Backlash for Gender Atypicality

Sanne Nauts

Backlash for Gender Atypicality

Sanne Nauts

Backlash for Gender Atypicality

Proefschrift

ter verkrijging van de graad van doctor

aan de Radboud Universiteit Nijmegen

op gezag van de rector magnificus prof. dr. Th.L.M. Engelen,

volgens besluit van het college van decanen

in het openbaar te verdedigen op dinsdag 17 maart 2015

om 12.30 uur precies

door

Sanne Nauts

geboren op 24 januari 1985

te Celle (Duitsland)

Promotor: Prof. dr. Daniël H.J. Wigboldus

Copromotor: Dr. Oliver Langner (Universitätsklinikum Schleswig-Holstein Lübeck, Duitsland)

Manuscriptcommissie:

Prof. dr. E.S. Becker

Prof. dr. S. Otten (Rijksuniversiteit Groningen)

Dr. B. Derks (Universiteit Leiden)

Contents

CHAPTER 1

General Introduction

Since the advance of the women's movement, women have gone from being "the slave of any boy whose parents forced a ring upon her finger" (Woolf, 1929/2012) to being CEOs and heads of state. In spite of the enormous progress that has been made, men and women still face highly divergent societal outcomes. For example, women remain underrepresented in high status positions that require agentic qualities (e.g., 3% of CEOs of Fortune 500-companies and less than 1% of heads of state is female; Catalyst, 2012; UN, 2013) while men remain underrepresented in low status positions that require communal qualities (e.g., 93% of American nurses is female; US Department of Health and Human Services, 2010). In fact, in order to end gender segregation in the labor market and reach full gender parity in all occupations, 65% of women would have to switch jobs (Rudman, & Glick, 2008). In other aspects of life (and death), men and women also face divergent societal outcomes, with men being much more likely than women to commit suicide (e.g., 79% of suicide victims in the US are male; CDC, 2012) or to be the victim of violent crime (e.g., 77% of homicide victims in the US are male; BJS, 2008). As these examples illustrate, people's outcomes in life remain heavily intertwined with their gender.

There are myriad reasons for men's and women's differential outcomes in life, ranging from biological factors to the differential treatment of men and women. From the minute they are born, boys and girls are expected to portray different traits and behaviors (Rubin, Provenzano, & Luria, 1974). Later in life, these gender stereotypes keep being enforced, so that men and women are discouraged from showing behaviors that are considered atypical for their gender. That is, research on *backlash for gender atypicality* (Rudman, 1998; Rudman, & Glick, 2001) suggests that gender atypical behavior (e.g., weakness or communality in men; agency or dominance in women) can have a range of negative consequences. Men, for example, may be disliked, effeminated, and casted off as weak or psychologically unstable if they engage in stereotypically feminine behaviors such as modesty, passiveness or self-disclosure (Costrich, Feinstein, Kidder, Marecek, & Pascale, 1975; Derlega, & Chaikin, 1976; Moss-Racusin et al., 2010). Likewise, women may be sabotaged, disliked, and turned down for leadership positions if they are agentic, assertive, or self-promoting (Heilman, Block, & Martell, 1995; Phelan,

Moss-Racusin, & Rudman, 2008; Rudman, 1998; Rudman, & Fairchild, 2004; Rudman, Moss-Racusin, Phelan, & Nauts, 2012a; Rudman, & Glick, 1999; 2011; for reviews, see Eagly, & Karau, 2002; Rudman, & Phelan, 2008; Rudman, Moss-Racusin, Glick, & Phelan, 2012b). Because people often refrain from showing gender atypical behavior out of a fear of being disliked (Moss-Racusin, & Rudman, 2010; Rudman, & Fairchild, 2004), backlash may straitjacket members of both genders by limiting the behavioral options that are available to them.

Backlash serves as a major impediment for reaching gender parity, but why are people motivated to penalize gender deviants? The Status Incongruity Hypothesis (SIH; Rudman et al., 2012a) suggests that people engage in backlash as a way of protecting the gender status quo. Whenever a woman enacts high status behavior (e.g., agency) or a man enacts low status behavior (e.g., communality), this behavior is incongruent with the status of their gender (it is *status incongruent*; Rudman et al., 2012a). Researchers distinguish between ascribed status (the status that people have as a result of their gender, age, or other demographic characteristics) and achieved status (the status that people have as a result of their personal achievements; Ridgeway, 2001). As a group, women are associated with low status (Rudman, & Kilianski, 2000), so that there is a mismatch between the ascribed status of women and the achieved status of qualified, agentic female leaders. Likewise, men are associated with high status, so that there is a mismatch between the ascribed status of men and the achieved status of communal, modest men. Status incongruent behavior jeopardizes the gender hierarchy, and people engage in backlash as a way of restoring this hierarchy. Hence, women are proscribed from high status behavior such as being dominant, stubborn or demanding, whereas men are proscribed from low status behaviors such as being weak, uncertain, or emotional.

The Status Incongruity Hypothesis builds on System Justification Theory (Jost, Banaji, & Nosek, 2004), which suggests that people are motivated to protect and maintain existing social structures. Protecting and rationalizing these social structures serves a palliative function, namely, to satisfy people's psychological needs for order, stability, and the reduction of guilt, dissonance, and anxiety (Jost, & Hunyady, 2002). Three related predictions follow from the SIH. First of all, according to the SIH, differences in people's motivation to protect the status quo should predict

backlash. Second, system justifying motives should predict backlash against men *and* women: put differently, the same underlying motives predict backlash against both genders. Third, the SIH suggests that backlash stems from a violation of *prescriptive stereotypes* (stereotypes describing how men and women should and should not behave), because these stereotypes are strongly aligned with status (Rudman et al., 2012a). Specifically, the SIH proposes that women are not allowed to engage in high status behavior such as dominance (which is reserved for men) and men are not allowed to engage in low status behavior such as weakness (which is reserved for women). By positing that backlash stems from a violation of a specific set of stereotypes (namely, prescriptive stereotypes), the SIH proposes that not all kinds of gender atypical behavior lead to backlash. Instead, gender atypical behavior leads to backlash only if the behavior poses a threat to the status quo.

With these propositions, the Status Incongruity Hypothesis provides an extension of another theoretical account of backlash, namely Role Congruity Theory (RCT; Eagly, & Karau, 2002). According to RCT, women who aim to obtain a leadership position face two hurdles (Eagly, & Karau, 2002). First of all, they must showcase their competence to counteract the *descriptive* stereotype that women are typically less competent than men. This first hurdle is frequently described as a *lack-of-fit* between the qualities that leaders are required to have, and the qualities women are societally expected to have. Women who highlight their competence and talk about their accomplishments can successfully overcome this first hurdle, so that agentic women are perceived as equally competent as their male counterparts (for a review, see Rudman, & Phelan, 2008; Rudman et al., 2012b). However, once they have passed this hurdle, they are met with a second hurdle, which consists of the *prescriptive* stereotype prescribing that women *should* be communal and *cannot* be dominant. As a result of this second hurdle, competent and qualified women are often met with backlash: they are disliked because they are perceived as too dominant (for a woman). Thus, aspiring female leaders are either regarded as insufficiently competent (if they show stereotypically feminine behavior) or as insufficiently nice (if they show stereotypically masculine behavior).

Like Role Congruity Theory (Eagly, & Karau, 2002), the Status Incongruity Hypothesis suggests that prescriptive stereotype violations can

result in backlash, but RCT and SIH highlight different reasons as to why this is the case. As such, the SIH extends RCT in two important ways. First of all, the SIH provides an integrative theory aimed at explaining backlash against *both* genders, while RCT focuses exclusively on backlash against female leaders. Second, the SIH provides a motivational account for backlash by suggesting that system justifying motives underlie the penalization of gender deviants. Specifically, the SIH proposes that status incongruent behavior jeopardizes the gender hierarchy, and people engage in backlash as a way of protecting the status quo. In contrast to this motivational account of backlash, RCT provides a more cognitive account of backlash. According to RCT, gender stereotypes may invoke a contrast effect, such that gender atypical behaviors are perceived more negatively *because* they are unexpected (Eagly, Makhijani, & Klonsky, 1992). In this view, dominant women are perceived as highly dominant because people draw more extreme inferences from unexpected, atypical behaviors (cf. Kelley, & Michela, 1980). In contrast, the SIH suggests that agentic women are perceived as particularly dominant because dominance is status incongruent for women and, therefore, this behavior jeopardizes the status quo. The SIH makes unique predictions that do not follow from RCT, namely, that 1) backlash should be exacerbated if people are particularly motivated to protect the status quo; 2) this should be the case for backlash against both genders and 3) backlash should be exacerbated if behavior constitutes a proscriptive stereotype violation (i.e., it is not merely unexpected, but also undesirable). This is in contrast to RCT, which proposes that expectancy, not status congruity, is key in backlash.

In the present dissertation, I will explore some key predictions of the SIH. Amongst others, I will study if negative responses to gender atypical men and women are exacerbated when people are motivated to protect the gender status quo (Chapters 2, 3, and 4) and explore if people respond differently to prescriptive and descriptive stereotype violations (Chapters 3 and 5). Before further describing these studies, this introduction will continue with a somewhat more elaborate review of backlash research and the SIH.

Backlash Against Women

Women who strive for a managerial position face a Catch-22: they need to behave agentically to prove that they are sufficiently competent for the job, but are disliked if they do. Consequently, they are less likely to be hired for leadership positions (for reviews, see Eagly, & Karau, 2002; Rudman, & Phelan, 2008; Rudman et al., 2012b). One reason why agentic women are liked less than their male counterparts is because they are perceived as excessively dominant relative to agentic men (the *dominance penalty*; Eagly et al., 1992; Rudman et al., 2012a). As a result of this, agentic female job applicants are generally rated as less hireable for a managerial job than their male counterparts, even if they show the exact same behavior (for reviews, see Eagly, & Karau, 2002; Rudman, & Phelan, 2008; Rudman et al., 2012b). Moreover, evaluators are likely to shift hiring criteria to match women's deficits, such that likeability is regarded as the most important hiring criterion for agentic women, but not for men or communal women (Phelan, Moss-Racusin, & Rudman, 2008). If women do get hired, they are faced with new obstacles. Subtle signs of disapproval are visible in the nonverbal behavior of subordinates (Butler, & Geis, 1990), and people subtly frown when encountering agentic women (Carranza, 2004). Competent female leaders may be sabotaged (Rudman et al., 2012a, Study 5), and compared to their male counterparts, they are less likely to receive promotions (Heilman, 2001). In sum, female gender vanguards are faced with career roadblocks that limit their chances of obtaining leadership positions, as well as their chances of succeeding in them.

On a more positive note, it is possible for women to circumvent backlash, and women who carefully combine agency with communality may have the same chances of being hired as their male counterparts (Heilman, & Okimoto, 2007; Rudman, & Glick, 2001). Women who succeed in balancing agency and communality in this way may develop an inclusive leadership style (*transformational leadership*) that is highly effective. This leadership style is more common amongst female than male managers, suggesting that learning to successfully circumvent backlash may help women to become more effective leaders (Eagly, Johannesen-Schmidt, & van Engen, 2003). Unfortunately, mixing agency and communion constitutes a difficult balancing act for women because highly competent women (unlike less competent women) are penalized for even the slightest

hint of agency (Rudman et al., 2012a, Study 5). As such, backlash forces many women to choose between being liked (when behaving communally) and being respected (when behaving agentically), a choice not faced by male job applicants. Men, however, may be faced with negative consequences of backlash in different contexts.

Backlash Against Men

Like women, men may face negative repercussions for engaging in gender atypical behaviors. For example, men may be regarded as "wimpy" if they succeed in traditionally feminine tasks (Heilman, & Wallen, 2010), and they may face social and economic sanctions if they take time off work to care for a sick child (Rudman, & Mescher, 2013). In employment interviews, communal men are liked less than communal women, although there is no evidence suggesting that this relative dislike is reflected in lower hireability ratings for men (Moss-Racusin, Phelan, & Rudman, 2010). Moreover, men may be disliked if they behave communally, but unlike women, they can avoid backlash by behaving in a traditionally masculine way (i.e., agentically). Additionally, there is evidence suggesting that men are more likely than women to get ahead in traditionally feminine occupations (Crocker, & McGraw, 1984), probably thanks to the fact that masculinity is associated with status (Banaji, & Hardin, 1996; Rudman, & Goodwin, 2004). As such, backlash against men does not seem to negatively impact men's chances of being hired, and, perhaps as a result of this, researchers in social and organizational psychology have largely ignored backlash against men as a problem worthy of being studied.

