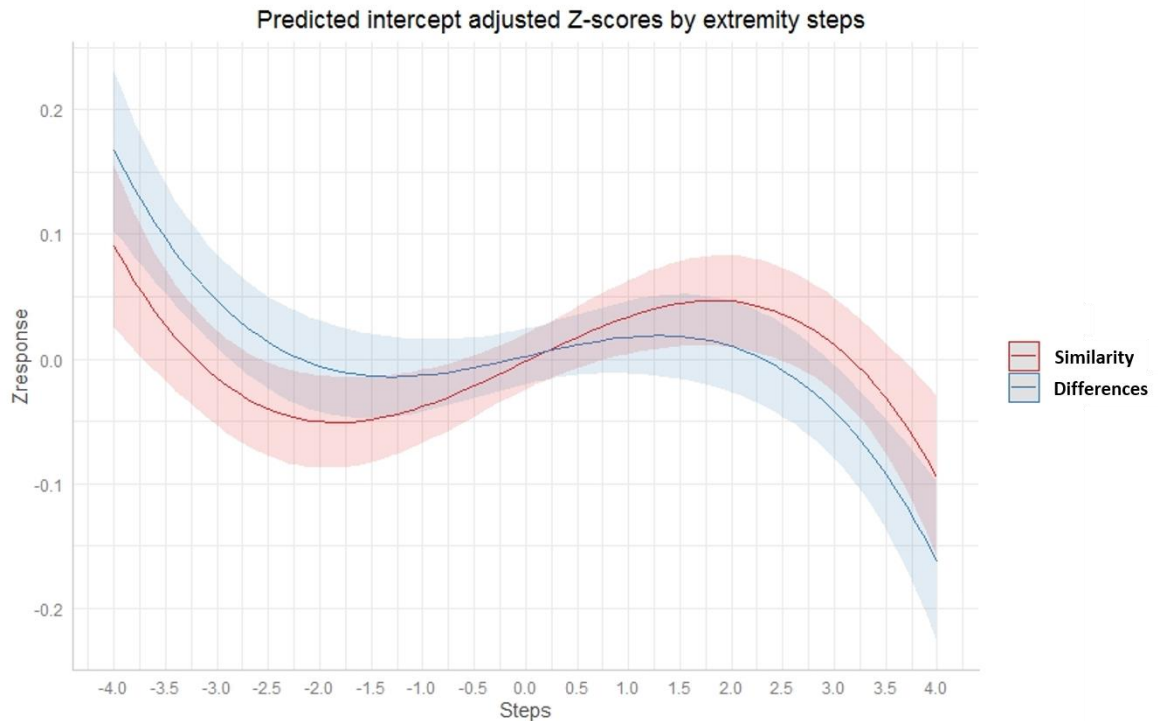


Figure 4.2. Predicted intercept adjusted marginal z-scores for the different focus conditions at each extremity step and their predicted 95%CI (created with the ggeffects package, Lüdtke, 2018)

To estimate the descriptive statistics of the area of assimilation and windows of assimilation for each condition, as well as their respective non-parametric percentile confidence intervals, a bootstrapping procedure (utilising the boot package; Canty & Ripley, 2019) with 3000 iterations was implemented, using a simplified model with maximum likelihood estimation and only random intercepts and linear slopes for participants and items. In each iteration the estimated curve was solved for zero to give the turning points where possible (using the polynom

package; Venables, Hornik & Maechler, 2016). For curves that only showed signs of assimilation the maximum measured value of $4SD$ was used. If only contrast effects were present, the value was kept at the minimum of $0SD$. To calculate the AoA, the definite integral between zero and the upward turning point was returned. The WoA was found by subtracting the downward turning point from the upward one. Results showed that the AoA, the cumulative strength of assimilation, was larger under the similarity focus condition, $0.10SD^2$, $SE = 0.02$, $Bias < -0.001$, 95% CI [0.06, 0.15], than the dissimilarity condition, $0.03SD^2$, $SE = 0.02$, $Bias = 0.002$, 95% CI [0.00, 0.06], but with percentile confidence intervals which overlapped in agreement with the main analysis. The WoA, indicating the range in which assimilation is more likely than contrast, was also found to be larger under a similarity focus, $6.35SD$, $SE = 0.28$, $Bias = -0.02$, 95% CI [5.77, 6.90], than a dissimilarity focus, $4.52SD$, $SE = 0.82$, $Bias = -0.18$, 95% CI [2.16, 5.33], with percentile confidence intervals that this time did not overlap.

Table 4.1.

All fixed effects and related statistics from mixed model analysis for similarity focus.

	<i>B</i> [95% CI]	<i>SE</i>	<i>df</i>	<i>t</i>	<i>p</i>
Fixed effects:					
x	0.0401 [0.0189; 0.0613]	0.0108	7.23	3.71	.007
x ³	-0.0040 [-0.0059; -0.0021]	0.0010	4.39	-4.10	.012
Intercept	0.0039 [-0.0177; 0.0000]	0.0110	4.80	0.36	.737

Table 4.2.

All fixed effects and related statistics from mixed model analysis dissimilarity focus.

	<i>B</i> [95% CI]	<i>SE</i>	<i>df</i>	<i>t</i>	<i>p</i>
Fixed effects:					
x	0.0191 [-0.0022; 0.0404]	0.0109	7.70	1.76	.118
x ³	-0.0038 [-0.0057; -0.0019]	0.0010	4.69	-3.86	.013
Intercept	0.0084 [-0.0137; 0.0305]	0.0113	5.22	0.74	.489

4.1.3. Discussion

In this first investigation of the effect of different comparative foci on the dynamic interactive pattern of assimilation and contrast, the hypothesised effect did not reach the standard level of significance providing insufficient evidence to determine that there was an influence. However, the descriptive pattern and follow up bootstrapping tests did yield some evidence that the condition might have affected the window of assimilation specifically. Considering these conflicting data points a replication attempt was performed to attempt to provide more conclusive findings.

One may take note of the relatively low data quality that was achieved from the online platform, with 22.7% of respondents meeting the pre-registered exclusion criteria severely reducing the power to detect the expected effect. Therefore, the next study will be a close replication, but using the alternative platform Prolific to recruit respondents which has been reported as having less dishonest and more diverse users, while still retaining data quality (Peer, Brandimarte, Samat, & Acquisti, 2017).

4.2. Study 2

This study will be a close replication of the previous one, with only the online platform used for recruiting participants being changed from MTurk to Prolific, which has been found to have a more diverse and honest population with comparable quality (Peer et al., 2017). As no changes were made to the general method or protocol, these sections will not be discussed here. Similarly the hypotheses, expectations and analyses for this study remain the same as in the previous study.

4.2.1. Method

Participants. A sample of 950 participants was this time sought on the Prolific platform for a slightly higher reward of £1.50 in an attempt to increase data quality. Five additional participant completed the study, but were not tallied by the platform resulting in 955 finished responses in total. This final sample consisted of 59.6 % females and ranged from 18 and 82 years of age ($M = 36.68$, $SD = 12.06$).

Data treatment. As in the previous study, non-numeric or empty responses were removed (3.6%) along with trials in which the attention check was failed (2.1%). Z-scores per participant were then created using the remaining responses where possible and truncated above 3 or below -3 (4.9%), with scores flipped for reverse coded items. All together these criteria resulted in 10.7% of trials being excluded from the analyses. Sixty eight participants did not correctly identify the condition they were assigned to and were removed, as well as 25 participants with low self-reported data quality. In total, all criteria combined called for the removal of 132 participants, resulting in a final sample of 823 to be used in the analyses. One might take note of the improved data quality using this alternative platform, with 13.8% of participants not providing usable data compared to the 22.7% in the previous study.

4.2.2. Results

An analyses identical to the one used in Study 1 was performed on these new data. Results showed that the similarity focus condition this time was shown to have a significant influence on the comparison pattern, $\chi^2(3) = 16.75$, $p < .001$, with the resulting curves descriptively again agreeing with the expectations that a focus on dissimilarities would lead to a more shallow curve, and seemingly narrower window of assimilation than a focus on differences (Figure 4.3).

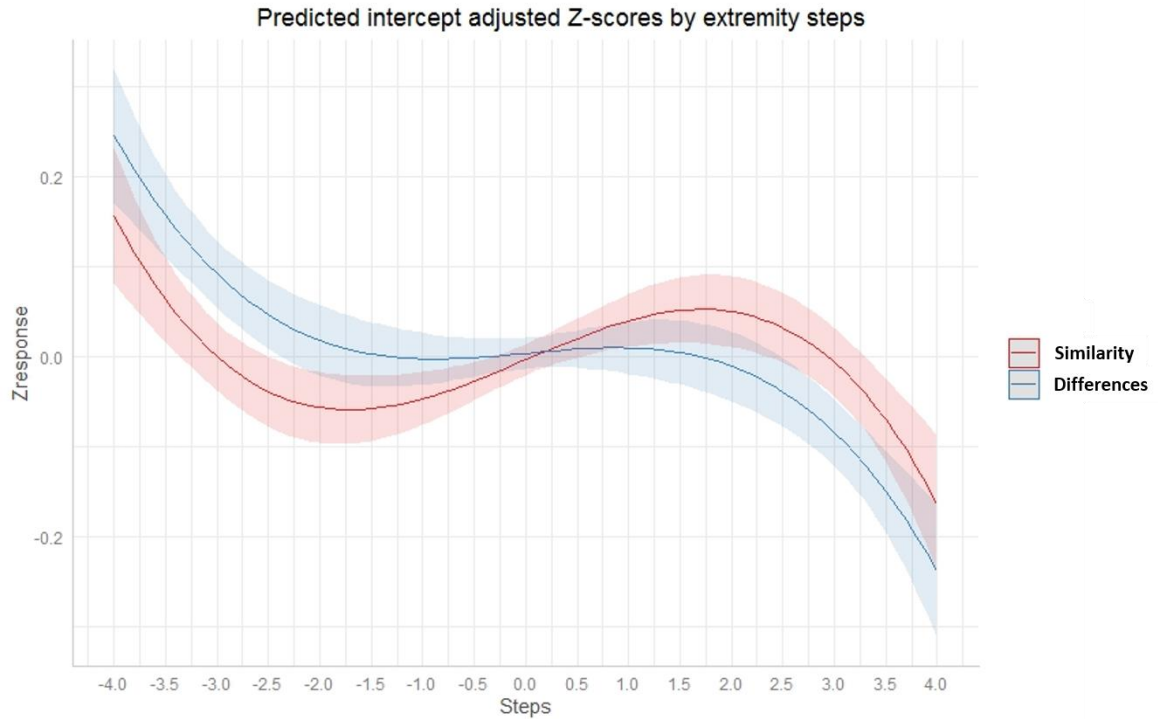


Figure 4.3. Predicted intercept adjusted marginal z-scores for the different focus conditions at each extremity step and their predicted 95% CI

Separate mixed models analyses using REML were then conducted to gain an indication of the pattern in both conditions separately. Results showed the positive linear and negative cubic effects to be significant in the similarity focus condition (Table 4.3), indicating both assimilation and contrast had likely occurred. On the contrary, the difference focus condition again did not show a significant linear effect, but did show a significant negative cubic effect (Table 4.4), thus only providing clear evidence for contrast effects for increasingly extreme standards.