Although backlash may not negatively affect men's chances of being hired, it may still affect men's well-being in other domains of life. Interestingly, although social psychologists have paid relatively little attention to studying backlash against men, research in developmental psychology suggests that gender atypical behavior is strongly sanctioned in boys. Feminine boys run the risk of being bullied, assaulted, and casted off as "sissies", resulting in high levels of psychological distress among these boys (Frosh, Phoenix, & Pattman, 2003; Phoenix, Frosh, & Pattman, 2003; Haldeman, 2000; Young, & Sweeting, 2004). Moreover, parents perceive cross-sexed behavior as more negative in boys than girls, worrying that gender atypical behavior in 5-year old boys is a sign of psychological

maladjustment and latent homosexuality (Martin, 1990; Sandnabba, & Ahlberg, 1999). Perhaps as a consequence, boys are over six times more likely than girls to be diagnosed with a gender identity disorder (Zucker, & Bradley, 1995). Because feminine behavior is associated with low status, several scholars have argued that it is difficult for parents to understand why boys would choose to voluntarily engage in such behaviors, unless there is something wrong with them (Feinman, 1981; Haldeman, 2000). In this view, gender atypical behavior may be more acceptable for girls than boys because stereotypically masculine (but not stereotypically feminine) behaviors are associated with desirable consequences such as status. Perhaps as a result of people's negative responses, gender atypical behavior is much less common in boys than girls (Sandberg, Meyer-Bahlburg, Ehrjardt, & Yager, 1993), suggesting that boys may be pressured to refrain from showing gender atypical behavior from an early age on.

In sum, research in developmental psychology suggests that backlash can have severe negative consequences for boys' emotional well-being and may strongly limit the behavioral options that are available to them. In contrast to the developmental literature, the literature in social and organizational psychology has largely ignored backlash against men because it does not seem to negatively affect men's chances on the labor market.

System Justification and Backlash

Unlike earlier theoretical accounts of backlash (e.g., Eagly, & Karau, 2002; Rudman, & Glick, 2001), the Status Incongruity Hypothesis presents an integrative theory that aims to uncover the motivational underpinnings of backlash against both genders. Backlash towards men and women may take different forms and may appear in different domains of life (e.g., in the workplace or elsewhere), but according to the SIH, they stem from the same underlying motive. The SIH predicts that people penalize communal men for the same reasons as they penalize agentic women, namely, as a way of putting them "back in their place" to protect the gender status quo. Women who engage in high status behavior threaten the gender hierarchy and may be regarded as usurping men's power. In a similar vein, men who engage in low status behavior threaten the status quo because men's high status position in society is legitimized by their ostensibly superior leadership skills. Thus, weak men compromise the very foundation on

which the gender status quo is build, namely, the belief that men legitimately have more power than women because women are too weak to lead. Backlash, then, serves to penalize gender deviants as a way of defending male hegemony.

System Justification Theory (SJT; Jost, Banaji, & Nosek, 2004) suggests that people have a strong motivation to defend existing societal hierarchies (such as the gender hierarchy) and to perceive them as legitimate and fair. Although people may not necessarily be conscious of their motivation to protect and legitimize the status quo, system justifying motives play an important role in shaping people's behavior (for a review, see Jost et al., 2004). Rationalizing the legitimacy of social structures (like the gender hierarchy) serves to reduce anxiety, cognitive dissonance, discomfort and guilt (Jost, & Hunyady, 2004). Interestingly, even members of disadvantaged groups (e.g., women, African Americans) are motivated to legitimize and protect the status quo. In fact, they may sometimes even be more motivated than members of advantaged groups (e.g., men, European Americans) to rationalize the very system that disadvantages them because they have a stronger need for dissonance reduction (Jost, Pelham, Sheldon, & Sullivan, 2002; Jost et al., 2004). Thus, both men and women may be motivated to protect the gender status quo, as doing so serves important palliative functions.

System Justification Theory (Jost et al., 2004) posits that people are generally motivated to defend and legitimize existing social structures, but there are individual differences in the strength of people's system justifying-motives, with some people being more motivated to protect the social system than others (Jost, & Kay, 2005; Kay, & Jost, 2003). The SIH predicts that people will be more likely to engage in backlash if they are more motivated to protect the status quo. Therefore, individual differences in people's need to protect the gender status quo, as measured with the Gender System Justification Beliefs Scale (GSJB-scale; Jost, & Kay, 2005), are expected to predict backlash. In Chapter 2 of the present dissertation, I will study if individual differences in GSJB are related to backlash against agentic female job applicants. In Chapter 4, I will additionally study if GSJB are related to memory for gender deviant behavior (which may be a precursor of backlash).

Interestingly, the motivation to protect a specific system (e.g., the gender status quo) can be temporarily heightened or lowered when an unrelated system (e.g., the hierarchy between different countries in the world) is threatened or reaffirmed (Kay et al., 2009). Next to individual differences in people's motivation to protect the gender status quo, I expect that a system threat-manipulation in which people read about the decline of their economy will increase backlash. Indeed, research suggests that a system threat-manipulation increases backlash against agentic female job applicants (Rudman et al., 2012a, Study 4). In the present dissertation, I will extend this research by studying if a system threat-manipulation affects backlash against atypical men (Chapter 2) and memory for gender deviant behavior (Chapter 3).

In addition to predicting that system justifying beliefs underlie backlash for gender atypicality, the Status Incongruity Hypothesis specifies which gender stereotypes are culpable in backlash. Specifically, the SIH suggests that people do not penalize gender atypical behavior because it is atypical (i.e., unexpected), but because it is considered to be a threat to the existing social structure. Thus, backlash effects should be most pronounced for stereotype violations that violate rules prescribing how men and women should (not) behave (i.e., *prescriptive* stereotypes), because these stereotypes are aligned with status. In Chapter 5, I will study how these prescriptive stereotypes affect the formation of trait inferences. And in Chapter 3, I present a first experiment aimed at exploring whether the prescriptive nature of gender stereotypes is key in backlash. This prediction of the SIH will be discussed next.

Descriptive and Prescriptive Stereotypes

Gender stereotypes typically consist of two components: a component describing how men and women are typically expected to behave (the *descriptive* component of gender stereotypes), as well as a component prescribing norms about how men and women should behave (the *prescriptive* component of gender stereotypes; Burgess, & Borgida, 1999; Prentice, & Carranza, 2002; 2004; Rudman et al., 2012a). Although almost all gender stereotypes are descriptive in nature, the extent to which they also contain a prescriptive component differs: stereotypes that are almost exclusively descriptive in nature are called *descriptive stereotypes*, stereotypes

that contain a large prescriptive component in addition to a descriptive component are called *prescriptive stereotypes*. An example of a descriptive stereotype is the stereotype that men enjoy watching sports (but not shopping), while women enjoy shopping (but not watching sports). Because this stereotype specifies how men and women are *expected* to behave, but not how they *should* behave, people may be surprised when seeing a female sports enthusiast or male shopaholic, but they are unlikely to respond with anger, disgust, or moral outrage (Rudman, & Glick, 2008). As such, descriptive stereotypes are *gender expectations*.

In addition to the descriptive component of gender stereotypes, many (but not all) gender stereotypes additionally contain a prescriptive component[1]. Prescriptions are *gender rules* describing how men or women *should* and s*hould not* behave. Men, for example, should be a little aggressive and assertive, while women should be warm and kind (Rudman et al., 2012a). Proscriptions are rules describing how men or women should *not* behave. Men, for example, should not be weak and emotional, while women should not be dominating or arrogant. Put differently, proscriptions are characteristics that men and women are not allowed to have. Table 1 contains an overview of the main features of the descriptive and prescriptive components of gender stereotypes. Both *gender expectations* (descriptive stereotypes) and *gender rules* (prescriptive and proscriptive stereotypes) can contribute to gender inequality, but do so through different processes, and in different situations. These processes will be described next.

Descriptive stereotypes and lack-of-fit

The descriptive component of gender stereotypes contributes to gender inequality because there is a perceived mismatch between women's qualities and the qualities that are required for success in traditionally masculine occupations. The descriptive component of stereotypes contains expectations about how men and women typically behave, such that men are expected to be agentic (e.g., intelligent, hard-working, ambitious), while women are expected to be communal (e.g., kind, warm, interested in

[1] Some prescriptive stereotypes do not have a clear descriptive component. Because these cases are rare, they are not the focus of the present dissertation.

children; Prentice, & Carranza, 2002). According to Role Congruity Theory, there is a lack-of-fit between the qualities expected from women and the qualities required from leaders (Eagly, & Karau, 2002; Heilman, 2001). When people think of a successful manager, they are more likely to think of a man (Heilman, Block, & Martell, 1995), suggesting that the stereotypical qualities of men largely overlap with the qualities that are valued in managers.

Table 1. Overview of the main differences between descriptive and prescriptive stereotypes.

descriptive stereotypes (gender expectations)	**prescriptive stereotypes (gender norms)**
• stipulate *expectations* about how men and women typically behave.	• stipulate *norms* about how men and women should (not) behave.
• positive and negative descriptions are not theoretically distinguished.	• the overarching term for positive and negative rules is *prescription*, but negative rules (the traits *forbidden* for men or women) are sometimes referred to as *proscriptions*.
• violation is perceived neutrally or positively.	• violation is perceived negatively.
• may contribute to gender inequality through lack-of-fit.	• may contribute to gender inequality through backlash.
• (almost) all gender stereotypes have a descriptive component.	• only some gender stereotypes have a prescriptive component (in addition to a descriptive component).
• the content of descriptive stereotypes has shown considerable change over time.	• the content of prescriptive stereotypes is largely resistant to change.

However, qualified, competent women can successfully defy descriptive stereotypes by self-promoting (Eagly, & Karau, 2002; for overviews, see Rudman, & Phelan, 2008; Rudman et al., 2012b). People generally respond positively to descriptive stereotype violations (Gill, 2004; Prentice, & Carranza, 2004), and these stereotypes have shown considerable change overtime, reflecting men's and women's changing social roles (Diekman, Goodfriend, & Goodwin, 2004). Thus, descriptive stereotypes are an obstacle to gender parity, but their influence may be limited because it is possible for women to disconfirm these stereotypes.

Prescriptive stereotypes and backlash

Prescriptive stereotypes stipulate gender rules prescribing which behaviors are (not) acceptable for men and women. As such, they serve to protect and maintain the gender status quo, and their violation is strongly policed[2]. Descriptive stereotype violations are unlikely to be perceived negatively: for example, people are unlikely to be outraged if a man is perfectionistic, or if a man indicates that he enjoys shopping. Likewise, people are unlikely to be outraged if a woman is lazy or indicates that she loves watching sports. Proscriptive stereotype violations, however, are likely to elicit strongly negative responses. For example, people will likely respond negatively to a man who is weak, complains when he breaks a nail, or starts screaming when he sees a mouse. Likewise, they are likely to respond negatively to a woman who is dominant, hits the table with her fist, or boasts about the number of sex partners she has had.

According to the Status Incongruity Hypothesis, the violation of proscriptive stereotypes is met with social sanctions because this behavior threatens the status quo. Men and women who engage in proscriptive stereotype violations enact status incongruent behavior, and backlash serves to put them "back in their place". By specifying which stereotypes are

[2] Some researchers distinguish between two types of prescriptive stereotypes: proscriptions (*negative* gender rules, as described above) and intensified prescriptions, which are *positive* gender rules that describe how men and women are required to behave (e.g., women are required to be interested in children; Prentice, & Carranza, 2002; 2004). Because the present dissertation is focused on backlash, I focus on proscriptions: for an overview of how intensified prescriptions could play a role in maintaining gender equality, please see Prentice, & Carranza (2002).

culpable in backlash, the SIH suggests that not all gender atypical behaviors should be met with strongly negative responses. This is important, as it has been suggested that people may be penalized for behaving in *any* way that is atypical for their gender (e.g., Alsop, Fitzsimons, & Lennon, 2002; Eagly, & Karau, 2002). The assumption that atypicality is key in backlash is apparent in the term "backlash for gender *atypicality*". The SIH proposes that backlash does not result from atypicality alone, suggesting that not all atypical behavior is out of bounds for men and women. Instead, men and women may only be penalized for behaviors that violate proscriptive stereotypes. Thus, contrary to what researchers thought when the term *backlash for gender atypicality* was introduced (Rudman, 1998; Rudman, & Glick, 1999), backlash may not be due to the atypicality of behaviors, but to their status incongruity (Rudman et al., 2012a).

Unlike descriptive stereotypes, prescriptive stereotypes are highly resistant to change, which may contribute to the continuing existence of gender inequality (Gill, 2004; Prentice, & Carranza, 2004). There are two reasons for this stability. First of all, gender rule violations are rare because people go to great lengths to avoid violating proscriptive stereotypes, as a way of avoiding backlash (Moss-Racusin, & Rudman, 2010; Rudman, & Fairchild, 2004). Through this process, proscriptive stereotypes shape behavior instead of following from it (which is the case for descriptive stereotypes). Second, because gender rules are strongly rooted in ideology, people resist changing them even if they are violated (Gill, 2004). Prentice and Carranza (2004) suggest that gender rules cannot be empirically disconfirmed because they reflect ideologies, not empirical facts. Due to their resistance to change and their role in causing backlash, proscriptive stereotypes may form an important roadblock for reaching gender parity.

The relationship between descriptive and proscriptive components of stereotypes

To illustrate the kinds of traits that are proscribed for men and women, Table 2 contains a list of proscriptions (taken from Rudman et al., 2012a). As is apparent from this table, proscriptions for women consist of traits such as being *aggressive*, *intimidating*, and *dominating*, constituting a negative, extreme version of agency (i.e., dominance).

Table 2. Traits that are proscribed for men and women (taken from Rudman et al., 2012a).

trait	typicality *d*	desirability *d*	status *d*
Men's proscriptions			
Emotional	-1.49	-1.12	-0.63
Naïve	-0.88	-1.03	-0.78
Weak	-1.02	-0.97	-1.32
Insecure	-1.06	-0.91	-0.96
Gullible	-1.02	-0.89	-1.07
Melodramatic	-1.22	-0.88	-0.01
Uncertain	-1.22	-0.80	-1.22
Moody	-0.71	-0.78	0.05
Superstitious	-0.74	-0.56	-0.64
Average	**-1.04**	**-0.88**	**-0.73**
Women's proscriptions			
Aggressive	0.43	1.03	1.36
Intimidating	0.89	0.98	1.21
Dominating	0.97	0.94	1.42
Arrogant	1.11	0.76	1.08
Rebellious	0.66	0.69	-0.40
Demanding	-0.15	0.65	1.24
Ruthless	0.64	0.65	0.59
Angry	0.71	0.65	-0.47
Controlling	0.42	0.61	1.33
Stubborn	0.42	0.55	0.65
Cold toward others	0.38	0.51	0.35
Self-centered	0.14	0.41	1.05
Cynical	0.28	0.41	0.12
Average	**0.53**	**0.68**	**0.73**

Note. Positive *d*-scores for typicality and desirability reflect stronger typicality or desirability for men than women, negative d-scores reflect stronger typicality or desirability for women than men. Positive effect sizes for status reflect high status, negative effect sizes reflect low status. By convention, small, medium and large effect sizes correspond to Cohen's ds of 0.20, 0.50 and 0.80, respectively; Cohen, 1988.

Proscriptions for men consist of traits such as *emotional*, *naïve*, and *weak*, constituting a negative, extreme version of communality (i.e., weakness).

The traits in Table 2 are accompanied by effect sizes describing the relative expectancy and desirability of these traits for men versus women. To determine these effect sizes, four separate groups of American[3] participants were asked to indicate how desirable specific traits are for men, how desirable they are for women, how typical they are for men, or how typical they are for women (see Rudman et al., 2012a, Study 1). The reported Cohen's *d*s reflect the effect sizes for the differences in desirability and expectancy for men versus women. As is apparent from Table 2, expectancy and desirability are correlated, so that negative traits and behaviors that are proscribed for a gender are often also considered atypical for that gender. Put differently, most proscriptions have both a descriptive component (i.e., they are more expected or typical for one gender or the other) as well as a prescriptive component (they are more or less desirable for one gender or the other). Being weak, for example, is considered as more undesirable for men than women, and it is also rated as less typical for men than women. Depending on the specific traits or behaviors that are tested, correlations between expectancy and desirability ranged from r = 0.34 to r = 0.87 (Nauts et al., unpublished; Rudman et al., 2012a). One reason as to why expectancy and desirability are correlated may be because people refrain from showing socially undesirable behavior for fear of backlash.

Prescriptive stereotypes and status

Next to the effect sizes for the relative differences in expectancy and desirability, Table 2 contains effect sizes for the relative status of each trait. This effect size was calculated by asking participants in a pretest to indicate the extent to which traits are associated with high or low status. As apparent from Table 2, proscriptions for women are generally high in

[3]Because research on gender prescriptions so far has been conducted using American participants (Prentice, & Carranza, 2002; Rudman et al., 2012a), it is unclear if these traits and behaviors are proscribed for men and women in countries other than the US. In the present dissertation, I will employ both Dutch and American samples of participants, but pretest stimuli in each country to ensure that the used behaviors are proscribed in both countries.

status, whereas proscriptions for men are generally low in status. This suggests that men are not allowed to portray low status behaviors, while women are not allowed to portray high status behaviors. Because men are generally considered to have more societal status than women, low status behavior is status incongruent for men, whereas high status behavior is status incongruent for women. Thus, the gender rules seem strongly aligned with status, so that what men *should be* is high in status and what they *shouldn't be* is low in status. What women *should be* is neutral or low in status and what they *shouldn't be* is high in status.

The Status Incongruity Hypothesis posits that status incongruent behavior threatens the gender status quo, and that backlash serves to penalize status incongruent behavior as a way of protecting the gender status quo. In so doing, the SIH specifies which stereotypes are culpable in backlash. Specifically, the SIH predicts that backlash results from gender rule violations, not from the violation of stereotypes that are unrelated to status (e.g., descriptive stereotypes or communality prescriptions for women). In line with the SIH, research by Rudman and colleagues (2012a; Study 4) suggests that backlash against agentic women is predicted by the dominance penalty, but not by differential ratings on traits that are not status incongruent for women (e.g., communality). Likewise, research by Moss-Racusin and Rudman (2010) suggests that backlash against men is predicted by a weakness penalty, but not by differential ratings on traits that are not status incongruent for men. In line with the SIH, these results suggest that backlash stems from a perceived violation of prescriptive stereotypes, thereby pinpointing exactly which types of traits and behaviors are likely to yield backlash.

In the present dissertation, I will extend this research by studying if proscriptive stereotypes uniquely predict backlash. Although these studies do not definitively answer the question whether backlash results from a violation of proscriptive stereotypes, they may provide some information about this premise. In Chapter 3, I study if backlash against men is more pronounced if stereotypes are proscriptive in nature, rather than purely descriptive. In so doing, I aim to establish which stereotypes contribute to backlash. Moreover, in Chapter 5, I study if proscriptive and descriptive gender stereotypes differentially affect the formation of spontaneous trait inferences (STIs). Spontaneous trait inferences are the impressions people

form of others without intention or awareness (Uleman, Newman, & Moskowitz, 1996). STI-formation can be biased by the stereotypical expectancies perceivers hold about a target (Wigboldus, Dijksterhuis, & Van Knippenberg, 2003; Wigboldus, Sherman, Franzese, & Van Knippenberg, 2004; Yan, Wang, & Zhang, 2012). Because STIs are formed without intention, biased STIs may be an important source of backlash that is difficult to control.

Overview of the Present Research

In the present dissertation, I present four empirical chapters on backlash and the Status Incongruity Hypothesis. In each of these chapters, I will present research in which backlash, or a possible antecedent of backlash, is tested using a novel methodological approach. Most backlash research has been conducted using videotaped employment interviews with agentic and communal male and female confederates (for an overview, see Rudman, & Phelan, 2008; Rudman et al., 2012b). This approach has yielded important insights, but it has its limitations. First of all, the ecological validity of this paradigm is limited. Second, the employment interview paradigm may not be optimally suitable to study backlash against men, as men may be penalized for different behaviors, and in different contexts, than women. By introducing novel paradigms to study backlash, I aim to address both of these concerns.

The first concern (low ecological validity) is addressed in Chapter 2, in which I present results of a study in which participants conducted live interviews with confederates. In this study, participants did not merely observe the interview (as they do in the classic backlash-paradigm) but interviewed a job applicant themselves. In this study, participants first completed a Gender System Justification Beliefs scale (GSJB; Jost, & Kay, 2005). Weeks later, they conducted a live phone interview with a male or female applicant who allegedly applied for a managerial position, and were asked to rate this applicant on indices of hireability, likeability, and dominance. Based on the Status Incongruity Hypothesis, I expected that people with higher GSJB-scores would rate agentic women as relatively less hireable and likeable. Moreover, I expected that women would be rated as relatively dominant (a *dominance penalty*) compared to men, but not as insufficiently communal (a *communality deficit*), in line with the SIH's

prediction that women should be penalized for status violations. The goal of this first chapter was to find evidence for the SIH's contention that system justifying motives exacerbate backlash, as well as to find evidence for backlash in a live employment interview.

The second concern (the unsuitability of classic approaches for testing backlash against men) was addressed in Chapters 3, 4 and 5. In these chapters, instead of using an employment interview paradigm, I investigated people's responses to a range of stereotype violating behaviors, such as scared, nervous and shy behavior for men and dominant, aggressive and rude behavior for women. While classic backlash-paradigms focus on a very limited number of behaviors (i.e., agentic or communal behavior in an employment interview), the novel approaches presented in this dissertation allow researchers to study a wider range of behaviors. Additionally, in the present dissertation, I tried to complement the direct, explicit measures of penalization that are used in classic backlash research with more indirect measures of backlash, such as a measure based on the mental representations that people formed of gender deviant male's faces (Chapter 3). Moreover, I studied processes that may contribute to people's relatively negative responses to gender deviants, such as memory for gender deviant behavior (Chapter 4) and spontaneous trait inferences (Chapter 5). Over the course of this dissertation, I will first present direct and explicit measures of backlash and gradually move to more indirect and spontaneous measures of (possible antecedents of) backlash.

Chapter 3 investigated mental representations of gender deviant men as a subtle and indirect measure of backlash. Specifically, I studied if, after a system threat prime, people formed mental representations of nervous/scared men as being weaker and more negative. In one study, I additionally tested if a system threat manipulation affected responses to men who violated descriptive stereotypes (i.e., perfectionistic/clumsy men). To study people's mental images of gender deviant men, I used a Reverse Correlation Image Classification Task (RCIC; Dotsch et al., 2008; Mangini, & Biederman, 2004), as well as a new task that was developed specifically for the present research, the Draw-a-Face-Task (DaFT). These measures allowed me to explore people's spontaneous inferences of gender deviant behavior in a data-driven fashion.

In Chapter 4, I studied a possible precursor of backlash, namely, people's memory for gender atypical behavior. I tested if individual differences in people's motivation to protect the gender status quo predicted their memory for proscriptive stereotype violations relative to stereotypical behaviors. I also tested if, after a system threat prime, participants showed better recall for proscriptive stereotype violations (relative to neutral behaviors). Memory is a possible antecedent of the penalization of gender atypical targets, and I expected that people who are motivated to protect the status quo would be more likely to remember behaviors that threaten the status quo (i.e., gender deviant behaviors).

In Chapter 5, I tested another possible antecedent of backlash, namely, spontaneous trait inferences (STIs). Because STIs are formed without intention (Uleman, Newman, & Moskowitz, 1996), they may be a particularly potent source of backlash. Previous research has suggested that stereotypes can affect the formation of spontaneous trait inferences (Wigboldus et al., 2003; 2004), but the results of this research may not be applicable to backlash, as the stereotypes that were employed in previous research did not have a clear proscriptive component. In Chapter 5, I studied if people spontaneously form stronger inferences of proscriptive (but not descriptive) stereotype violations (e.g., scared men, dominant women). If people form stronger STIs based on this particular type of gender deviant behavior, this may be a precursor of prejudiced responses to gender deviants.

In sum, the goal of the present dissertation is to study why people engage in backlash. Specifically, I aim to investigate if backlash is exacerbated when people are motivated to protect the status quo, which would suggest that system justifying motives may underlie backlash. Throughout four empirical chapters[4], I will introduce novel research paradigms to study backlash and possible precursors of backlash.

[4] These chapters were written as separate journal articles, and may therefore show some overlap.

CHAPTER 2

System Justification and Backlash against Agentic Women

This chapter is based on Study 3 of:
Rudman, L.A., Moss-Racusin, C.A., Phelan, J.E., & Nauts, S. (2012). Status incongruity and backlash effects: Defending the gender hierarchy motivates prejudice against female leaders, *Journal of Experimental Social Psychology, 48, 165-179.*

"For when a woman is strong, she is strident.
If a man is strong, he's a good guy."
Margaret Thatcher

Women who apply for a leadership position face a difficult Catch-22: they need to behave agentically in order to prove that they are sufficiently competent for the job, but are disliked if they do (Rudman, 1998; Rudman, & Glick, 2001). As a result, women who apply for leadership positions may be less likely to be hired than their male counterparts (for reviews, see Rudman, & Phelan, 2008; Rudman, Moss-Racusin, Glick, & Phelan, 2012b), and may be sabotaged even if they do get hired (Rudman, Moss-Racusin, Phelan, & Nauts, 2012a, Study 5). These negative responses to female agency (termed *backlash for gender atypicality*) constitute a major impediment for women who aspire to obtain positions of power. Due to backlash, female job applicants are often forced to choose between being liked (if they show communal behavior) or being perceived as competent (if they show agentic behavior; Rudman, & Glick, 2008).

According to the Status Incongruity Hypothesis (SIH; Rudman et al., 2012a) backlash serves to preserve the gender status quo. The SIH builds on System Justification Theory (SJT; Jost, Banaji, & Nosek, 2004), which posits that people are motivated to protect existing social structures in order to satisfy psychological needs for certainty and stability (Jost, & Hunyady, 2002). According to the SIH, agentic women jeopardize the gender hierarchy by portraying high status behaviors (such as dominance) that are reserved for leaders and men. By penalizing agentic women, they are "put back in their place" as a way of defending the gender status quo. Put differently, the SIH postulates that backlash functions to defend male hegemony by discouraging women from obtaining high status positions. If system justifying motives indeed underlie prejudiced responses towards agentic women, individual differences in people's motivation to protect the gender status quo should predict backlash.