The bootstrapping procedure with 3000 iterations showed that the AoA was larger under the similarity focus, $0.11SD^2$, $SE = 0.02$, $Bias < -0.001$, 95% CI [0.07, 0.15], than under the dissimilarity focus condition, $0.01SD^2$, $SE = 0.01$, $Bias = 0.002$, 95% CI [0, 0.03], with percentile

confidence intervals that did not overlap. These results agree with the main analyses as well as the expectation that a dissimilarity focus would reduce the strength of assimilation overall compared to a similarity focus. In agreement, the window of assimilation was also found to be wider in the similarity focus condition, $5.94SD$, $SE = 0.21$, $Bias = -0.01$, 95% CI [5.48, 6.31], than in the dissimilarity focus condition, $3.07SD$, $SE = 1.14$, $Bias = -0.29$, 95% CI [0, 4.24], with percentile confidence intervals that once again did not overlap. Taken together these results provide convincing evidence that a difference in similarity focus affects both the strength of the aggregated assimilative effects as well as shifting the window in which assimilation is the dominant tendency.

Table 4.3.

All fixed effects and related statistics from mixed model analysis for similarity focus.

	<i>B</i> [95% CI]	<i>SE</i>	<i>df</i>	<i>t</i>	<i>p</i>
Fixed effects:					
x	0.0492 [0.0215; 0.0769]	0.0141	7.68	3.48	.009
x ³	-0.0056 [-0.0081; -0.0030]	0.0013	5.37	-4.32	.006
Intercept	0.0023 [-0.0132; 0.0177]	0.0079	4.84	0.29	.787

Table 4.4.

All fixed effects and related statistics from mixed model analysis for dissimilarity focus.

	<i>B</i> [95% CI]	<i>SE</i>	<i>df</i>	<i>t</i>	<i>p</i>
Fixed effects:					
x	0.0105 [-0.0143; 0.0353]	0.0127	6.62	0.83	.437
x ³	-0.0045 [-0.0068; -0.0021]	0.0012	4.91	-3.73	.014
Intercept	0.0097 [-0.0059; 0.0253]	0.0080	5.18	1.22	.276

4.2.3. Discussion

This replication study produced results that were largely in line with the previous study, but this time the focal test did reveal a clear and significant effect in the expected direction.

Separate analyses showed that the positive linear effect, a sign of assimilation, was only found in the similarity focus condition, but did not appear in the dissimilarity focus condition. Visual inspection of the resulting patterns also clearly showed more pronounced assimilative effect in the former than the latter condition. Lastly, the bootstrapping tests affirmed the previously found difference in the windows of assimilation, with a larger WoA in the similarity focus condition than the dissimilarity focus condition. However, this time, it also showed a more convincing difference between the areas of assimilation.

The results thus show a clear shift in the comparative pattern when focusing on similarities versus differences in line with the theoretical mechanism underlying the SAM. More importantly, the results highlight the importance of standard selection for the investigation of moderating effects. There is no simple flip in the pattern from only assimilative effects to only contrastive effects, rather a shift in focus from similarities to dissimilarities corresponds to a shift in the window in which assimilation is likely to occur, while also reducing the aggregated strength of assimilation in this same window. However, even under a dissimilarity focus contrast effects do not dominate in comparisons with moderate standards in this context. In fact, the estimated windows of assimilation in these data indicate that one would expect to find consistent assimilation under a similarity focus, while simultaneously finding contrast under a dissimilarity focus, only for a small area between $1.5SD$ and $3SD$. Thus, only in this range would one expect to find a clear cross over effect when evaluating the moderator.

4.3. General discussion

Previous work has often treated moderating variables as either leading to assimilation or contrast invariably. However, the current work shows that these variables might often be better conceptualised as affecting the equilibrium between the two comparative tendencies, which

naturally vary to a differing degree as the relative distance to the target increases from more moderate to more extreme standards. As a result, the exact standard that one uses, can profoundly affect the perceived influence of these moderating variables and, thereby, the inferences that are drawn. Across two studies, the current work provides a more nuanced picture of the moderating role of comparative foci by uncovering a shift in the dynamic interactive pattern of assimilation and contrast modelled across a large range of extremity steps for facial evaluations of trustworthiness. As Study 1 was somewhat inconclusive regarding the effect of the focus condition on the curvilinear pattern, a replication in Study 2 was conducted to provide clearer evidence of this shift for both the overall pattern, as well as for the area of assimilation and the window of assimilation specifically.

In addition to the fact that these studies reaffirm the robustness of the influence of focusing on similarities versus differences, they also clearly show that this effect does not represent a simple flip from assimilation to contrast. Rather the pattern shows a suppression of assimilative tendencies and narrowing of the window of assimilation under a dissimilarity focus compared to the similarity focus. However, even under a dissimilarity focus assimilative effects do still occur for more moderate standards, albeit with significantly reduced aggregated strength. The current data cannot exclude the possibility that contrast effects remained the same at each step in the spectrum, while only the prevalence or strength of the assimilative effect were influenced or vice versa. However, this pattern is also in line with the theoretically predicated notion that the dissimilarity focus increases the number of contrastive judgments even for more moderate standards, but not to the extent that they became dominant over the assimilative tendencies. Again, this non-linearity is expected if a moderating variable influences the likelihood of one comparative outcome to occur over another rather than strictly excluding one of these tendencies

altogether, which would form an exceptionally strong claim for any variable that cannot be directly manipulated due to imperfect construct validity of the manipulation (Cronbach & Meehl, 1955; Campbell, 1957; Cook & Campbell, 1979).

These findings have a number of far reaching consequences when one considers the large body of research in which only a single or narrow band of comparison standards are selected. The most important one is the revelation that wide bands representing the standards' relative distance would not be expected to produce contrast for a dissimilarity focus ($0SD$ to $1.5SD$) or assimilation under a similarity focus (anything $3SD$ or above) in the current context. Consequently, a full moderation, in which assimilation is found for a similarity focus and contrast is found for a dissimilarity focus, would only be expected to occur for standards roughly in the range of $1.5SD$ to $3SD$ from the target. This fact is not as problematic if one only considers the foci in relation to one another, i.e. outcomes for a similarity versus dissimilarity focus for the same standard, although one may note that the size of the relative difference need not be equal at all steps. However, when considering one focus in isolation or in comparison to a control condition, one may detect no noticeable comparative effects if the standards used are simply representative of the point in the extremity spectrum where assimilation and contrast effects cancel each other out. This can lead to erroneous conclusions even on an aggregated level, and the potential appearance of no or limited signs of assimilation under a similarity focus (Gerber et al., 2018).

It should again be noted here that the current work has only evaluated the manipulation of a similarity focus through clear direct instructions to participants, and the results thus say little about the efficacy of the procedural priming procedures more commonly used to induce the separate foci (e.g., Mussweiler, 2001a). Whether the patterns found here extend to this more

subtle procedure remains to be seen, but the principle issues raised with regard to the narrow selection of comparison standards certainly also apply to these studies, and studies investigating wholly different moderators as well. Broader still, these issues present significant hindrances to the study of comparative processes in any form, as well as the interpretation of aggregated effect found in meta-analyses as relative distances are rarely sampled broadly or even defined clearly.

Another notable finding is that the size of the window of assimilation under the similarity focus was highly similar to the one found for the trustworthiness dimension previously where no manipulation of focus occurred (Barker & Imhoff, 2020). This may indicate that the default focus is indeed one of similarities as suggested by Mussweiler (2003) and in contrast to the meta-analytical findings in Gerber et al. (2018), which found contrast effect to be dominant overall studies. However, this discrepancy could also be a result of the specific nature of the CJT-41 paradigm, such as the evaluative dimension under consideration, as the feelings of trust or distrust has been found to affect the comparative process (Posten & Mussweiler, 2013). Similarly, the simultaneous presentation of standards may itself induce more assimilation than contrast (Wedell, Parducci, & Geiselman, 1987). Although this shows that the current work should certainly not be used to make strong claims about which tendency is indeed dominant overall, it does highlight the fact that the aggregated strength of effects across existing studies is similarly not a sufficient basis for this claim. The dominance of contrast across the literature may simply be another indication that the standards used in these studies are those which are more likely to induce contrast effects. Without broader tests across a range of comparison standards as the norm in comparative research, this alternative explanation will remain ever present and can impede the investigations of similar research questions.

Notwithstanding its own limitations in its generalisability, the current work overall highlights the strength of the CJT-41 paradigm's set-up. By taking a more holistic approach to the measurement of comparative outcomes, it gains the ability to be used not just as a tool to model patterns of assimilation and contrast for various items and dimensions (Barker & Imhoff, 2020), but can also be used as a strong tests of theoretically relevant moderating variables. However, the CJT-41 paradigm is but one approach that can be taken to investigate comparison patterns in more holistic fashion. Any set-up which includes multiple comparison standards, with a relatively precise manipulation of their extremity information, should be capable of achieving similarly strong tests. For instance, many investigations may not need the same granularity that the CJT-41 provides, and could further reduce strain on participants by only including certain bands of extremity in which one expects the effect to occur. Adopting such an approach will help future research produce more nuanced inferences and improve the generalisability of their findings.