People do not engage in backlash arbitrarily, but only if they feel there is a justification for it (Rudman, & Fairchild, 2004). One way to justify why an agentic woman is not hired is by casting her off as "too dominant" (termed the *dominance penalty*; Eagly, Makhijani, & Klonsky, 1992). The dominance penalty is apparent in the epithets that are used to refer to

powerful women (e.g., "castrating bitch"; Kanter, 1977), as well as in the landmark case against Ann Hopkins, who was denied promotion because she was perceived as too masculine and dominant (Price Waterhouse v. Hopkins; Fiske, Bersoff, Borgida, Deaux, & Heilman, 1991). Recent research indicates that the dominance penalty fully mediates backlash, suggesting that agentic women are disliked *because* they are perceived as too dominant (Rudman et al., 2012a, Study 2). According to the Status Incongruity Hypothesis, the dominance penalty is key in backlash against women, because dominance is a *proscription* for women (i.e., it is a gender rule describing how women are not allowed to behave; Prentice, & Carranza, 2002; 2004; Rudman et al., 2012a, Study 1). Because dominance is not accepted in women, casting a woman off as dominant is sufficient justification for not hiring her.

Women seem to be proscribed from portraying dominant behaviors because dominance is a high status characteristic that is *status incongruent* with women's low status in society (Rudman et al., 2012a, Study 1). As such, dominant women are perceived as a threat to the gender status quo. The Status Incongruity Hypothesis suggests that the dominance penalty is pivotal in backlash because dominance is strongly aligned with status. Moreover, the SIH predicts that individual differences in people's motivation to protect the status quo are related to the dominance penalty. This is in contrast with Role Congruity Theory (RCT; Eagly, & Karau, 2002), which posits that backlash is the result of a perceived *communality deficit*, in that agentic women are regarded as insufficiently communal, friendly and modest. RCT proposes that this communality deficit (as well as the dominance penalty) stems from a perceptual contrast effect: because people are more likely to draw extreme inferences based on unexpected behaviors (Kelley, & Michela, 1980), they will draw more extreme inferences of women's agentic behavior, compared to men's agentic behavior. In contrast to RCT, the SIH provides a motivational account for backlash by suggesting that agentic women are disliked not because their behavior is unexpected, but because their behavior threatens the status quo. Because communality is neutral in status (Rudman et al., 2012a; Study 1), the Status Incongruity Hypothesis posits that the dominance penalty, not a communality deficit, plays a pivotal role in backlash.

To test if individual differences in people's motivation to protect the gender status quo predict backlash, participants in the present study first completed a measure of gender system justification beliefs (Jost, & Kay, 2005). Several weeks later, they interviewed a male or female confederate who responded to interview questions in an agentic, self-promoting way[1] (cf. Rudman, & Glick, 2001). After that, participants completed indices of likeability, hireability, and several traits (e.g., dominance, communality). We expected that individual differences in system justifying motives predict dominance ratings, likeability and hireability, but not communality. Moreover, we expected that the dominance penalty accounts for women's lower likeability ratings, and that women's lower likeability ratings account for their lower hireability ratings.

Method

Overview and Design

Participants interviewed a male or female confederate who was allegedly practicing for a phone interview for the position of marketing manager.

---

[1]In the original design of the study, we included another condition, namely, whether the applicant script was self-promoting or ingratiating. In this ingratiating script, ingratiating comments were added to the self-promoting script to investigate if ingratiation could soften female agency and diminish backlash effects. We choose not to report the data for this condition in the present dissertation or in Rudman et al., 2012a, because preliminary data analyses suggested that we had been unsuccessful in training confederates in this condition, as there was an effect of confederate on the dominance ratings of our female confederates, $F(1,33) = 3.44$, $p = .05$, $\eta_p^2 = .18$. Put differently, differences in ratings between confederates could not be fully attributed to their gender. As indicated by previous research (e.g., Vonk, 1998), ingratiation can easily be perceived as brown-nosing, and it can sometimes invoke a "slime-effect". We put a lot of effort into creating a script that was ingratiating, but did not invoke a "slime-effect" and elaborately trained confederates to present the ingratiating comments according to script. Nevertheless, the data suggest that not all our confederates were successful in avoiding the "slime effect", and that there were individual differences in confederates' ability to ingratiate successfully. Regardless of their gender, ingratiation seemed to "work" for some confederates while it backfired for others.

In the self-promotion condition, there were no differences in the ratings of our two male confederates, and there were no differences in ratings between our three female confederates, F's 0.05 to 2.41, *ns*. This suggests that confederates had been successfully trained to deliver the self-promoting script in the same way, and that differences in ratings between male and female confederates can likely be attributed to their gender.

Confederates answered interview questions in a highly scripted way, providing answers that were strongly agentic and self-promoting. After the interview, participants rated the alleged job applicant on indices of hireability, likeability, and traits that are, amongst others, related to communality and dominance. Participants interviewed a male or a female confederate, yielding a simple 2 (confederate gender: male or female) between subjects-design with gender system justification beliefs as continuous predictor.

Participants

Seventy-one Rutgers University participants (36 men) participated in the study in exchange for partial course credit.

Materials

Applicants. Five confederates (three women, two men[2]) acted as job applicants. All confederates (aged 22 to 26) were Caucasian and wore casual business attire during the experiment.

Job description. Job applicants allegedly applied for a position as marketing manager. A pretest (N = 40) suggested that participants estimated that approximately 63% of marketing managers are male, indicating that the position of marketing manager was considered to be slightly male-dominated. The marketing manager was described in a company advertisement as follows: "You are responsible for the formulation and execution of a marketing strategy, together with your team of eight experienced marketers. You coordinate market analyses aimed at identifying consumer needs and introduce new products and services to strengthen our position in the market." The qualifications were listed as "You have a masters degree in marketing; you have strong analytical skills and like to take the initiative; you are innovative and creative; you have excellent communication skills; and you can manage and inspire a team". The job description was designed as requiring both agentic and communal

[2]Due to scheduling issues resulting from technical problems (i.e., an electricity black-out in the greater New Brunswick-area), confederates were not interviewed by the exact same number of participants. Male confederates A. and B. were interviewed by 15 and 21 participants, respectively. Female confederates A., B. and C. were interviewed by 8, 13 and 14 participants, respectively.

qualities to reflect a feminized job description for a managerial position (cf. Rudman, & Glick, 2001).

Interview scripts. Confederates were trained to answer interview questions according to a script. The scripts contained standard interview questions (e.g., "can you name your two most important qualities and a point for improvement?") with answers that were strongly agentic and self-promoting (see Appendix 1 for excerpts).

Deviations from script. The present research employed live interviews, which has the advantage of being more ecologically valid than the standard interview videos frequently used in backlash research. A possible downside, however, is that such live interviews may provide less experimental control, as participants may create different interview environments for female versus male applicants. Though such effects may be interesting, in the present study, we aimed to provide a standardized situation to study whether the perception of the same behaviors would be different when they were performed by a female, compared to a male, job applicant. To test if there were major differences in the interview environment created by participants, two independent coders unobtrusively coded for 56% of the interviews in what order the interview questions were asked, how many acknowledgements interviewers provided (e.g., saying "uhum", "okay", or "thank you, Steve") and how friendly and respectful the interviewer sounded. Coders also kept log sheets noting any deviations from the script (e.g., participants skipping or changing questions). Inter rater reliabilities for these indices ranged from mediocre (a =.41 for respectfulness) to perfect (a = 1.00 for question order). Because interviews were coded live (as recording conversations would violate local IRB-protocols), differences in ratings could not be resolved through discussion, and averaged ratings were used for all analyses. However, the pattern of results is identical regardless of whether the ratings of Coder 1, Coder 2, or the average of both ratings is used.

Applicant ratings. Participants rated the applicant on indices of liking, hireability, competence, and several trait indices on 6-point Likert scales (anchors: 1 not at all; 6 very much). Question order was the same for all participants.

The liking index (α = .81) consisted of the following three items: "How much did you like the applicant?"; "Would you characterize this

person as someone you want to get to know better?" and "Would the applicant be popular with colleagues?". The hireability index (α = .89) consisted of the following three items: "How likely is it that you would choose to interview the applicant for the job?", "How likely is it that the applicant would be hired for the job?" and "How likely is it that you would hire the applicant for the job?". The competence index (α = .72) consisted of the following six items: "Did the applicant strike you as someone who has strong analytical skills?", "Did the applicant strike you as a self-starter?", "Did the applicant strike you as a good listener?", " Would you characterize this person as someone likely to get ahead in their career?", "Estimate the percentage of marketing problems the applicant would be able to solve independently" and "Estimate the percentage of subordinates who would feel comfortable seeking help from the applicant". The latter two questions used a 6-point scale with percentages as anchors (e.g., 0-17%). We included questions about both agentic and communal qualities in the competence index to reflect the demands described in the job advertisement.

Next to indices of likeability and hireability, we included indices of dominance and communality to test the SIH's prediction that the dominance penalty (but not a perceived communality deficit) predicts backlash.[3] These scales were based on the research by Rudman et al. (2012a; Studies 1 and 2). Participants were asked to indicate on a 6-point Likert scale to what extent the applicant struck them as someone with certain qualities. The index for communality (α = .78) consisted of the following five items: *warm, sensitive to the needs of others, supportive, cooperative, and friendly.*

The index for dominance consisted of the following six items (α = .82): *dominating, intimidating, arrogant, self-centered, manipulative* and *cold toward others.*

Gender System Justification Beliefs. Participants' motivation to protect and maintain the gender status quo was measured using a gender specific version of a system justification-questionnaire, the Gender System Justification Beliefs questionnaire (GSJB; Jost, & Kay, 2005). This scale consisted of seven items (e.g., "In general, relations between men and

[3]We included additional traits that are not relevant to the present research, and therefore, will not be discussed in the present Chapter. More information on these measures is available in Rudman et al., 2012a (Study 3).

women are just and fair" and "Society is set up so that men and women usually get what they deserve") and had reasonable reliability (a = .75). The scale was administered weeks before the experiment was conducted, as part of a large departmental pretest.

Procedure

Participants were instructed that they would interview a recent Rutgers graduate as part of an interview training project. Their role was to help this person prepare for a job interview for the position of marketing manager. Because the interview for which the applicant was practicing would be taking place over the phone, participants were told that they would also conduct the practice-interview through the phone. After giving informed consent, participants read the job description and received a stack of nine questions to ask to the applicant. Next, the confederate briefly entered the participant's cubicle, introduced him- or herself as Steven or Susan Anderson, gave participants a walkie talkie, and left again, after which participants could start the phone interview. We choose to use a phone interview using walkie talkies instead of conducting live face-to-face interviews to be better able to standardize the interview setting. Walkie talkies were chosen instead of a phone because the use of a walkie talkie made it impossible for participants to interrupt confederates (e.g., to ask additional questions) while he or she was talking, since walkie talkies are a one-way communication device. This was done to further standardize the experimental situation across participants. Because the experiment took place in a lab with poor cell phone reception, the use of walkie talkies instead of a phone did not invoke any suspicion on part of participants.

After completing the interview, participants rated the applicant on hireability, likeability, competence, and the trait indices. Finally, participants were debriefed and thanked for their participation.

Results

Deviations from Script

To test whether participants behaved differently towards female and male applicants, we tested whether there were any effects of applicant gender on question order, the number of deviations from the script, number of acknowledgements, and ratings of friendliness and respectfulness. Overall,

there were very few deviations from the script, suggesting that we succeeded in creating a setting that was relatively standardized. There were no significant effects of applicant gender on the number of times participants deviated from the script, or the order in which questions were asked (all χ^2s > 2.00, *ns*). Moreover, there were no effects of applicant gender on the number of acknowledgements that participants provided to applicants, or on how respectful or friendly they sounded (all *F*s < 1). Together, these data do not provide evidence for a differential treatment of male and female applicants in our study. This suggests that our paradigm provided a relatively controlled setting, although it is off course possible that interviews differed in ways not measured in the present study.

Applicant Ratings

To investigate if applicant evaluations differed depending on the applicant's gender, we conducted separate ANOVAs with applicant gender as independent variable and indices for likeability, hireability and competence as dependent variables. Based on previous backlash research, we expected that there would not be significant differences in competence-ratings for agentic female and male applicants, but that female applicants would be rated as significantly less likeable and hireable, as well as more dominant, than their male counterparts. As depicted in Table 1, and in line with our expectations, there was no significant effect of applicant gender on ratings of competence, $F(1, 70) = 1.33$, *ns*. However, there were effects on likeability and hireability: female applicants were rated as significantly less likeable than male applicants, $F(1,70) = 4.05$, $p < .05$, Cohen's $d = 0.43$, as well as less hireable, $F(1,70) = 5.92$, $p < .05$,Cohen's $d = 0.51$, suggesting that participants engaged in backlash against agentic women. Moreover, women were rated as significantly more dominant than men, $F(1,70) = 5.41$, $p = .02$, Cohen's $d = -0.46$, suggesting that agentic women received a *dominance penalty*. In line with prior backlash research, participant sex did not significantly interact with applicant sex or GSJB on any of the used indices, all *F*s < 1.06, all *p*s > .24. Thus, there was no evidence suggesting that male or female participants were more likely to engage in backlash against agentic women.

Based on the Status Incongruity Hypothesis, and contrary to Role Congruity Theory, we did not expect differences in ratings of male and

female applicants for communality. Indeed, we did not find evidence for a perceived communality deficit ($F < 1$).