Furthermore, other set-ups may remedy some of the notable downsides to the current paradigm, not least of which are the large samples that are required to produce consistent patterns resulting from the small effect sizes that are produced. This is likely due to the limited informational content that the facial stimuli provide as noted in previous work (Barker & Imhoff, 2020). The paradigm, thus, represents a minimal scenario of comparative judgments, which might underestimate the size of comparison effects as they happen in other contexts. Including additional knowledge of the target in some unobtrusive way within the paradigm may increase the potential for the activation of standard-consistent knowledge and the strength of the comparative effects. This would also form a fruitful test of the selective accessibility mechanism

itself and capture how the balance of informational content may affect the comparative pattern itself.

Nonetheless, the current work has provided a clear picture of the dynamic interactive pattern of assimilation and contrast for facial evaluation under both a dissimilarity and similarity focus, uncovering a striking shift and shallowing of the comparison pattern when focusing on former compared to the latter. The results not only provide support for the robustness of the moderating role of different comparative foci, but they also show the effectiveness of the current approach to perform strong tests of theoretical predictions across a wide range of standard extremities. In doing so it has highlighted the problematic nature of using narrowly selected standards for the investigation of moderators of comparative judgments, while providing one way in which this issue could be remedied in future research that aims to take a more holistic approach to comparative research.

Chapter 5: General Discussion

This dissertation has aimed to highlight and address some fundamental shortcomings in the standard approach to comparative research, which hamper both the generalisability of findings and the creation of strong tests of theory. As an alternative, it has argued for a new more holistic way of investigating these effects which takes into account the dynamic interactive pattern of assimilation and contrast, and its sensitivity to item selection and other contextual factors. In support of this call for a new standard, it has provided tangible recommendations for future research regarding standard, dimension, and item selection; offered a first concrete definition of what an extreme standard might be; and has developed a paradigm that grants one way in which future comparative research might avoid some of these pitfalls.

These points were made clear across three chapters. Chapter 2 highlighted the sensitivity of the comparison process to the exact judgment under consideration using the newly developed CJT paradigm. Standards of similar extremity evoked highly heterogeneous comparison outcomes across five different judgments. More precisely, the judgment item related to the Trustworthiness dimension produced assimilation in response to moderate ($1SD$ from the target) standards, but no contrast from extreme ($4SD$) standards. The items related to the Competence and Dominance dimensions showed contrast from moderate standards, but extreme standards only induced contrast for Dominance. The item related to Extraversion was alone in showing the classically predicted interaction with assimilation to moderate and contrast from extreme standards. Finally the Likability item produced no comparative effects to either extremity level. The large heterogeneity found in this chapter clearly indicates that the use of multiple judgments and dimensions need to be employed in order to make generalisable claims about comparative effects, even when the relative distance of the standards are ostensibly equal in all situations.

Chapter 3 corroborated this sensitivity to the judgment being made and provided the first fine grained look at the proposed dynamic interactive pattern of assimilation and contrast. The uncovered patterns showed that, within the given context, standards greater than $3SD$ can be considered extreme enough to produce contrast, while assimilation is strongest around $1SD$ to $2SD$. However, the findings presented in this chapter also indicate that standards that fall in between what might be considered ‘moderate’ and ‘extreme’ can produce apparent null effects, and that where this exact point falls is again highly sensitivity to item level differences. Taken together, these findings emphasise the necessity for comparative research to select far broader ranges of standards, thereby refraining from the arbitrary dichotomisation of the relative standing of standards into abstract categories of ‘moderate’ or ‘extreme’. Instead, a more holistic approach should be considered in which the relative standing of the standard is conceptualised as a continuous moderating variable.

Finally, Chapter 4 showed how future research may take such a broad holistic look at comparative effects using a set up like the CJT-paradigm. In doing so, it revealed that, rather than leading to only assimilation or only contrast, moderators such as a comparative focus on similarities or dissimilarities cause a shift in the dynamic interactive pattern, affecting the equilibrium at each point in the extremity spectrum. This shift, as opposed to an absolute flip, means that, within the current context, wide stretches on the extremity spectrum seemingly do not produce consistent contrast under a dissimilarity focus ($0SD$ to $1.5SD$), nor assimilation under a similarity focus ($3SD$ or above). Hence, studies that select only narrow bands of standards from this spectrum would be expected to find the classical pattern, i.e. assimilation for a similarity and contrast for a dissimilarity focus, only in the stretch of roughly $1.5SD$ to $3SD$ from the target. This fact highlights the need to take the relative standing of the comparative

standard into account even when it is not the focal variable of interest. At minimum, future research should clearly define the relative standing of their standards in order to avoid erroneous conclusions, and to emphasise the limits of generalisability and potentially construct validity as well (Wells, & Windschitl, 1999).

In sum, the current dissertation has provided ample evidence to support the need for a more holistic approach in comparative research, while also providing new key insights into the pattern of comparative outcomes, and how they are affected by item selection and moderating variables. The following sections will consider the broader implications for comparative research and other fields, as well as offering some final recommendations. Limitations and fruitful lines for future research will then be discussed, ending in some final concluding thoughts.

5.1. Implications

The work presented in this dissertation has clear sweeping implications for the field of comparative research in general, and in particular for the investigation of comparative outcomes. Most generally it implies that a large part of the literature needs to be considered as reflecting the comparative process only within the exact context of the study, standards, dimensions, and items that were used. Until they are revisited and confirmed using a more holistic approach with broader representative item, and standard selection, such findings should not be assumed to be robust in different contexts. Similarly, this represents a call to future research to refrain from using these limited approaches to studying comparison processes.

In particular, previous findings related to the influence of standard extremity may be less robust and more sensitive to item selection than previously thought. A key issue with these studies, and many others in the field, is that they often rely solely on a single comparative judgment and a single comparison standard per condition, severely reducing their generalisability

and the construct validity of their manipulations (Wells & Windschitl, 1999). This issue has been made imminently clear in Chapter 2, where similar standards produced vastly different results dependent on the judgment item that were considered. Although the work presented in Chapter 3 later confirmed the theoretically expected comparative pattern across multiple items, this might not necessarily be the case for other effects in the literature which have been investigated using limited items, dimensions, and comparisons as was common in the past.

A similar point, more specific to comparative research itself, relates to the narrow selection of comparison standards to represent upward or downward, and moderate or extreme standards. Here the implications of the current work are most profound. By moving away from the abstract dichotomous categorisation of standards as moderate and extreme, a more nuanced interactive pattern emerged. In light of this, it becomes obvious that no comparative judgment is isolated from this influence of standard extremity as each standard affects the equilibrium of comparative effects to a unique degree. Comparative effects in the literature must, therefore, be treated as fundamentally linked to the precise relative standing of the standard that was presented. Future research should acknowledge this fact by at minimum reporting the relative standing of their standards, but would be advised to select a broad range of standards and include the relative standing within their models to be meaningfully compare them across studies. Similarly, this means that existing findings which have not accounted for this factor will be limited in their conclusions. Even when a large number and range of standards are used (e.g., Gerber, Wheeler, & Sulz, 2018), the failure to consider their relative standing could lead to weak or null effects on aggregate, as the assimilative effects produced by more moderate standards are counteracted by the contrastive influences of the more extreme ones. Consequently, it becomes clear that comparative outcomes can rarely if ever be investigated in isolation from the relative

standing of the standards used, and should rather be considered in the more holistic way proposed in this dissertation. This more nuanced approach is not just paramount from a theoretical standpoint, but will better predict the effect of real world comparisons that can affect important societal issues such as the wellbeing of medical patients (e.g., Stanton et al., 1999), body satisfaction and eating disorders (e.g., Wasilenko, Kulik & Wanic, 2007), or judgments of racism (O'Brien et al., 2010).

Although the current work has only demonstrated the influence of these design factors on comparative outcomes, other branches of comparative research might similarly benefit from adopting a more holistic approach. For instance, the extremity of standards has recently been suggested as being of influence in the comparative effects on self-esteem (Fardouly, Pinkus, & Vartanian, 2017) and motivation (Diel & Hoffmann, 2019). In light of these advances, a broader selection of standards will need to be considered in these areas as well, with explicit modelling of the relative standing if non-linear dynamics similar to the ones presented here are present. Even in absence of such complex dynamics, many, if not all comparison areas should report the relative standing of their standards to aid comparison between studies, generalisation and replication efforts.

Similarly, other moderating variables of a continuous nature that have been dichotomised in the past (e.g., Warm vs. Cold, Steinmetz & Mussweiler, 2011; or Fluent vs. Disfluent, Häfner & Schubert, 2008), may need to be re-evaluated using a similar fine grained approach to pinpoint their boundary conditions. Even outside of the comparative domain, these moderators may benefit from being investigated in a more continuous fashion in the future, such as the broad findings related to physical warmth and interpersonal closeness (IJzerman & Semin, 2009), or

fluency and judgments of truth (Reber & Schwarz, 1999). Indeed, this advice will hold in any situation in which a continuous variable can lead to opposing effects.

A similar point could be made for research outside the social real. For instance, there are numerous visual illusions that produce both assimilative and contrastive effects that could be investigated in a similar manner (e.g., the Delboeuf illusion and the size illusion of quadrilaterals in Goto et al., 2007). In a similar vein, contrast effects have been reported in the enumeration of dot patterns (Cordes, Goldstein & Heller, 2014), but as of yet no assimilation effect have been reported. Adjusting various parameters, and modelling them according to the principles laid out here, could yet produce assimilative patterns, or may lead to the conclusion that something in the nature of these types of judgments precludes assimilative effects from occurring. In both cases, this more holistic approach is likely to lead to theoretical advances in comparative theories.

In addition to these key implications for study design, the current work has also extended comparative processing into the domain of spontaneous facial evaluations of social traits, indicating the fundamental role that these processes play in even the briefest of everyday interactions with strangers. In this way, even a mere glance at another person can affect the way in which we see others in regard to dimensions such as Dominance, Trustworthiness and Extraversion, in addition to the well documented influence of attractiveness (e.g., Wedell, Parducci, & Geiselman, 1987; Thornton & Moore 1993). Even though the downstream consequences of these spontaneous facial comparisons were not investigated here, these findings come among growing concerns about the effect of social media use on wellbeing (for an overview see Verduyn, et al., 2017) and its links with social comparison (e.g., Vogel, Rose, Roberts, & Eckles, 2014; Fardouly et al., 2017). Notwithstanding the fact that the findings presented here are obviously too far removed from such a real world situation to legitimise any

conclusions regarding their effect in these situations, they do emphasise that even in lack of any knowledge about a target or standards, mere brief encounters with facial images can directly influence our judgments on complex social traits.

This fact could also be of influence in areas strictly outside of the normal comparative paradigm, such as choice preferences. For instance, recent work has shown that choices of facial attractiveness can be influenced by the extremity of a third distractor face through the process of divisive normalisation (Furl, 2016), where the representation of all stimuli intensities are normalised within the current informational ecology (Carandini & Heeger, 2012). Although speculative, it is not unlikely, considering the findings presented here, that complex social traits, such as those reflected in the CJT stimuli, would also be susceptible to divisive normalisation, and thus would behave similarly in such paradigms. If so, this would broaden the fundamental process of normalisation, which operates even at the neuronal level, to also affect abstract representations of key social traits. Therefore, it may also prove to be an important process underlying contrast effects in the comparison process itself. However, there are some caveats that must be added when considering these findings and their implications as will be discussed in the next section.

5.2. Limitations & Future Research

As noted in the introduction, a number of key design choices were made in order to more precisely control the relative distance of the standards, as well as to present information in an unobtrusive way. However, although the implications for study design and the interpretation of prior research broadly hold regardless, these choices do themselves limit the generalisability of the precise findings to the context in which they were found. For instance, the limited selection of items and domains, the choice of manipulations, the simultaneous presentation of images, the

focus on facial evaluations, and the use of open responses all limit the conclusions regarding the comparative outcomes to the context that they created. Indeed, rather than considering the presented patterns and estimates as a conclusion, for instance regarding what may constitute an extreme standard in all cases, the current work should only be seen as a first step towards forming more robust definitions and more precise inferences about such matters.

In a similar way, the findings here are limited by the populations in which they were found, and so may not reflect a universal process that holds in other populations. Although both online and lab samples were collected from different national populations (Germany, the US and the UK), all were from western backgrounds, a common issue in psychological research (Henrich, Heine & Norenzayan, 2010). It is quite possible that non-western samples may produce remarkably different comparative patterns. For instance, one may speculate on the effect that a more collectivist cultural upbringing may have on the comparative pattern, since it has been suggested to increase the general tendency to compare especially with upward standards (Chung & Mallery, 1999; Baldwin & Mussweiler, 2018), and that the related interdependent self-construal can result in less contrastive tendencies (Cheng & Lam, 2007). Therefore, future research should be encouraged to broaden samples to include non-western populations in order to identify the universality of these comparative processes.

To a lesser extent the maximum extremity that was used ($4SD$) also limits our findings to the standards inside this range. Although the population of standards that are excluded is an exceptionally small percentage of the total population, the current standards might still be considered as obtainable and relevant where others may not. Previous work has emphasised the need for standards to be diagnostic, with exceptionally extreme exemplars sometimes deemed as irrelevant (Festinger, 1954; Lockwood & Kunda, 1997; Tesser, 1988). This can lead them to no

longer be used in the same comparative manner. On the other hand, as noted throughout this dissertation, the literature on the influence of extreme standards often includes only the most extreme exemplars in history (e.g., Hitler vs. Santa Claus; Herr, 1986). Based on the current work, it cannot be determined how extreme standard might need to be to no longer be deemed relevant, or if this effect may only occur under other circumstances. Nevertheless, such a dynamic would only further strengthen the call for a broader selection of comparative standards in future research. It also highlights the fact that the issue of vague boundaries of important moderators is not limited to the concept of moderate and extreme standards. What is deemed unobtainable or irrelevant is an equally abstract concept for which the exact conditions are not clearly defined. Hence, if the expected attenuation in response to a particularly extreme standard occurs, it can be deemed unobtainable, but if contrast occurs it must have still been deemed obtainable enough, forcing results to again only speak to auxiliary theories (Meehl, 1990). Notwithstanding, the restriction of the range of standards used in the current dissertation does not permit inferences of comparative effect beyond those boundaries. Additional research would be necessary to extend the findings to a more inclusive range, which may or may not induce this additional effect.

Another reason that this effect might not have been found is the use of other-focused comparisons rather than self-focused ones. As mentioned previously, this was done to ensure that both targets and standards were symmetrical regarding their informational content. Nevertheless, a focus on the other, rather than the self, places the current findings somewhat outside of the most common practices in social comparison research. However, it is unlikely that self-related comparisons, which dominate the literature, would not produce substantially different comparative patterns than those found here. Both self-enhancement and other motivational

mechanisms may differently affect the likelihood for assimilation or contrast to occur in specific situations (e.g., Alicke, 1985; Kunda, 1990; Beauregard & Dunning, 1998, 2001). Such variables will undoubtedly also affect the underlying interactive pattern of these comparative outcomes. For instance, extreme upward comparisons may result in assimilation, as participants may ‘bask in reflected glory’ in some situations (Tesser, 1988). Furthermore, even in absence of these motivational factors, the inherent informational imbalance, at the heart of the decision to move away from self-related judgments, is likely to have a profound effect on the comparative process. For instance, more ambiguous stimuli have been found to be more likely to assimilate towards standards than those of which some knowledge exists (Herr, Sherman, & Fazio, 1983), and informational asymmetries have been theorised to underlie a number of self-enhancement biases in comparative judgments (Chambers & Windschitl, 2004). Whether and how all these factors influence the interactive pattern will need to be probed in future research. Manipulating the informational content of standards and target separately within a paradigm similar to the CJT, would allow the precise modelling of their consequences on the comparative outcomes in a highly controlled manner that is minimally affected by any motivational factors, a benefit that can only be achieved in other-related judgment settings.

Somewhat related to issues of informational content, and its effect on the comparative pattern, is the limited information that participants can infer from facial stimuli alone. Although these facial stimuli have been shown to clearly convey trait information (Oosterhof & Todorov, 2008), their informational content is per definition limited to superficial judgments, as participants have no further knowledge of the target or standards past behaviour. Within the SAM framework (Mussweiler, 2003), the lack of prior knowledge that can be made selectively accessible during the biased hypothesis testing may diminish the size of the effects, relying at

most on simulated behaviour. In fact, as there is no previous knowledge of the targets in this context, the results imply that the selective accessibility mechanism, if indeed the true mechanism of comparative outcomes, must also ease the generation of hypothetical behaviours in line with the similarity hypothesis. Therefore, the accessibility mechanism cannot be limited to actual knowledge of the target, but also imagined knowledge. Other models, such as the reflective versus evaluative model (REM; Markman & McMullen, 2003), have emphasised the potential importance of mental simulation in enhancing the accessibility of counterfactual cognitions that are consistent with the standard. The findings presented here seem to agree that this might be an important part of the comparative process, but that the effects produced seem to be substantially weaker than if factual knowledge did exist. Therefore, providing some broader knowledge about the targets past actions, may elicit larger effects and help investigate the possibility that these two effects work in synchrony. For instance, previous work focusing on unknown targets have often used descriptions of ambiguous behaviour as additional knowledge, which is neutral in regards to the judgment, but is reinterpreted in line with or in contrast to the activated comparative information (e.g., Herr 1986; Mussweiler & Damisch, 2008). Somehow including this type of ambiguous information within the current paradigm may increase the strength of the comparative effects by supplementing the simulated knowledge.

One interesting query in such a study would be if the increased knowledge would only affect the area of assimilation, but not the window of assimilation itself. This would be expected under the SAM since the holistic assessment of similarities, which is thought to determine the direction of the comparative effect, should precede the selective accessibility mechanism. This mechanism would only later call upon and interpret these ambiguous behaviours in line with the initially formed similarity/dissimilarity hypothesis before a judgment is made. Therefore, the

point at which a standard is extreme enough to elicit an initial judgment of dissimilarity should remain the same with additional neutral information, as long as it does not affect how similar the standard and target are judged.

The alternative IEM (Schwarz & Bless, 2007), on the other hand, clearly states that any knowledge that is accessible at the time can be included in the formation of a mental representation. Potentially simulated counterfactual behaviour can, therefore, more easily be incorporated in the IEM framework. However, it too would predict small effect for these types of comparisons, since the size of comparative effects are directly linked to the informational content that is included or excluded from the mental representations of both the target and the standard. Thus, it logically follows that the reduced informational content of facial stimuli would be expected to produce only the small comparative effects that we see reflected in this dissertation. However, the fact that a mental representation of the standard is also constructed seems to make a distinct prediction separate from those made by the SAM. Namely, it may be that broadening the informational content of not just the target, but also the standard could lead to stronger effect at each point in the extremity spectrum in an additive way. Future research may manipulate the informational content, but not strength, of both the target and standard orthogonally to investigate the separate effects each has on the dynamic interactive pattern underlying the comparative outcomes. This would clarify not only the possible consequences of informational asymmetries between standards and targets, as discussed previously, but also be an intriguing way to compare these two theories and offer more precise inferences about the comparative process. Therefore, although the current work itself does not offer a decisive answer on the workings of the comparative process or which model is better suited to describe it, it does offer a valuable and flexible tool for future work to answer these questions.

5.3. Conclusion

Despite the discussed limitations, the current dissertation has uncovered the previously obscured dynamic interactive pattern of assimilation and contrast, which is fundamentally linked to the relative standing of the standard to which a comparison is made. This finding, along with its demonstrated sensitivity to changes in study design, such as item selection and comparative mind-set, demonstrates the limitations of previous designs with limited item and standard selection. Although this high sensitivity also means that the precise patterns found here are likely to be substantially different in different research contexts, this fact only emphasises the points raised regarding the limits of the generalisability of previous findings, and strengthens the call for a new standard of comparative research. As a whole, the current dissertation has made a clear case for a more holistic approach to the investigation of comparative effects, while also providing a basic paradigm that can be used for strong tests of theory by modelling the complete dynamic interactive patterns that underlie comparative outcomes.