Table 1. *Average evaluations of male and female job applicants.*

	male applicant	female applicant	Cohen's d
competence	4.88	4.75	**0.23**
SD	*0.45*	*0.61*	
liking	4.89	4.53	**0.43***
SD	*0.65*	*0.84*	
hireability	5.43	5.04	**0.51***
SD	*0.60*	*0.82*	
communality	4.16	4.15	**0.02**
SD	*0.60*	*0.71*	
dominance	3.03	3.44	**-0.46***
SD	*0.76*	*0.98*	

Note. Effect sizes (Cohen's *d*) represent applicant sex differences. By convention, small, medium, and large effect sizes correspond to Cohen's *d* of .20, .50, and .80, respectively (Cohen, 1988). Positive effect sizes indicate higher ratings for male compared to female applicants. Effect sizes with an asterix (*) refer to significant differences between male and female applicants at $p < .05$.

Gender System Justification Beliefs

To test if gender system justifiers were particularly likely to administer the dominance penalty, we conducted a regression analysis with dominance as dependent variable and gender (0 = male, 1 = female), participants' Gender System Justification-score (GSJB), and the interaction between applicant gender and GSJB as predictors. In line with our expectations, there was a significant applicant gender x GSJB interaction, $\beta = -.33$, $p < .01$. For ratings of male applicants, there was no significant relation between participants' GSJB-score and their dominance-ratings, $r(29) = -.15$, *ns*. For

ratings of female applicants, there was a positive relation between participants' GSJB and their dominance-ratings, $r(31)^4 = .46$, $p = .01$. In line with our expectations, people with a stronger motivation to protect the gender status quo gave higher dominance-ratings to agentic women, but not to agentic men. In sum, individual differences in participants' motivation to protect the status quo affected the dominance penalty for women.

Communality ratings were submitted to the same analysis, but for perceived communality, there was no significant interaction between applicant gender and individual differences in people's motivation to protect the status quo ($\beta = .13$, *ns*). Although null results should always be interpreted with caution, these results are in line with the SIH, which posits that gender system justifiers should penalize agentic women with a dominance penalty, but not with a communality deficit.

Subsequently, we tested if gender system justification beliefs affected hireability ratings for female applicants. In line with our expectations, there was a significant applicant gender x GSJB interaction, $\beta = -.30$, $p < .05$. For ratings of male applicants, there was no relation between their GSJB-score and ratings on hireability, $r(29) = 0.14$, *ns*. For ratings of female applicants, there was a significant relation between participants' GSJB-score and ratings on hireability, $r(31) = -0.38$, $p < .05$. In sum, people with a higher motivation to protect the gender status quo gave lower hireability ratings to agentic women, but not to agentic men.

Finally, we tested if gender system justification beliefs affected the likeability of female applicants. In line with our expectations, there seemed to be a GSJB x applicant gender interaction, but this interaction was only marginally significant, $\beta = -.23$, $p < .08$. Contrary to our predictions, for ratings of male applicants, there was a positive relation between applicant's GSJB-score and ratings on likeability, $r(29) = 0.35$, $p = .06$. For ratings of female applicants, there was no relation between participants' GSJB-score and ratings on likeability, $r(31) = -.18$, $p = .32$. In sum, contrary to our expectations, we did not find evidence suggesting that individual differences in system justification beliefs predicted liking of self-promoting female applicants.

---

[4] Degrees of freedom differ across analyses due to missing data for GSJB.

Accounting for Backlash Against Female Leaders

In prior backlash research, differences in hireability between agentic men and women were fully mediated by differences in liking for agentic men and women, suggesting that agentic women were less likely to be hired *because* they were not liked (Rudman, & Glick, 1999; 2001). To test if this was also the case in the present study, we used PRODCLIN to compute confidence intervals based on an asymmetrical distribution of the mediated (indirect) effect (MacKinnon, Fritz, Williams, & Lockwood, 2007). We also tested if dominance and communality mediated liking. According to the SIH, dominance, but not communality, should mediate liking.

Table 2. *Mediation analyses for marketing manager applicants, including Standard Errors (SE) and 95% Confidence Intervals (95% CI).*

Path/effect	B	SE	95% CI
Model 1			
c (applicant gender → hire)	-.29*	.12	
a (applicant gender → like)	-.24*	.11	
b (like → hire)	.58***	.10	
c'	-.15	.10	
a x b (mediation effect)	-.14*	.07	-.283, -.015
Model 2			
c (applicant gender → like)	-.24*	.11	
a (applicant gender → dominant)	.20*	.10	
b (dominant → like)	-35***	.12	
c'	-.18	.11	
a x b (mediation effect)	-.07*	.04	-.167, -.002
Model 3			
c (applicant gender → like)	-.24*	.11	
a (applicant gender → communal)	-.01	.10	
b (communal → like)	.80***	.09	
c'	-.23*	.08	
a x b (mediation effect)	-.01	.08	-.166, .149

Note. Applicant gender was coded as 0 (*male*) 1 (*female*). Estimates are unstandardized. Confidence intervals for $a \times b$ are based on an asymmetrical distribution. Intervals that do not include zero support rejecting the null hypothesis that $a \times b = 0$. *$p < .05$; ** $p < .0$; ***$p < .001$.

As depicted in Table 2 (Model 1), the effect of applicant gender on hireability was reduced to nonsignificance after accounting for liking, and the 95% confidence interval did not include zero. This suggests that liking fully mediated hireability, replicating prior backlash research. Moreover, the effect of applicant gender on liking also reduced to nonsignificance after accounting for dominance, and the 95% confidence interval did not include zero. This suggests that dominance fully mediated liking, so that agentic women were disliked *because* they were perceived as too dominant. Communality did not mediate the effect of applicant gender on liking (Model 3): the 95% confidence for the mediated effect included zero. In sum, mediation analyses suggest that liking was mediated by dominance, but not communality.

Discussion

Replicating prior backlash research, the present study suggests that agentic female job applicants were regarded as less likeable and hireable than their male counterparts, and that this effect was mediated by the dominance penalty. Importantly, individual differences in people's motivation to protect the gender status quo predicted hireability ratings and the dominance penalty (but, contrary to our expectations, not likeability), such that gender system justifiers rated women as relatively more dominant and less hireable. In line with the Status Incongruity Hypothesis, people who were more strongly motivated to protect the gender status quo were more likely to penalize women who posed a threat to the status quo (i.e., agentic job applicants).

In line with other research (e.g., Rudman et al., 2012a, Study 2), we did not find evidence for a perceived communality deficit. Although any null results should be interpreted carefully, the fact that several studies have failed to find evidence for a perceived communality deficit while finding evidence for a dominance penalty may be interpreted as a careful indication that the dominance penalty is pivotal in backlash. In line with the SIH, gender system justifiers were more likely to perceive agentic women as overly dominant, and this fully mediated backlash. These results suggest that backlash may stem from a motivation to penalize people who engage in status incongruent behavior.

However, the data of this study should be interpreted carefully. First of all, we did not find the expected effect of system justification beliefs on

liking for agentic women. Moreover, due to the correlational nature of the present study, alternative explanations for the observed effects are possible, and more experimental research is needed to investigate if system justification beliefs predict backlash. Finally, although the paradigm employed in the present research is more ecologically valid than classic backlash-paradigms, we cannot rule out that idiosyncratic differences between the confederates apart from their gender may have influenced our results.

Replicating prior backlash research, we did not find evidence for an effect of participant sex on backlash. This lack of participant sex-differences in backlash has been somewhat of a puzzle to researchers because women generally score lower on measures of sexism (Glick, & Fiske, 2001), and are therefore expected to be more accepting of female agency. The Status Incongruity Hypothesis provides a possible explanation for this conundrum by pointing to system justifying motives as a cause of backlash. System Justification Theory suggests that low status groups (such as women) are at least as motivated to protect the status quo as high status groups (Jost, & Hunyady, 2002). Acknowledging that the status quo is unjust may evoke feelings of guilt, anxiety, discomfort and dissonance on the part of low status groups, causing them to rationalize and protect the very system that disadvantages them. Psychologically, women are as vested in the gender status quo as men are and, thus, may be just as likely to fend of threats to the status quo by penalizing gender deviants. If backlash stems from a motivation to protect the status quo, as the SIH proposes, this would explain why men and women both engage in backlash against agentic female job applicants.

By suggesting that individual differences in the motivation to protect the status quo moderate the dominance penalty and women's lower hireability ratings, the present research is the first to indicate why people are motivated to penalize agentic women. Women who portray high status behaviors such as agency endanger the gender status quo in which men have more status than women for ostensibly legitimate reasons. By showing that they are able to lead, women jeopardize the social system, and they are penalized as a way of protecting this system. In this way, backlash may be a motivated process that serves a clear purpose: to defend and protect the existing social system and, with it, people's need for certainty, clarity, and

the absence of guilt and anxiety. Casting strong, capable women off as overly dominant serves to discourage women from showing these behaviors, so the stereotype that women are weak (Glick et al., 2004) can be maintained. Although the foresight of being perceived as strident and dislikeable did not ward off Margaret Thatcher, research suggests that many women go to great lengths to avoid showing behavior that can evoke backlash (Moss-Racusin, & Rudman, 2010; Rudman, & Fairchild, 2004). Competence, for women, seems to come at a cost, and backlash may force women to choose between being respected and being liked. By sanctioning agentic women, women who apply for a leadership position are put back in their place. Moreover, the sanctions imposed on agentic women keep women in their place by providing a powerful deterrent for a new generation of potential female leaders, who may be less likely to aspire a leadership position for fear of backlash. In this way, backlash seems to fulfill its system justifying function, providing people with certainty and clarity in exchange for inequality.

Appendix 1
Excerpts from Applicant Scripts

Q1. Can you give an example of a project you did in your former position, and what your role in this project was?
At DWG, I developed a new way to conduct market analysis that improved our response by 20%. Instead of conducting market analysis by phone, I initiated a project to switch to modern technologies, such as e-mail and using Facebook. We reached a much younger group of customers and dramatically improved our marketing plan.

Q2. Can you name your two most important qualities and a point for improvement?
One of my most important qualities is that I am good at analyzing complex situations. At DWG, I was often confronted with difficult situations in which I had to incorporate perspectives of different people within the company, and determine what was important and needed to be addressed immediately and what could wait. I also know what I want and I like to make quick decisions, not debate endlessly about all the options that are on the table. It is inefficient to keep repeating arguments people already heard a couple of times; if you have all the necessary information, at a certain point you just need to stop talking and decide what to do.

A point of improvement? Let me think. Uh…I think a point of improvement is that I can be a little impatient now and then. Making decisions quickly is important, but you should also wait for others if they need a little more time to reach a certain solution.

CHAPTER 3

Picturing Men who Scream at Mice: System Threat and Mental Representations of Gender Deviant Men

This chapter is based on Nauts,S., Langner, O., Dotsch, R. & Wigboldus, D.H.J. (under review). Picturing men who scream at mice: System threat and mental representations of gender deviant men.

Women were made to be loved, and must not aim at respect, lest they should be hunted out of society as masculine.
Mary Wollstonecraft (A Vindication of the Rights of Women, 1792/2004).

Agentic women who apply for leadership positions face social sanctions (Rudman, Moss-Racusin, Phelan, & Nauts, 2012a), corroborating Wollstonecraft's contention that women "must not aim for respect". Recent research suggests that communal men may, too, be sanctioned (Moss-Racusin, Phelan, & Rudman, 2010), suggesting that men were made to be respected, but perhaps, must not aim for love.

Men are stereotypically expected to be "bad but bold", and socially approved forms of masculinity emphasize dominance and toughness (Connell, 1995; Glick et al., 2004). A failure to behave in accordance with these gender rules may lead to social sanctions (termed *backlash for gender atypicality*; Rudman, 1998; Rudman, & Glick, 2001), and boys as young as eleven may be penalized if they fail to "stand tall like a man" (Frosh, Phoenix, & Pattman, 2003; Phoenix, Frosh, & Pattman, 2003). Men may be penalized if they take time off work to care for a sick child, or if they are modest, passive, or self-disclosing (Costrich, Feinstein, Kidder, Marecek, & Pascale, 1975; Derlega, & Chaikin, 1976; Moss-Racusin et al., 2010; Rudman, & Mescher, 2013), and extant norms of masculinity are in part culprit for problems ranging from boys' underachievement in schools (Frosh et al., 2003; Phoenix et al., 2003) to men's relatively high rates of suicide and substance abuse (Cleary, 2012).

Although backlash research has long focused exclusively on women (Rudman, & Phelan, 2008), recently, the Status Incongruity Hypothesis (SIH; Rudman et al., 2012a) has provided an integrative theory of backlash against both genders. Unfortunately, the burden of evidence for the SIH has been derived from research on backlash against women (with exception of Moss-Racusin et al., 2010). Most notably, the SIH suggests that backlash serves to protect the gender status quo, but there is as of yet no empirical evidence for the role of system justifying motives in causing backlash against men. Although men and women are penalized for different behaviors (Moss-Racusin et al., 2010), the SIH forecasts that backlash against men and women stems from the same motive. In the present

research, we tested whether backlash against men is exacerbated if people are motivated to protect the status quo.

In the present research, backlash was studied using a novel, data-driven approach to study participants' mental representations of the facial appearance of atypical men. An important advantage of this approach is that it allowed us to study spontaneous impressions without asking participants to verbalize their impressions (something people may not always be able or willing to do: Nisbett, & Wilson, 1977). Study 1 introduced a novel task to capture participants' mental representations of atypical men, and in Study 2, we used a Reverse Correlation Image Classification Task (RCIC; Dotsch, Wigboldus, Langner, & Van Knippenberg, 2008[1]) to study participants' mental images. In a RCIC task, randomly varied images are presented, allowing for an unconstrained search for facial features correlated with social attributions.

The SIH and Backlash Against Men

The SIH is based on System Justification Theory (SJT), which posits that people are motivated to protect and maintain the status quo between groups in society, as doing so serves an important palliative function (e.g., to reduce guilt and anxiety; Jost, & Hunyady, 2003). The SIH builds on SJT by suggesting that backlash stems from a motivation to protect the gender status quo. According to the SIH, high status behavior (e.g., dominance) is proscribed for women, whereas low status behavior (e.g., weakness) is proscribed for men. Dominant women and weak men jeopardize the current gender status quo in which men have more status than women. People can protect this status quo by sanctioning gender deviants. In other words, women who engage in high status behavior may threaten the status quo by being be too powerful (for a woman). Men who engage in low status behavior may likewise threaten the status quo, as men's high status position in society is legitimized by their ostensibly superior strength and (leadership) skills. Thus, men who are weak, scared, or uncertain compromise the very foundation on which the gender status quo is build,

[1]Some researchers refer to a RCIC task as Reverse Correlation Task, or RCT; both terms refer to the same paradigm.

namely, the belief that men legitimately have more power than women because women are too weak to lead. Backlash, then, serves to put gender deviants back in their place as a way of defending male hegemony.

According to the SIH, backlash is exacerbated when the motivation to protect the status quo is heightened (either chronically or experimentally induced). Indeed, people with higher scores on a system justification-scale are more likely to penalize dominant women, and backlash against dominant women increases after a system threat manipulation (Rudman et al., 2012a). The present research aims to extend the SIH by studying whether a system threat manipulation also increases backlash against atypical men. We exposed participants to a system threat prime in which people read about the rise or decline of the economy, to temporarily lessen or heighten their motivation to protect the status quo (cf. Kay et al., 2009; Rudman et al., 2012a). After this system threat prime, participants were exposed to a vignette about a male or female target engaging in behavior that is proscribed for men (i.e., behavior that is considered atypical for males, as well as undesirable). As a novel way to ascertain backlash, we captured participants' mental representation of the target person's facial appearance.

Mental Representations of Gender Deviant Men

Backlash against gender deviant men can take different forms. For example, gender deviant men may be demoted (Rudman, & Mescher, 2013), disliked (Costrich et al., 1975), effeminated (Frosh et al., 2003; Phoenix et al., 2003), regarded as weak (Moss-Racusin et al., 2010) and their psychological stability may be questioned (Derlega, & Chaikin, 1976). The present research took a radically different approach to studying backlash by investigating people's spontaneous representations of the facial appearance of gender deviant men. More specifically, we explored whether people formed mental representations of gender deviant men as having more weak, negative or feminine facial features. To do so, we used two different tasks: the Draw-a-Face Task in Study 1, and the Reverse Correlation Image Classification Task (Dotsch et al., 2008) in Study 2.

Faces are omnipresent in our daily lives, and people spontaneously draw complex social inferences from faces with remarkable ease and rapidity. Inferences about the trustworthiness of faces, for example, seem

to be formed within 50 ms (Todorov, Pakrashi, & Oosterhof, 2009). Further, people strongly agree on what constitutes a trustworthy face (Todorov, Said, Engell, & Oosterhof, 2008), and believe in the accuracy of their inferences even if they are obviously wrong (Olivola, & Todorov, 2010). Inferences of facial appearance predict important societal outcomes (for an overview, see Todorov, Olivola, Dotsch, & Mende-Siedlecki, in prep.) such as who wins elections (Ballew, & Todorov, 2007; Hall, Goren, Chaiken, & Todorov, 2009; Todorov, Mandisodza, Goren, & Hall, 2005), and who is promoted in the military (Mazur, Mazur, & Keating, 1984).

People's impressions of faces can have important societal outcomes, even if these inferences are not necessarily accurate. The opposite effect can also occur: people's representations of faces can be influenced by the ideas they have about a group, even if those ideas are not necessarily accurate. There is a growing body of evidence suggesting that stereotypes can influence what people imagine a face looks like. For example, people's mental representations of the faces of homosexual men contain more feminine features than their representations of the faces of heterosexual men (Dotsch, Wigboldus, & Van Knippenberg, 2011). Moreover, Indian children have mental representations of Brahmin faces (a high status group in India) that contain more positive features than their mental representations of Dalit faces (a low status group in India; Dunham, Srinivasan, Dotsch, & Barner, 2014). Lastly, prejudiced Dutch participants hold mental representations of Moroccans (a stigmatized group in The Netherlands) that contain more criminal facial features (Dotsch et al., 2008; Dotsch et al., 2011). Thus, far from being an objective reality, people's mental representations of faces are influenced by their stereotypic beliefs.

Mental representations of facial appearance can be studied using data driven tasks such as the Reverse Correlation Image Classification Task (RCIC; Dotsch et al., 2008; Dotsch et al., 2011; Mangini, & Biederman, 2004; Todorov, Dotsch, Wigboldus, & Said, 2011). An important advantage of data driven tasks is that they do not force participants to verbalize their impression of a target person, as people may not necessarily be able or willing to do so (Dotsch, & Todorov, 2012). Data driven methods such as a RCIC are indirect measures that allow participants to form an unconstrained impression that is not guided by the researcher's questions.

This feature of data driven tasks allows researchers to gain insight in the impressions that perceivers spontaneously form of gender deviant men.

In the present research, we employed two data driven tasks to capture participants' mental representations of gender deviant men. For Study 1, we developed a new task, the Draw-a-Face Task (DaFT) to study whether, after a system threat prime, participants imagined gender deviant men to be more weak, feminine and dislikeable. In Study 2, we used a Reverse Correlation Image Classification Task to study inferences of weakness, femininity, and valence. In both studies, participants were exposed to an elaborate behavioral vignette that contained neutral information as well as two stereotype violations. After reading this vignette, we captured participants' mental image of the target person using the DaFT or RCIC task. Next, a second, independent, group of participants rated the pictures generated by the first group on toughness, femininity and likeability (Study 1) or weakness, femininity and positivity (Study 2). We expected that, after a system threat prime, participants would form mental representations of gender deviant men as more weak, feminine, dislikeable and negative.

Study 1: Draw-a-Face Task

In Study 1, we developed a new task to study the effect of a system threat manipulation on people's mental representation of the facial appearance of gender deviant men. In this task, the Draw-a-Face Task (DaFT), participants could compose a face by selecting elements (e.g., hair, eyes, etc.) from a list of alternatives, yielding images that are similar to those that can be created using phantom image generation software used by the police. This task is short and easy-to-use for participants, has high face validity, and delivers ecologically valid images of faces.

Study 1 consisted of two phases. In the image generation phase (Phase 1), participants were exposed to a system threat-manipulation, read a vignette about the target person, and then created an image of the target person using the DaFT. In the image rating phase of the experiment (Phase 2), a second, independent, group of participants rated the images created in the image generation phase on toughness, masculinity, and likeability. We expected that, after a system threat, people would mentally represent gender deviant men as relatively weaker, more feminine, and less likeable.

Method

Participants. Seventy-six Dutch Radboud University students participated in the image generation phase of the study in exchange for a €5 gift certificate. For seven participants, the data of the DaFT were erroneously not saved, resulting in a final sample of 69 participants (16 men). Participants in this phase of the study were between 18 and 65 years old (average age 21).

One hundred and eighty-two participants (104 men) participated in the image rating- phase of the study through Amazon's Mechanical Turk in exchange for $0.50. Participants in this phase of the study were between 18 and 65 years old (average age 36).

Materials and procedure.

System threat manipulation. The system threat manipulation consisted of an alleged newspaper article about the rise (system boost-condition) or decline (system threat-condition) of the Dutch economy (cf. Kay et al., 2009; Rudman et al., 2012a). Participants were told that the study consisted of two unrelated parts and that, in this first part, they would read a newspaper article to pretest materials for an upcoming study. Different fonts were used throughout the experiment to bolster the impression that the system threat-manipulation was not part of the same study as the target vignette and DaFT. In the system boost-condition, participants read an article about the rise of the Dutch economy; in the system threat-condition, they read an article about the decline of the Dutch economy (see Appendix 1). Following that, participants wrote down for three minutes why they thought the author's position was justified. Finally, to bolster our cover story, they indicated how well-written, compelling, interesting, and understandable they thought the article was[2].

Target information. The system threat manipulation was followed by a vignette about a target person (whom we named "M."). Participants were instructed to read the information about M. and try to form an impression of him or her, as they would have to answer questions about

[2]These ratings did not differ across conditions for Study 1 (all Fs < 1) or Study 2 (Fs 0.69 to 1.36, p's .23 to .68).

this person later in the experiment. The vignettes (see Appendix 2) consisted of two behaviors that entailed a proscriptive stereotype violation for men, as well as some general information about M., a 20-year old male or female student who was described as a typical Dutch freshman. We added some generic information about M. to give participants sufficient information to form an impression.

As stereotypic behaviors, we selected behaviors that were proscribed for males (i.e., that were considered relatively unexpected and undesirable for men). To do so, we conducted a pretest in which four participant samples (*N*s 20 to 27) indicated how typical several behaviors are for men, how typical they are for women, how desirable they are for men, or how desirable they are for women. We calculated the effect size (Cohen's *d*) for differences in typicality and desirability for men and women (cf. Rudman et al., 2012a) and finally selected the following behaviors (translated from Dutch): "started to scream when he/she saw a mouse" and "had a stomach ache when he/she had to give a presentation". These behaviors were regarded as strongly atypical and strongly undesirable for men compared to women (average Cohen's *d*s = 2.81 and 1.29, respectively)[3].

Draw-a-Face Task. The DaFT was developed specifically for the present research, as a way to capture participants' mental representations of faces. In this task, participants could compose a face by selecting different components, such as hair, a nose, eyebrows, eyes, glasses, a jaw, and a mouth (see Figure 1 for a screenshot of the task; see Figure 2 for examples of images participants created using the DaFT). The task consisted of several tabs with lists of elements that participants could select. In total, the DaFT consisted of 247 elements (86 hairstyles, 15 noses, 34 pairs of eyebrows, 35 pairs of eyes, 7 glasses, 21 jaws and 49 mouths), resulting in over a billion possible feature combinations. All of these elements could be individually resized (their width and height could be separately altered to make them wider/ narrower), moved and rotated, so that participants could create a face that was in line with their mental representation of M. We created the individual components by selecting elements from faces (taken

[3]By convention, small, medium and large effect sizes correspond to a Cohen's *d* of 0.20, 0.50 and 0.80, respectively (Cohen, 1988)

from the Radboud Faces Database; Langner et al., 2010 and Karolinska Emotional Faces Database; Lundqvist, & Litton, 1998) and adjusting the hue of these pictures. Although the DaFT has an intuitive user interface and participants report finding it easy to use, we showed participants a short instruction film before they commenced the task to ensure that they understood how it worked.

Figure 1. Screenshot of the Draw-a-Face Task.

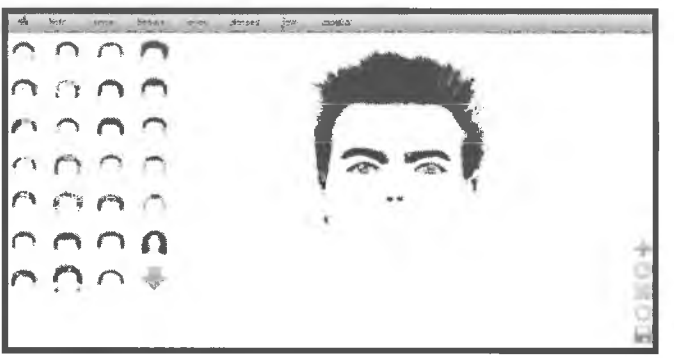
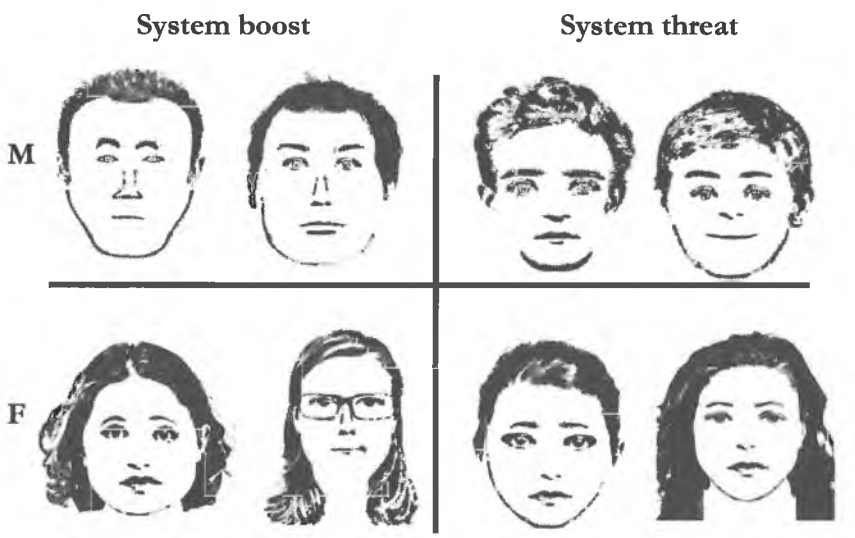


Figure 2. Examples of faces that were created in each condition by participants using the Draw-a-Face Task.