References

- Ahrens, A. H. (1991). Dysphoria and social comparison: Combining information regarding others' performance. *Journal of Social and Clinical Psychology, 10*, 190-205.
<https://doi.org/10.1521/jscp.1991.10.2.190>
- Alicke, M. D. (1985). Global self-evaluation as determined by the desirability and controllability of trait adjectives. *Journal of Personality and Social Psychology, 49*, 1621–1630.
<https://doi.org/10.1037/0022-3514.49.6.1621>
- Alves, H., Koch, A., & Unkelbach, C. (2016). My friends are all alike—The relation between liking and perceived similarity in person perception. *Journal of Experimental Social Psychology, 62*, 103–117. <https://doi.org/10.1016/j.jesp.2015.10.011>
- Alves, H., Koch, A., & Unkelbach, C. (2017). The “common good” phenomenon: Why similarities are positive and differences are negative. *Journal of Experimental Psychology: General, 146*, 512. <https://doi.org/10.1037/xge0000276>
- Alves, H., Koch, A., & Unkelbach, C. (2018). A cognitive-ecological explanation of intergroup biases. *Psychological Science, 29*, 1126-1133.
<https://doi.org/10.1177/0956797618756862>
- Bailis, D. S., & Chipperfield, J. G. (2006). Emotional and self-evaluative effects of social comparison information in later life: How are they moderated by collective self-esteem? *Psychology and Aging, 21*, 291-302. <https://doi.org/10.1037/0882-7974.21.2.291>
- Baillargeon, R. (1991). Reasoning about the height and location of a hidden object in 4.5- and 6.5-month-old infants. *Cognition, 38*, 13-42. [https://doi.org/10.1016/0010-0277\(91\)90021-U](https://doi.org/10.1016/0010-0277(91)90021-U)

-
- Baldwin, M., & Mussweiler, T. (2018). The culture of social comparison. *Proceedings of the National Academy of Sciences, 115*, E9067-E9074.
<https://doi.org/10.1073/pnas.1721555115>
- Ballem, C.C., & Todorov, A. (2007). Predicting political elections from rapid and unreflective face judgments. *Proceedings of the National Academy of Sciences, 104*, 17948–17953.
<https://doi.org/10.1073/pnas.0705435104>
- Barker, P., & Imhoff, R. (2020). *Comparing Curves: Describing the Dynamic Interaction of Assimilation and Contrast in Facial Evaluations*. Manuscript submitted for publication.
- Barker, P., Dotsch, R., & Imhoff, R. (2020). Assimilation and Contrast in Spontaneous Comparisons: Heterogeneous Effects of Standard Extremity in Facial Evaluations. *International Review of Social Psychology, 33*, 11. <http://doi.org/10.5334/irsp.402>.
- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*, 1–48.
<https://doi.org/10.18637/jss.v067.i01>
- Beauregard, K. S., & Dunning, D. (1998). Turning up the contrast: Self-enhancement motives prompt egocentric contrast effects in social judgments. *Journal of Personality and Social Psychology, 74*, 606–621.
- Beauregard, K. S., & Dunning, D. (2001). Defining self-worth: Trait self-esteem moderates the use of self-serving trait definitions in social judgment. *Motivation and Emotion, 25*, 135–161. <https://doi.org/10.1037//0022-3514.74.3.606>
- Bell, B. A., Morgan, G. B., Schoeneberger, J. A., Kromrey, J. D., & Ferron, J. M. (2014). How low can you go? An investigation of the influence of sample size and model complexity on point and interval estimates in two-level linear models. *Methodology: European*

- Journal of Research Methods for the Behavioral and Social Sciences*, 10, 1–11.
<https://doi.org/10.1027/1614-2241/a000062>
- Berger S., Graham, N., & Zeileis, A. (2017). Various versatile variances: An object-oriented implementation of clustered covariances in R, Working Papers, *Faculty of Economics and Statistics, University of Innsbruck*,
<https://EconPapers.repec.org/RePEc:inn:wpaper:2017-12>.
- Biernat, M. (2003). Toward a broader view of social stereotyping. *American Psychologist*, 58, 1019. <https://doi.org/10.1037/0003-066X.58.12.1019>
- Bless, H., & Schwarz, N. (1998). Context effects in political judgement: Assimilation and contrast as a function of categorization processes. *European Journal of Social Psychology*, 28, 159-172. [https://doi.org/10.1002/\(SICI\)1099-0992\(199803/04\)28:2<159::AID-EJSP860>3.0.CO;2-4](https://doi.org/10.1002/(SICI)1099-0992(199803/04)28:2<159::AID-EJSP860>3.0.CO;2-4)
- Bless, H., & Schwarz, N. (2010). Mental construal and the emergence of assimilation and contrast effects: The inclusion/exclusion model. In *Advances in experimental social psychology* (Vol. 42, pp. 319-373). Academic Press. [https://doi.org/10.1016/S0065-2601\(10\)42006-7](https://doi.org/10.1016/S0065-2601(10)42006-7)
- Bless, H., Schwarz, N., & Wänke, M. (2003). The size of context effects in social judgment. In J. P. Forgas, K. D. Williams, & W. von Hippel (eds.), *Social judgments: Implicit and explicit processes* (pp. 180-197). Cambridge, UK: Cambridge University Press
- Brewer, M. B., & Weber, J. G. (1994). Self-evaluation effects of interpersonal versus intergroup social comparison. *Journal of Personality and Social Psychology*, 66, 268.
<https://doi.org/10.1037/0022-3514.66.2.268>

- Brown, J. D., Novick, N. J., Lord, K. A., & Richards, J. M. (1992). When Gulliver travels: social context, psychological closeness, and self-appraisals. *Journal of Personality and Social Psychology, 62*, 717. <https://doi.org/10.1037/0022-3514.62.5.717>
- Buckingham, J. T., & Alicke, M. D. (2002). The influence of individual versus aggregate social comparison and the presence of others on self-evaluations. *Journal of Personality and Social Psychology, 83*, 1117-1130. <https://doi.org/10.1037/0022-3514.83.5.1117>
- Buunk, A. P., Groothof, H. A. K., & Siero, F. W. (2007). Social comparison and satisfaction with one's social life. *Journal of Social and Personal Relationships, 24*, 197-205. <https://doi.org/10.1177/0265407507075410>
- Buunk, B. P., & Ybema, J. F. (1997). Social comparisons and occupational stress: The identification-contrast model. In B. P. Buunk & F. X. Gibbons (Eds.), *Health, coping, and well-being: Perspectives from social comparison theory* (p. 359–388). Lawrence Erlbaum Associates Publishers.
- Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin, 54*, 297–312. <http://dx.doi.org/10.1037/h0040950>
- Canty, A., & Ripley, B. (2019). boot: Bootstrap R (S-Plus) Functions. R package version 1.3-24.
- Carandini, M., & Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nature Reviews Neuroscience, 13*, 51-62. <https://doi.org/10.1038/nrn3136>
- Cash, T. F., Cash, D. W., & Butters, J. W. (1983). "Mirror, mirror, on the wall...?": Contrast effects and self-evaluations of physical attractiveness. *Personality and Social Psychology, 9*, 351-358. <https://doi.org/10.1177/0146167283093004>

-
- Chambers, J. R., & Windschitl, P. D. (2004). Biases in social comparative judgments: the role of nonmotivated factors in above-average and comparative-optimism effects. *Psychological Bulletin, 130*, 813. <https://doi.org/10.1037/0033-2909.130.5.813>
- Cheng, R. W. Y., & Lam, S. F. (2007). Self-construal and social comparison effects. *British Journal of Educational Psychology, 77*, 197-211. <https://doi.org/10.1348/000709905X72795>
- Chung, T., & Mallery, P. (1999). Social comparison, individualism-collectivism, and self-esteem in China and the United States. *Current Psychology, 18*, 340-352. <https://doi.org/10.1007%2Fs12144-999-1008-0>
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Boston, MA: Houghton Mifflin.
- Corcoran, K., Epstude, K., Damisch, L., & Mussweiler, T. (2011). Fast similarities: efficiency advantages of similarity-focused comparisons. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*, 1280. <https://doi.org/10.1037/a0023922>
- Corcoran, K., Hundhammer, T., & Mussweiler, T. (2009). A tool for thought! When comparative thinking reduces stereotyping effects. *Journal of Experimental Social Psychology, 45*, 1008-1011. <https://doi.org/10.1016/j.jesp.2009.04.015>
- Cordes, S., Goldstein, A., & Heller, E. (2014). Sets within sets: The influence of set membership on numerical estimates. *Journal of Experimental Psychology: Human Perception and Performance, 40*, 94-105. <https://doi.org/10.1037/a0034131>
- Corneille, O., Yzerbyt, V. Y., Pleyers, G., & Mussweiler, T. (2009). Beyond awareness and resources: Evaluative conditioning may be sensitive to processing goals. *Journal of Experimental Social Psychology, 45*, 279-282. <https://doi.org/10.1016/j.jesp.2008.08.020>

- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302. <https://doi.org/10.1037/h0040957>
- Diel, K., & Hofmann, W. (2019). Inspired to perspire: The interplay of social comparison direction and standard extremity in the context of challenging exercising goals. *Social Cognition*, *37*, 247-265. <https://doi.org/10.1521/soco.2019.37.3.247>
- Dunning, D., & Hayes, A. F. (1996). Evidence for egocentric comparison in social judgment. *Journal of Personality and Social Psychology*, *71*, 213. <https://doi.org/10.1037/0022-3514.71.2.213>
- Dunning, D., Meyerowitz, J. A., & Holzberg, A. D. (1989). Ambiguity and self-evaluation: The role of idiosyncratic trait definitions in self-serving assessments of ability. *Journal of Personality and Social Psychology*, *57*, 1082. <https://doi.org/10.1037/0022-3514.57.6.1082>
- Epstude, K., & Mussweiler, T. (2009). What you feel is how you compare: How comparisons influence the social induction of affect. *Emotion*, *9*, 1. <https://doi.org/10.1037/a0014148>
- Faith, M. S., Leone, M. A., & Allison, D. B. (1997). The effects of self-generated comparison targets, BMI, and social comparison tendencies on body image appraisal. *Eating Disorders*, *5*, 128-140. <https://doi.org/10.1080/10640269708249216>
- Fardouly, J., Pinkus, R. T., & Vartanian, L. R. (2017). The impact of appearance comparisons made through social media, traditional media, and in person in women's everyday lives. *Body Image*, *20*, 31-39. <https://doi.org/10.1016/j.bodyim.2016.11.002>
- Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191. <https://doi.org/10.3758/BF03193146>