In the image rating phase of the study, a second, independent, group of participants rated all pictures that had been created in Phase 1 of the study. The image rating phase of the study had a between subjects-design, so that participants rated all pictures either on toughness (N = 67) or on masculinity (N = 54) or on likeability (N = 59). All questions were answered on 7-point Likert scales (anchors: very weak-very tough; very feminine-very masculine; very negative-very positive), and the order of pictures was randomized.

Results

Toughness. Before conducting the analyses, we averaged participants' ratings for each image (cf. Dotsch et al., 2008). We expected that, after a system threat manipulation, gender deviant men would be rated as relatively weaker (compared to women). We conducted a 2 (system threat prime: boost versus threat) x 2 (target gender: male versus female) between subjects ANOVA to test this. In line with our expectations, there was a significant interaction between system threat and target gender, $F(1,68) = 5.26$, $p = .03$, $\eta_p^2 = .08$.

Table 1. *Average toughness-ratings for male and female target images created in the system boost- and system threat-condition.*

	Toughness		
	Male target	**Female target**	**Cohen's *d***
System Boost	4.17	3.44	**1.23**
SD	*0.70*	*0.46*	
System Threat	3.79	3.66	**0.25**
SD	*0.58*	*0.43*	

Note. Larger values for Cohen's *d* reflect greater toughness ratings for male compared to female images.

The pattern of results was in line with our expectations: As depicted in Table 1, in the system boost-condition, gender deviant men were rated as tougher than women, $F(1, 68) = 15.08$, $p < .001$, Cohen's $d = 1.23$. However, in the system threat-condition, the male-female difference was not significant, $F < 1$.

After being threatened by the decline of the Dutch economy, participants formed mental representations of gender deviant men as relatively weaker.[4]

Masculinity and likeability. We conducted analyses similar to the above for the average image ratings of masculinity and likeability. Against our prediction, we neither found evidence for system threat leading to more feminine mental representations of gender deviant men, nor for system threat affecting the likeability of gender deviant men, *F*s < 1 for both interactions.

Discussion

In line with our expectations, participants who had been threatened with the decline of the Dutch economy formed mental representations of gender deviant men as having relatively more weak facial features. In line with the Status Incongruity Hypothesis, these results suggest that a system threat manipulation exacerbates backlash against gender deviant men. In so doing, the present study extends the SIH by providing empirical evidence suggesting that backlash against men also stems from a motivation to protect the status quo. Contrary to our expectations, we did not find effects on femininity and likeability. After a system threat-manipulation, gender deviant men were rated as weaker (in comparison to women), but not as significantly more feminine or less likeable. It is unclear why men were not rated as less likeable or more feminine after a system threat. Possibly, the DaFT was insufficiently sensitive to pick up differences in likeability and femininity. In Study 2, we resorted to a different task that may be more suitable to pick up these differences.

Study 2: Reverse Correlation Image Classification Task

In Study 2, we employed a different task, a RCIC, to explore participants' mental images of gender deviant men after a system threat prime. In Study 2, instead of measuring likeability (as we did in Study 1), we measured positivity, as this judgment is more commonly used in face perception research because it is an important determinant of approach-responses to

[4] We did not expect any significant differences between male and female participants and, indeed, there was no significant system threat x target gender x participant sex interaction, $F(1,68) = 2.59$, $p = .11$.

faces (Todorov, Baron, & Oosterhof, 2008). We expected that, after a system threat manipulation, participants would form mental representations of gender deviant men (compared to women) that were more negative, weak, and feminine.

The role of prescriptive stereotypes

In addition to using a different task, in Study 2, we also added new behavioral stimuli to study which components of gender stereotypes are culpable in backlash. The stimuli that were used in Study 1 were all strong proscriptive stereotype violations for males: in Study 2, we added descriptive stereotype violations. Gender stereotypes typically consist of a descriptive component describing how men and women are *expected* to behave, as well as a prescriptive component stipulating how men and women *should* behave (negative prescriptions that proscribe how men and women should *not* behave are termed proscriptions; Burgess, & Borgida, 1999; Prentice, & Carranza, 2002, 2004; Rudman et al., 2012a). In Study 2, we explored whether each of these two components contributed to backlash.

Expectancy violation and norm violation may both contribute to backlash, albeit in different ways. The descriptive component of gender stereotypes may invoke a contrast effect (Eagly, Makhijani, & Klonsky, 1992); according to attribution theory, people form more extreme inferences of unexpected behaviors (Kelley, & Michela, 1980). From this perspective, backlash effects could alternatively be explained by simple contrast effects and without the need to invoke a motivational account as proposed by the SIH. If backlash is driven by contrast effects, weak men may be regarded as particularly weak (and negative) simply because their behavior is atypical.

Alternatively, the SIH predicts that the prescriptive component of gender stereotypes is the culprit in backlash. From this perspective, gender deviants are strongly penalized because gender stereotypes entail norms about how men and women should behave. This is in line with other theorizing suggesting that gender stereotypes are particularly impervious to change due to their strongly proscriptive nature, which causes men and women to refrain from behaving atypically to avoid penalization (Prentice, & Carranza, 2004). Thus, according to the SIH, backlash should be

stronger for behavior that is strongly norm-violating than for behavior that is not strongly norm-violating.

In Study 2, we added behaviors that were strongly unexpected for males, but did not invoke a clear norm violation. We expected that participants would have the most negative mental images of gender deviant males if they engaged in norm-violating behavior and participants felt threatened by the decline of the Dutch economy. Put differently, we expected that, after a system threat prime, backlash effects would be particularly strong if men's behavior constituted a proscriptive stereotype violation.

The design of Study 2 was highly comparable to the design of Study 1, with two notable exceptions: we added descriptive stereotype violations and used a Reverse Correlation Image Classification Task (RCIC) to capture participant's mental representations of gender deviant men. Like Study 1, Study 2 consisted of two phases: an image generation phase and an image rating phase.

Method

Participants. One hundred and forty-six Dutch Radboud University students (63 men) participated in the image generation phase of the study in exchange for partial course credit or a €7,50 gift certificate. One participant was removed from this sample because he failed to follow the experimenter's instructions, resulting in a final sample of 145 participants (62 men)[5]. Participants in Phase 1 were between 17 and 36 years old (average age 22).

Forty-seven Dutch Radboud University students (13 men) participated in the image rating phase of the study in exchange for partial course credit or a €5 gift certificate. Participants in Phase 2 were between 18 and 36 years old (average age 22).

Materials and procedure. We selected proscriptive and descriptive stereotype violations based on the same pretest as in Study 1. As proscriptive stereotype violations, we used the same stimuli as in Study 1.

[5]The pattern of significant and non-significant findings remains the same regardless of whether this particular participant is removed from the sample or not.

As descriptive stereotype violations, we selected the following behaviors: "accidentally caused the power to short-circuit" and "completely rewrote an essay, because he/she thought it was not perfect yet". These behaviors were regarded as strongly atypical for men compared to women (average Cohen's d = 1.26), but the difference in desirability between men and women was much smaller (average Cohen's d = 0.40) than for the proscriptive stereotype violations (average Cohen's d = 1.29). Thus, the proscriptive and descriptive stereotypes were regarded as strongly unexpected for men, but the proscriptions entailed a strong norm violation, whereas the descriptive stereotypes did not. Although there was a large difference in how norm-violating the used proscriptions and descriptions were, even the descriptive stereotypes were slightly norm-violating: the stimuli we selected were descriptive in a relative sense.

As in Study 1, participants were first exposed to a system threat manipulation, in which they read an alleged newspaper article about the rise (system boost-condition) or decline (system threat-condition) of the Dutch economy. After this task, participants read a vignette about M. The vignette was exactly the same as in Study 1, except for in the descriptive stereotype violation-conditions. In these conditions, the proscriptive stereotype violations were replaced by descriptive stereotype violations. After reading the vignette about M., participants completed a short filler task in which they had to solve 15 anagrams of European cities.

Reverse Correlation Image Classification Task. After the filler task, participants completed a forced choice version of the Reverse Correlation Image Classification Task (RCIC; Dotsch et al., 2008; Dotsch, & Todorov, 2012; Mangini, & Biederman, 2004). In this task, participants were presented with two noisy images of faces, side by side. Participants' task was to select the image that most looked like M. Participants completed 450 trials of a RCIC task (five blocks of 90 trials each, with short breaks in between). These trials consisted of a base face (the average face of all male and female faces in the Karolinska Emotional Faces Database; Lundqvist, & Litton, 1998; see Figure 3) on which random noise patterns were superimposed. In a given trial, the same noise pattern was added to the base face in picture A and subtracted from the base face in picture B

(see Figure 4). Different random noise pattern were used across trials[6]. Superimposing noise patterns on the base face distorts the facial features, making each picture look subtly different.

For each participant, we created a classification image (CI; see examples in Figure 5) by averaging all noise patterns of those faces they had selected as resembling M. Superimposing the CI of a participant on the original base face is supposed to depict that participant's mental representation of M's face. Responses on trials for which the response latency was smaller than 300 ms (2.63% of trials) were excluded.

The 145 images of mental representations that resulted from the RCIC task were judged by a second, independent, group of participants in Phase 2. These images were rated on positivity, masculinity and weakness on a 7-point scale (anchors: very negative-very positive; very feminine-very masculine; not weak at all-very weak). Participants first completed a block in which they rated all pictures on one feature (e.g., on masculinity), then proceeded to the next block in which they rated all pictures on another feature (e.g., positivity), etc. Block order was counterbalanced across participants, and the order of pictures within the blocks was randomized.

Figure 3. Base face used in all trials of a RCIC.

[6]These random noise patterns consisted of truncated sinusoid patches that were a function of over 4000 parameters, namely: orientation (0°, 30°, 60°, 90°,120° and 150°), spatial frequency (1,2,4,8 and 16 patches per image, with each patch spanning two sinusoid cycles) and phase (Π/2), with random contrasts (amplitudes).

Figure 4. Example of a trial used in a RCIC. Face A and B consist of the same base face to which random noise is either added (A) or subtracted (B).

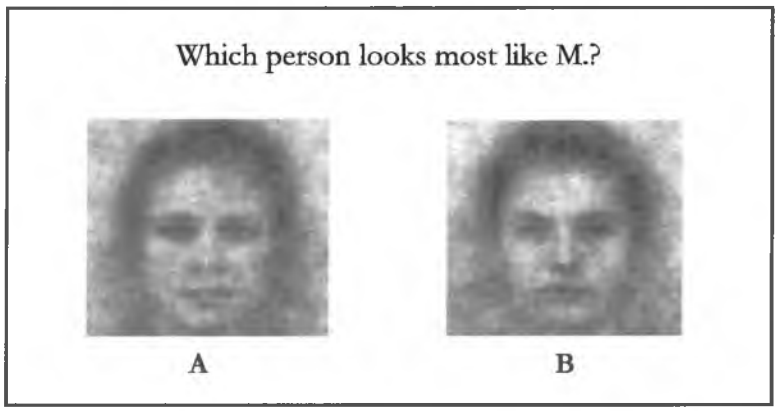


Figure 5. Examples of classification images that were created by participants using the Reverse Correlation Image Classification Task (RCIC).

Results

Positivity. Before conducting the analyses, we averaged participants' ratings for each image (cf. Dotsch et al., 2008). We conducted a 2 (system threat prime: boost versus threat) x 2 (target gender: male versus female) x 2 (kind of stereotype: description versus proscription) ANOVA to investigate the effects of system threat on the relative positivity of men who engaged in descriptive and proscriptive stereotype violations. We expected that backlash effects would be strongest after a system threat, and for proscriptive stereotype violations. Put differently, we expected that participants' mental representations of men would be more negative than participants' mental representations of women, especially if men engaged in

a proscriptive stereotype violation and participants had been threatened by the decline of the Dutch economy. Contrary to our expectations, there was no significant three-way interaction between system threat prime (boost versus threat), target gender (male versus female) and stereotype (descriptive versus proscriptive), $F(1,144) = 1.37$, $p = .24$. The data of Study 2 did not suggest that backlash effects were significantly stronger for proscriptive stereotype violations in the system threat condition compared to the other conditions.