-
- Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, 7, 117–140.
<https://doi.org/10.1177/001872675400700202>
- Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, 11, 77–83.
<https://doi.org/10.1016/j.tics.2006.11.005>
- Förster, J., Liberman, N., & Kuschel, S. (2008). The effect of global versus local processing styles on assimilation versus contrast in social judgment. *Journal of Personality and Social Psychology*, 94, 579. <https://doi.org/10.1037/0022-3514.94.4.579>
- Furl, N. (2016). Facial-attractiveness choices are predicted by divisive normalization. *Psychological Science*, 27, 1379–1387. <https://doi.org/10.1177/0956797616661523>
- Gentner, D., & Markman, A. B. (1994). Structural alignment in comparison: No difference without similarity. *Psychological Science*, 5, 152–158. <https://doi.org/10.1111/j.1467-9280.1994.tb00652.x>
- Gentner, D., & Rattermann, M. J. (1991). Language and the career of similarity. In S. A. Gelman & J. P. Byrnes (Eds.), *Perspectives on language and thought: Interrelations in development* (p. 225–277). Cambridge University Press.
<https://doi.org/10.1017/CBO9780511983689.008>
- Gerber, J. P., Wheeler, L., & Suls, J. (2018). A social comparison theory meta-analysis 60+ years on. *Psychological Bulletin*, 144, 177. <https://doi.org/10.1037/bul0000127>
- Gibbons, F. X., & Buunk, B. P. (1999). Individual differences in social comparison: Development of a scale of social comparison orientation. *Journal of Personality and Social Psychology*, 76, 129. <https://doi.org/10.1037/0022-3514.76.1.129>

- Goto, T., Uchiyama, I., Imai, A., Takahashi, S. Y., Hanari, T., Nakamura, S., & Kobari, H. (2007). Assimilation and contrast in optical illusions. *Japanese Psychological Research*, 49, 33-44. <https://doi.org/10.1111/j.1468-5884.2007.00330.x>
- Häfner, M., & Schubert, T. W. (2009). Feel the difference! The influence of ease experiences on the direction of social comparisons. *Journal of Experimental Social Psychology*, 45, 291-294. <https://doi.org/10.1016/j.jesp.2008.09.008>
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, 466, 29-29. <https://doi.org/10.1038/466029a>
- Herr, P. M. (1986). Consequences of priming: Judgment and behavior. *Journal of Personality and Social Psychology*, 51, 1106. <https://doi.org/10.1037/0022-3514.51.6.1106>
- Herr, P. M., Sherman, S. J., & Fazio, R. H. (1983). On the consequences of priming: Assimilation and contrast effects. *Journal of Experimental Social Psychology*, 19, 323–340. [https://doi.org/10.1016/0022-1031\(83\)90026-4](https://doi.org/10.1016/0022-1031(83)90026-4)
- Higgins, E. T. (1987). Self-discrepancy: A theory relating self and affect. *Psychological Review*, 94, 319–340. <https://doi.org/10.1037/0033-295X.94.3.319>
- Hsee, C. K. (1996). The evaluability hypothesis: An explanation of preference reversals between joint and separate evaluations of alternatives. *Organizational Behavior and Human Decision Processes*, 67, 247–257. <https://ssrn.com/abstract=930092>
- IJzerman, H., & Semin, G. R. (2009). The thermometer of social relations: Mapping social proximity on temperature. *Psychological Science*, 20, 1214-1220. <https://doi.org/10.1111/j.1467-9280.2009.02434.x>

-
- Imhoff, R., & Koch, A. (2017). How orthogonal are the Big Two of social perception? On the curvilinear relation between agency and communion. *Perspectives in Psychological Science, 12*, 122–137. <https://doi.org/10.1177/17456916166657334>
- Imhoff, R., Woelki, J., Hanke, S., & Dotsch, R. (2013). Warmth and competence in your face! Visual encoding of stereotype content. *Frontiers in Psychology, 4* (386). <https://doi.org/10.3389/fpsyg.2013.00386>
- Johnson, C., Lammers, J. (2012). The powerful disregard social comparison information. *Journal of Experimental Social Psychology, 48*, 329-334. <https://doi.org/10.1016/j.jesp.2011.10.010>
- Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review, 93*, 136. <https://doi.org/10.1037/0033-295X.93.2.136>
- Koch, A., Imhoff, R., Dotsch, R., Unkelbach, C., & Alves, H. (2016). The ABC of stereotypes about groups: Agency/socioeconomic success, conservative–progressive beliefs, and communion. *Journal of Personality and Social Psychology, 110*, 675. <https://doi.org/10.1037/pspa0000046>
- Kruglanski, A. W. (1996). Motivated social cognition: Principles of the interface. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 493–520). New York: Guilford Press.
- Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*, 480–498. <https://doi.org/10.1037/0033-2909.108.3.480>
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1–26. <https://doi.org/10.18637/jss.v082.i13>

- Lin, L. L., & Kulik, J. A. (2002). Social comparison and women's body satisfaction. *Basic and Applied Social Psychology, 24*, 115-123.
https://doi.org/10.1207/S15324834BASP2402_4
- Lockwood, P., & Kunda, Z. (1997). Superstars and me: Predicting the impact of role models on the self. *Journal of Personality and Social Psychology, 73*, 91.
<https://doi.org/10.1037/0022-3514.73.1.91>
- Lockwood, P., & Kunda, Z. (1999). Increasing the salience of one's best selves can undermine inspiration by outstanding role models. *Journal of Personality and Social Psychology, 76*, 214-228. <https://doi.org/10.1037/0022-3514.76.2.214>
- Lockwood, P., Wong, C., McShane, K., & Dolderman, D. (2005). The impact of positive and negative fitness exemplars on motivation. *Basic and Applied Social Psychology, 27*, 1-13.
https://doi.org/10.1207/s15324834basp2701_1
- Lüdecke, D. (2018). ggeffects: Tidy data frames of marginal effects from regression models. *Journal of Open Source Software, 3*, 772. <http://doi.org/10.21105/joss.00772>.
- Lüdecke, D., Ben-Shachar, M., & Makowski, D. (2020). "Describe and understand your model's parameters." CRAN. <https://doi.org/10.5281/zenodo.3731932>
- Markman, A. B., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology, 25*, 431-467.
<https://doi.org/10.1006/cogp.1993.1011>
- Markman, K. D., & McMullen, M. N. (2003). A reflection and evaluation model of comparative thinking. *Personality and Social Psychology Review, 7*, 244-267.
https://doi.org/10.1207/S15327957PSPR0703_04

-
- Martin, R., Suls, J., & Wheeler, L. (2002). Ability evaluation by proxy: Role of maximal performance and related attributes in social comparison. *Journal of Personality and Social Psychology, 82*, 781-791. <https://doi.org/10.1037/0022-3514.82.5.781>
- McFarland, C., Buehler, R., & MacKay, L. (2001). Affective responses to social comparisons with extremely close others. *Social Cognition, 19*, 547-586.
<https://doi.org/10.1521/soco.19.5.547.19911>
- Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry, 1*, 108-141.
<https://www.jstor.org/stable/1448768>
- Mendes, W. B., Blascovich, J., Major, B., & Seery, M. (2001). Challenge and threat responses during downward and upward social comparisons. *European Journal of Social Psychology, 31*, 477-497. <https://doi.org/10.1002/ejsp.80>
- Mori, K., & Mori, H. (2011). No confederates needed: Social comparison without collaboration. *Social Behavior and Personality, 39*, 543-552. <https://doi.org/10.2224/sbp.2011.39.4.543>
- Mussweiler, T. (2001a). 'Seek and ye shall find': Antecedents of assimilation and contrast in social comparison. *European Journal of Social Psychology, 31*, 499-509.
<https://doi.org/10.1002/ejsp.75>
- Mussweiler, T. (2001b). Focus of comparison as a determinant of assimilation versus contrast in social comparison. *Personality and Social Psychology Bulletin, 27*, 38-47.
<https://doi.org/10.1177/0146167201271004>
- Mussweiler, T. (2003). Comparison processes in social judgment: mechanisms and consequences. *Psychological Review, 110*, 472. <https://doi.org/10.1037/0033-295X.110.3.472>

- Mussweiler, T., & Bodenhausen, G. V. (2002). I know you are, but what am I? Self-evaluative consequences of judging in-group and out-group members. *Journal of Personality and Social Psychology, 82*, 19. <https://doi.org/10.1037/0022-3514.82.1.19>
- Mussweiler, T., & Epstude, K. (2009). Relatively fast! Efficiency advantages of comparative thinking. *Journal of Experimental Psychology: General, 138*, 1. <https://doi.org/10.1037/a0014374>
- Mussweiler, T., & Posten, A. C. (2012). Relatively certain! Comparative thinking reduces uncertainty. *Cognition, 122*, 236–240. <https://doi.org/10.1016/j.cognition.2011.10.005>
- Mussweiler, T., & Rüter, K. (2003). What friends are for! The use of routine standards in social comparison. *Journal of Personality and Social Psychology, 85*, 467. <https://doi.org/10.1037/0022-3514.85.3.467>
- Mussweiler, T., & Strack, F. (1999). Hypothesis-consistent testing and semantic priming in the anchoring paradigm: A selective accessibility model. *Journal of Experimental Social Psychology, 35*, 136-164. <https://doi.org/10.1006/jesp.1998.1364>
- Mussweiler, T., & Strack, F. (2000). The ‘relative self’: Informational and judgmental consequences of comparative self-evaluation. *Journal of Personality and Social Psychology, 79*, 23. <https://doi.org/10.1037//0022-3514.79.1.23>
- Mussweiler, T., Rüter, K., & Epstude, K. (2004a). The man who wasn’t there: Subliminal social comparison standards influence self-evaluation. *Journal of Experimental Social Psychology, 40*, 689–696. <https://doi.org/10.1016/j.jesp.2004.01.004>
- Mussweiler, T., Rüter, K., & Epstude, K. (2004b). The ups and downs of social comparison: Mechanisms of assimilation and contrast. *Journal of Personality and Social Psychology, 87*, 832. <https://doi.org/10.1037/0022-3514.87.6.832>

-
- Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39.
- O'Brien, L. T., Crandall, C. S., Horstman-Reser, A., Warner, R., Alsbrooks, A., & Blodorn, A. (2010). But I'm no bigot: How prejudiced White Americans maintain unprejudiced self-images. *Journal of Applied Social Psychology*, *40*, 917-946.
<https://doi.org/10.1111/j.1559-1816.2010.00604.x>
- Ohmann, K., Stahl, J., Mussweiler, T., & Kedia, G. (2016). Immediate relativity: EEG reveals early engagement of comparison in social information processing. *Journal of Experimental Psychology: General*, *145*, 1512. doi: 10.1037/xge0000222
- Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, *105*, 11087–11092.
<https://doi.org/10.1073/pnas.0805664105>
- Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153-163. <https://doi.org/10.1016/j.jesp.2017.01.006>
- Pelham, B. W., & Wachsmuth, J. O. (1995). The waxing and waning of the social self: Assimilation and contrast in social comparison. *Journal of Personality and Social Psychology*, *69*, 825. <https://doi.org/10.1037/0022-3514.69.5.825>
- Philippot, P., Schwarz, N., Carrera, P., De Vries, N., & Van Yperen, N. W. (1991). Differential effects of priming at the encoding and judgment stage. *European Journal of Social Psychology*, *21*, 293-302. <https://doi.org/10.1002/ejsp.2420210403>

- Posten, A. C., & Mussweiler, T. (2013). When distrust frees your mind: The stereotype-reducing effects of distrust. *Journal of Personality and Social Psychology, 105*, 567.
<https://doi.org/10.1037/a0033170>
- R Core Team (2018). R: A language and environment for statistical computing [Computer software]. *R Foundation for Statistical Computing*, Vienna, Austria. Retrieved from <https://www.R-project.org/>.
- Raat, A.N., Kuks, Jan B.M., Van Hell, E.A., & Cohen-Schotanus, J. (2013). Peer influence on students' estimates of performance: Social comparison in clinical rotations. *Medical Education, 47*, 190-197. <https://doi.org/10.1111/medu.12066>
- Reber, R., & Schwarz, N. (1999). Effects of perceptual fluency on judgments of truth. *Consciousness and Cognition: An International Journal, 8*, 338–342.
<https://doi.org/10.1006/ccog.1999.0386>
- Schwarz, N., & Bless, H. (1992a). Constructing reality and its alternatives: Assimilation and contrast effects in social judgment. In L. L. Martin & A. Tesser (Eds.), *The construction of social judgments* (pp. 217–245). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Schwarz, N., & Bless, H. (1992b). Scandals and the public's trust in politicians: Assimilation and contrast effects. *Personality and Social Psychology Bulletin, 18*, 574-579.
<https://doi.org/10.1177/0146167292185007>
- Schwarz, N., & Bless, H. (2007). Mental construal processes: The inclusion/exclusion model. In D. Stapel & J. Suls (Eds.), *Assimilation and contrast in social psychology* (pp. 119–141). Philadelphia, PA: Psychology Press.

-
- Simonsohn, U. (2018). Two lines: A valid alternative to the invalid testing of u-shaped relationships with quadratic regressions. *Advances in Methods and Practices in Psychological Science, 1*, 538–555. <https://doi.org/10.1177/2515245918805755>
- Smith, W. P., & Sachs, P. R. (1997). Social comparison and task prediction: Ability similarity and the use of a proxy. *British Journal of Social Psychology, 36*, 587-602. <https://doi.org/10.1111/j.2044-8309.1997.tb01151.x>
- Strull, T. K., & Gaelick, L. (1983). General principles and individual differences in the self as a habitual reference point: An examination of self-other judgments of similarity. *Social Cognition, 2*, 108-121. <https://doi.org/10.1521/soco.1983.2.2.108>
- Stanton, A. L., Danoff-Burg, S., Cameron, C. L., Snider, P. R., & Kirk, S. B. (1999). Social comparison and adjustment to breast cancer: An experimental examination of upward affiliation and downward evaluation. *Health Psychology, 18*, 151. <https://doi.org/10.1037/0278-6133.18.2.151>
- Steinmetz, J., & Mussweiler, T. (2011). Breaking the ice: How physical warmth shapes social comparison consequences. *Journal of Experimental Social Psychology, 47*, 1025-1028. <https://doi.org/10.1016/j.jesp.2011.03.022>
- Tesser, A. (1988). Toward a self-evaluation maintenance model of social behavior. *Advances in Experimental Social Psychology, 21*, 181-227. [https://doi.org/10.1016/S0065-2601\(08\)60227-0](https://doi.org/10.1016/S0065-2601(08)60227-0)
- Tesser, A., & Campbell, J. (1982). Self-evaluation maintenance and the perception of friends and strangers. *Journal of Personality, 50*, 261-279. <https://doi.org/10.1111/j.1467-6494.1982.tb00750.x>

- Thornton, B., & Moore, S. (1993). Physical attractiveness contrast effect: Implications for self-esteem and evaluations of the social self. *Personality and Social Psychology Bulletin, 19*, 474-480. <https://doi.org/10.1177/0146167293194012>
- Todorov, A., Dotsch, R., Porter, J. M., Oosterhof, N. N., & Falvello, V. B. (2013). Validation of data-driven computational models of social perception of faces. *Emotion, 13*, 724. <https://doi.org/10.1037/a0032335>
- Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science, 308*, 1623–1626. <https://doi.org/10.1126/science.1110589>
- Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual Review of Psychology, 66*, 519–545. <https://doi.org/10.1146/annurev-psych-113011-143831>
- Tversky, A. (1977). Features of similarity. *Psychological Review, 84*, 327. <https://doi.org/10.1037/0033-295X.84.4.327>
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124-1131.
- Upshaw, H. S. (1978). Social influence on attitudes and on anchoring of congeneric attitude scales. *Journal of Experimental Social Psychology, 14*, 327–339. [https://doi.org/10.1016/0022-1031\(78\)90029-X](https://doi.org/10.1016/0022-1031(78)90029-X)
- Veldhuis, J., Konijn, E. A., Seidell, J. C. (2012). Weight information labels on media models reduce body dissatisfaction in adolescent girls. *Journal of Adolescent Health, 50*, 600-606. <https://doi.org/10.1016/j.jadohealth.2011.10.249>

-
- Venables, B., Hornik, K., & Maechler, M. (2016). Polynom: A collection of functions to implement a class for univariate polynomial manipulations. R package version 1.3-9. <https://CRAN.R-project.org/package=polynom>
- Verduyn, P., Ybarra, O., Résibois, M., Jonides, J., & Kross, E. (2017). Do social network sites enhance or undermine subjective well-being? A critical review. *Social Issues and Policy Review, 11*, 274-302. <https://doi.org/10.1111/sipr.12033>
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1-48. <https://doi.org/10.18637/jss.v036.i03>
- Vogel, E. A., Rose, J. P., Roberts, L. R., & Eckles, K. (2014). Social comparison, social media, and self-esteem. *Psychology of Popular Media Culture, 3*, 206. <https://doi.org/10.1037/ppm0000047>
- Wasilenko, K. A., Kulik, J. A., & Wanic, R. A. (2007). Effects of social comparisons with peers on women's body satisfaction and exercise behavior. *International Journal of Eating Disorders, 40*, 740-745. <https://doi.org/10.1002/eat.20433>
- Wedell, D. H., Parducci, A., & Geiselman, R. E. (1987). A formal analysis of ratings of physical attractiveness: Successive contrast and simultaneous assimilation. *Journal of Experimental Social Psychology, 23*, 230-249. [https://doi.org/10.1016/0022-1031\(87\)90034-5](https://doi.org/10.1016/0022-1031(87)90034-5)
- Wells, G. L., & Windschitl, P. D. (1999). Stimulus sampling and social psychological experimentation. *Personality and Social Psychology Bulletin, 25*, 1115-1125. <https://doi.org/10.1177/01461672992512005>
- Wiggins, J. S., Phillips, N., & Trapnell, P. (1989). Circular reasoning about interpersonal behavior: Evidence concerning some untested assumptions underlying diagnostic

classification. *Journal of Personality and Social Psychology*, 56, 296.

<https://doi.org/10.1037/0022-3514.56.2.296>

Willis, J., & Todorov, A. (2006) First impressions: Making up your mind after 100 ms exposure to a face. *Psychological Science*, 17, 592–598. <https://doi.org/10.1111/j.1467-9280.2006.01750.x>

Wood, S., & Scheipl, F. (2017). Gamm4: Generalized additive mixed models using 'mgcv' and 'lme4'. R package version 0.2-5. <https://CRAN.R-project.org/package=gamm4>

Wundt, W. (1980). Outlines of psychology. In *Wilhelm Wundt and the making of a scientific psychology* (pp. 179-195). Springer, Boston, MA. (Originally published in 1897)

Zeileis, A. (2004). “Econometric computing with HC and HAC covariance matrix estimators.” *Journal of Statistical Software*, 11(10), 1-17. <http://doi.org/10.18637/jss.v011.i10>

Zhang, S., & Markman, A. B. (2001). Processing product unique features: Alignability and involvement in preference construction. *Journal of Consumer Psychology*, 11, 13–27. http://doi.org/10.1207/S15327663JCP1101_2


Appendix B



Figure A1. Means and standard errors of Z transformed responses for each extremity step and dimension.

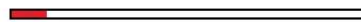
Appendix C

Please identify the judgment target:



How often a month does the target enforce his opinion?

Fortschrittsbalken



Wie oft setzt die zu beurteilende Person seine Meinung durch im Verlauf von 6 Monaten?

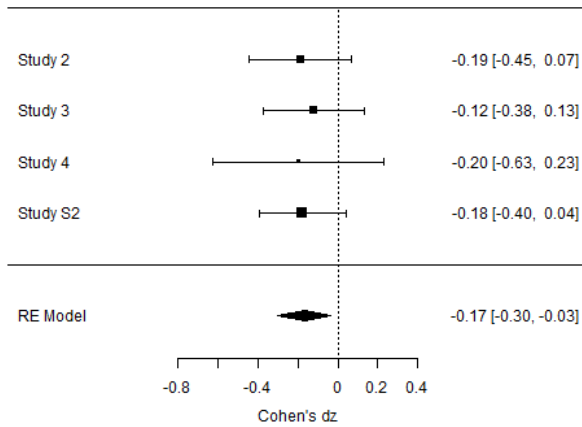


Enter

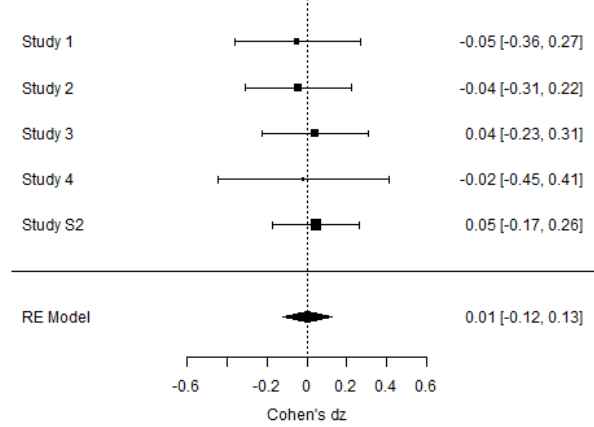
Figure A2. Example of a CJT trials online (top) and in the lab (bottom).

Appendix D

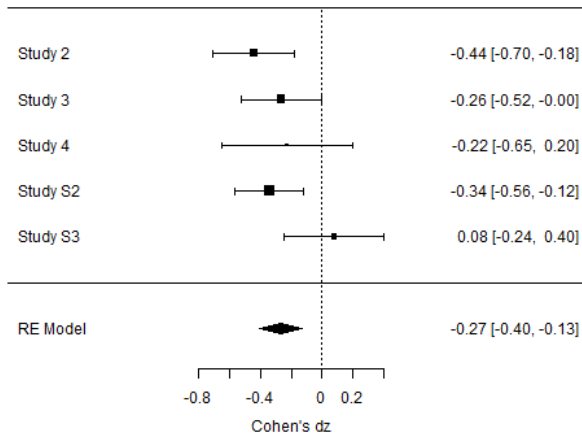
Competence:
Moderate



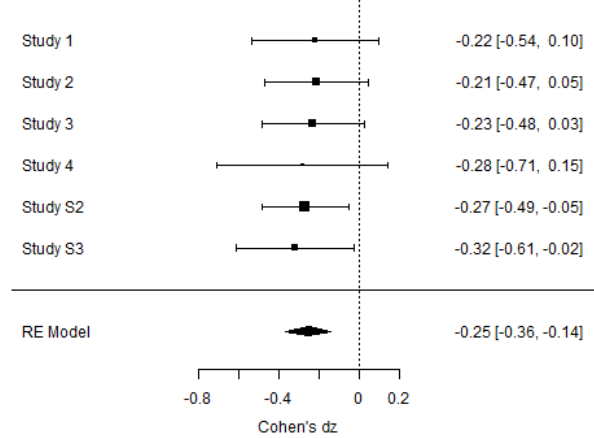
Extreme



Dominance:
Moderate

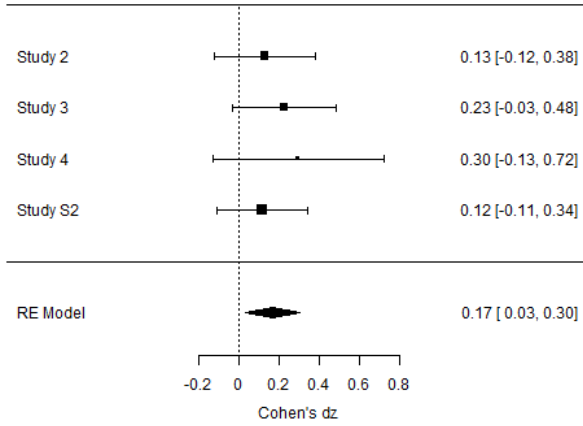


Extreme

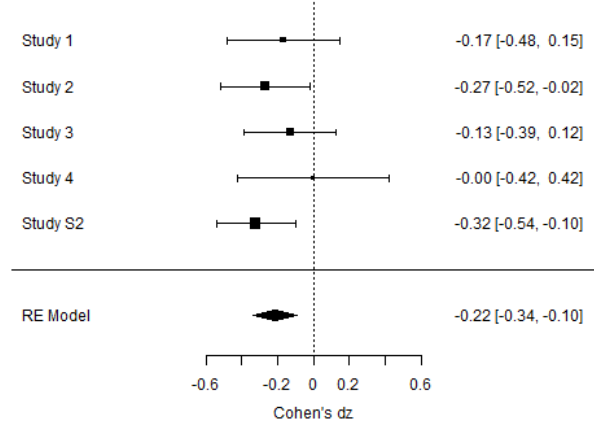


Extraversion:

Moderate

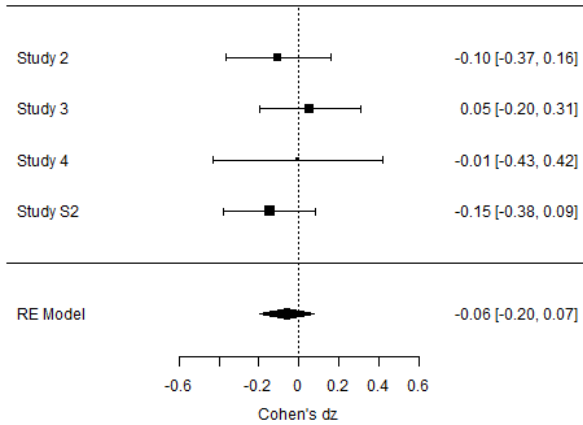


Extreme

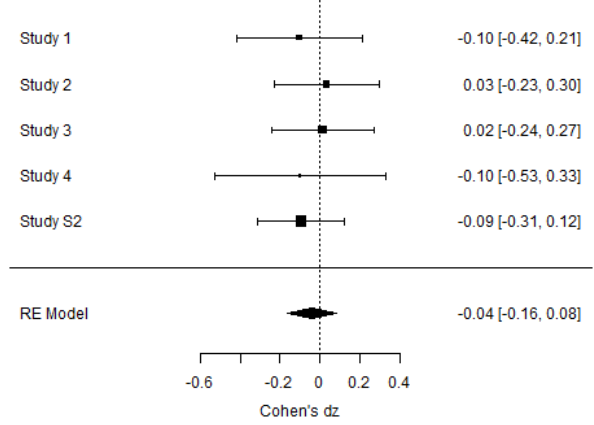


Likability:

Moderate

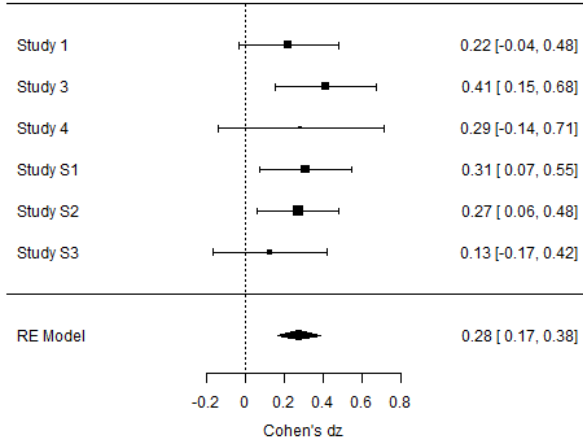


Extreme

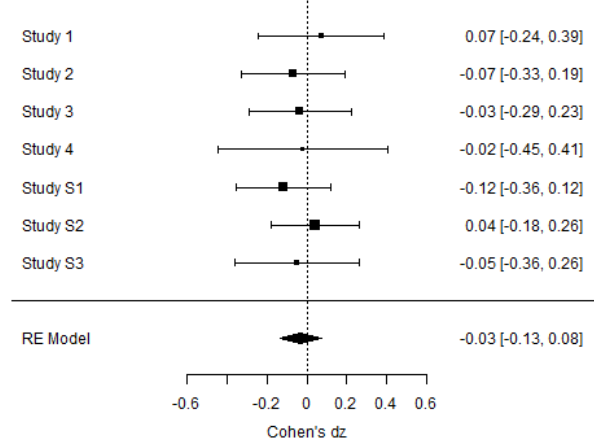


Trustworthiness:

Moderate



Extreme



Appendix E

Table A1.

Items selected to represent Dominant and Submissive behaviours for use in Study 3.4.

Dominant:

1. How many times does the target enforce his opinion in 6 month?
2. How many times does the target take the lead in a group in 6 months?
3. How many times does the target try to persuade people to think like they do in 6 months?

Submissive:

1. How many times does the target follow someone else's instructions in 6 months?
 2. How many times does the target try to speak quietly in 6 months?
 3. How many times does the target try to avoid making decisions in 6 months?
-

Table A2.

Items selected to represent Trustworthy and Untrustworthy behaviours for use in Study 3.5 & 4.1.

Trustworthy:

1. How many times does the target keep his promises in 6 months?
2. How many times does the target return things that they have borrowed in 6 months?
3. How many times does the target act friendly and smile at people in 6 months?

Untrustworthy:

1. How many times does the target not show up to an appointment in 6 month?
 2. How many times does the target try to manipulate or trick other people in 6 months?
 3. How many times does the target do something illegal in 6 months?
-