Although the omnibus interaction did not support our hypothesis, the simple main effects were strongly in line with our expectations. Looking at the simple main effects, we found evidence for backlash effects only for the condition with the system threat-manipulation and the proscriptive stereotype violation. As listed in Table 2, participants threatened with the decline of the Dutch economy formed more negative mental representations of men (compared to women) who engaged in a proscriptive norm violation, $F(1, 53) = 3.93$. $p = .05$, Cohen's $d = 0.73$. In none of the other conditions the ratings of men and women differed significantly, all Fs < 1. In line with our expectations, we only found evidence for backlash after a system threat, and when behavior was strongly norm-violating. We did not find evidence for backlash when behavior was not strongly norm-violating, or when behavior was norm-violating but participants had been primed with a system boost. Put differently, backlash only emerged when participants' motivation to protect the status quo was temporarily heightened, and when men's behavior contained a strong proscriptive component. Men were not penalized if their behavior was merely unexpected (i.e., when they rewrote a paper and caused the power to short-circuit), but only if it strongly violated gender rules (i.e., when they screamed at a mouse and had a stomach ache before giving a presentation) and participants were motivated to defend the gender hierarchy.[7]

As Table 2 illustrates, the data leave several questions unanswered. For example, the pattern of means suggests that, in line with our expectations, mental images of male proscriptive stereotype violators

[7]We did not expect any differences between male and female participants and, indeed, there was no significant system threat x target gender x participant sex interaction, $F < 1$.

become more negative, $F(1,46) = 14.84$, $p < .001$. Unexpectedly, however, mental images of women who engage in these behaviors become more positive after a system threat prime, $F(1,46) = 44.64$, $p < .001$.We did not anticipate this finding, but it is possible that weak, stereotypical women are seen as more positive when the status quo is threatened. More research is needed to shed light on these effects. Moreover, proscriptions and descriptions differed in terms of how unexpected they were: in future research, stimuli should be selected that are equal in terms of expectancy, but not desirability.

Table 2. *Mean positivity ratings of Classification Images for the system boost- and system threat-conditions for proscriptive and descriptive stereotype violations, and difference in rating between men and women (Cohen's d).*

	Descr. Female Target	Descr. Male Target	*d*	Proscr. Female Target	Proscr. Male Target	*d*
System Boost	3.73	3.54	**0.30**	3.75	3.81	**-0.09**
SD	*0.68*	*0.62*		*0.65*	*0.64*	
System Threat	4.18	3.99	**0.25**	4.03	3.58	**0.73**
SD	*0.80*	*0.69*		*0.57*	*0.67*	

Note. Larger values for Cohen's *d* reflect greater positivity ratings for female compared to male images.

Masculinity and weakness. We conducted analyses similar to the above for the average image ratings of masculinity and weakness. We expected that male participants would be rated as particularly feminine if they engaged in a proscriptive norm violation and participants' worldviews had been threatened. We did not find evidence for this effect: there was no significant three-way interaction between system threat, target gender, and kind of stereotype, $F < 1$. We also expected that male participants would be rated as particularly weak if they engaged in a proscriptive stereotype violation and participants' worldviews had been threatened. We did not find evidence for this effect, either: there was no significant three-way interaction between system threat, target gender, and kind of stereotype, $F < 1$.

Discussion

In line with our expectations, we found evidence of backlash if men's behavior was norm-violating and participants had been primed with a system threat: in all other conditions, we did not find evidence of backlash. These findings suggest that people formed more negative mental representations of males compared to females if their behavior was norm-violating, and if participants felt threatened, corroborating the SIH's contention that the prescriptive component of gender stereotypes plays a pivotal role in explaining backlash. If behavior was merely unexpected, but not strongly proscribed for males, people did not engage in backlash against atypical men. However, these data should be interpreted carefully because the omnibus interaction was not significant and we unexpectedly did not find any effects on masculinity and weakness. Moreover, more research is needed to study the possibility that threats to the status quo do not only affect people's evaluation of atypical men, but also their evaluation of stereotypical women. Because women who behave in accordance with gender stereotypes may reaffirm the status quo, such an effect would be in line with the Status Incongruity Hypothesis.

General Discussion

The goal of the present line of research was to investigate whether system justifying motives underlie backlash against atypical men. According to the SIH, agentic women are penalized because they threaten the gender hierarchy, and men may be penalized for similar reasons (Rudman et al., 2012a). Empirical evidence for this latter link was lacking, and the present research set out to investigate whether backlash against men is exacerbated after a system threat prime. In two studies, we found evidence suggesting that a system threat prime increases backlash against gender deviant men, corroborating the SIH's contention that backlash serves to protect the status quo. This suggests that scared or nervous men compromise the legitimacy of the gender hierarchy, and that backlash serves to put them back in their place.

Although it may seem intuitive that agentic, powerful women pose a threat to the gender hierarchy, the present research suggests that weak men may likewise threaten male hegemony. Men who engage in weak, low status behavior (e.g., being very nervous before giving a presentation)

compromise the perceived legitimacy of men's superior societal status by challenging the notion that men are more fit to lead. When people's belief in the social system is threatened (e.g., because they read about the decline of the economy), this increases their motivation to bolster the gender hierarchy, causing them to lash out at atypical men. Interestingly, the present research suggests that, after a system threat prime, participants *spontaneously* formed more weak (Study 1) or negative (Study 2) impressions of gender deviant men. By using data driven methods, we were able to study backlash in an indirect and unconstrained way, without forcing participants to form an impression of any particular characteristic of the target person.

Next to suggesting that system threat exacerbates backlash against gender deviant men, the present research extends backlash research by suggesting that men are penalized outside of an employment context, and for behaviors other than modesty. In the present research, men were penalized for being scared (screaming at a mouse) and nervous (having a stomach ache before giving a presentation). Moreover, the present research suggests that backlash may not merely be the result of a contrast effect. In Study 2, we used behaviors that were unexpected for men, but that differed in their extent of norm violation. Backlash effects only emerged after a system threat, and if behavior entailed a proscriptive stereotype violation, corroborating the SIH's notion that the prescriptive components of gender stereotypes play a pivotal role in backlash. However, the findings of Studies 1 and 2 are not entirely consistent: in Study 1, we found effects on ratings of weakness, but not likeability, whereas in Study 2, we found effects on positivity, but not weakness. We did not find effects on masculinity in either study. Although speculative, this divergence could be due to methodological differences between the DaFT and RCIC task. For example, it may be easier to vary facial weakness in the DaFT than in a RCIC task. Facial weakness is strongly influenced by the relative width and height of a face (the Width-to-Height ratio or WtH; Carré, & McCormick, 2008). In the DaFT, participants can easily change the width and height of a face through a simple point-and-click-interface. In a RCIC task, the exact same changes will appear infrequently and only by chance, as they would need to be aligned with random variations in many image locations. Especially with the relatively small number of Reverse Correlation-trials in

Study 2 (450), a RCIC task is unlikely to find consonant WtH-ratio variations. On the other hand, a RCIC task may be more sensitive than the DaFT to the low spatial frequency information in faces which is indicative of valence and trustworthiness (Bar, Neta, & Linz, 2006), potentially making it a more sensitive measure of valence.

Alternatively, it is possible that the divergences between Study 1 and 2 reflect ambivalence in people's attitudes towards gender deviant men. Stereotypes of men contain an element of threat (they are seen as "bad, but bold"; Glick et al., 2004), while stereotypes of women are univalently positive (Eagly, & Mladinic, 1994). Gender deviant men threaten the status quo, but they may also be regarded as weaker and, therefore, as less threatening. In line with the idea that gender deviance is univalently negative for women, but not for men, gender deviant men are liked less in some studies (e.g., Moss-Racusin et al., 2012), but not others (e.g., Derlega, & Chaikin, 1976). Gender deviant women, on the other hand, are generally disliked (for an overview, see Rudman, & Phelan, 2008; Rudman, Moss-Racusin, Glick, & Phelan, 2012). Future research could further investigate the content of people's impressions of gender deviant men, for example by using a more qualitative approach to studying mental representations. By asking participants to write down their impressions of faces generated with the DaFT or RCIC task and using content analysis to analyze these impressions, researchers could gain more insight in the potentially ambivalent nature of impressions of gender deviant men.

Although the results of the present research leave several questions unanswered, we believe that they present an exciting new toolkit for studying gender stereotypes. Next to exploring a theoretical question regarding backlash against gender deviant men, the present research introduced the Draw-a-Face-Task (DaFT) as a new paradigm to study mental representations. We believe that the DaFT may be valuable to researchers in the domain of stereotyping and person perception because it has several practical advantages over a RCIC task. For example, the DaFT yields ecologically valid images, does not require programming skills and can be administered in a couple of minutes (in contrast, a RCIC task is long and tedious for participants, as researchers use hundreds or even thousands of trials; e.g, Dotsch et al., 2011; Jack, Garrod, Yu, Caldara, & Schyns 2012). Although the DaFT has some disadvantages over a RCIC task (e.g.,

participants are limited by the facial elements provided to them), we believe that the DaFT's practical advantages make it a more accessible task than a RCIC, allowing more researchers to reap the benefits of studying mental representations.

In sum, the present research suggests that backlash against men is exacerbated after a system threat manipulation. When people had been threatened by the decline of the economy, they imagined norm-violating men as having relatively weak (Study 1) or negative (Study 2) facial features. Men and women may be penalized for different behaviors, but our data suggest that they are penalized for the same reason. In the present research, men were not allowed to be nervous or scared, suggesting that gender stereotypes straitjacket men by confining them through norms of toughness. Perhaps ironically, people lashed out at scared men most harshly when they, themselves, felt threatened.

Appendix 1

Alleged newspaper article used as system boost manipulation in Studies 1 and 2. This article (translated from Dutch) was entitled "International position of the Netherlands on the rise":

> *More and more people in the Netherlands are satisfied with the state of their country, and the Dutch feel more confident about their country than ever before. Whether it is thanks to the relatively quick economic recovery after the crisis, increasing influence within the European Union, or the belief that the government will be able to leverage global changes to their advantage and keep government finances healthy, the Dutch are, as a whole, very satisfied. Many citizens feel that the country is socially, economically, and politically stable. It seems that the economic and political circumstances in The Netherlands are better than those in surrounding countries, especially now that the Dutch economy is recovering from the crisis more quickly than economies in surrounding countries. A recent poll by Statistics Netherlands indicates that fewer people than ever before indicate a willingness to leave The Netherlands and emigrate to another country."*

Alleged newspaper article used as system threat manipulation in Studies 1 and 2. This article (translated from Dutch) was entitled "International position of The Netherlands in decline":

> *More and more people in the Netherlands are disappointed with the state of their country, and the Dutch feel more uncertain about their country than ever before. Whether it is due to the crisis and deteriorating economic circumstances, decreasing influence within the European Union, or a general fear and anxiety that the government will be unable to leverage global changes to their advantage and keep government finances healthy, the Dutch are, as a whole, deeply dissatisfied. Many citizens feel that the country has reached a low point socially, economically, and politically. It seems that the economic and political circumstances in surrounding countries are better than those in The Netherlands, especially now that the Dutch economy is recovering from the crisis less quickly than economies in surrounding countries. A recent poll by Statistics Netherlands indicates that more people than ever before indicate a willingness to leave The Netherlands and emigrate to another country."*

Appendix 2

Vignette used in Studies 1 and 2 (translated from Dutch). Parts that constitute a proscriptive stereotype violation are put between square brackets. In Study 2, in the descriptive stereotype violating conditions, the sentences between brackets were replaced by the following sentences: "accidentally caused the power to short-circuit" and "completely rewrote an essay, because he/she thought it was not perfect yet".

"M. is a 20-year old man/woman from Nijmegen. Last year, he/she moved into a dormitory and started studying at Radboud University. Originally, he/she is from a village in Gelderland, but his/her village is too far from Nijmegen to commute, and M. thought it would be fun to move into a dormitory. M. enjoys life as a student. He/She at first had to get used to it, but in general, he/she is happy that he/she moved to Nijmegen. He/She enjoys the study program, although studying is quite time-consuming because M. has to give lots of presentations and has to write lots of essays. That does not always go well:[the other night, he/she had a bad stomach ache because he/she had to give a presentation.] But M. still has enough time left to work as a waiter, and off course to go out. He/She is not the kind of person who goes partying until 7 am twice a week, but M. does enjoy going to dorm parties and other fun parties. M. has nice roommates, and they often cook together. His/her room is not very big, but he/she has everything that he/she needs there. Sometimes, something goes wrong: the other day, he/she [had to scream when he/she saw a mouse]. Still, M. really likes living in a dorm: he/she finds it much more enjoyable than living with his/her parents. Next year, M. will start his/her sophomore year, after which he/she will have to choose a major. M. is not sure what major he/she would like to pick, so he/she plans to figure that out in the coming year. "

CHAPTER 4

Forgive and Forget? System Justification and Memory for Stereotype-Inconsistent Behavior

This chapter is based on Nauts, S., Rudman, L.A., Langner, O., & Wigboldus, D.H.J. (in prep). Forgive and forget? System justification and memory for stereotype-inconsistent behaviors.

Acknowledgements

We would like to thank Bart Meuleman, Inge Huijsmans and Veronique Louhenapessy for their help in coding participants' responses.

Despite the success of the women's movement, women are strongly underrepresented in high status positions that require agentic qualities (e.g., only 3% of the CEOs of Fortune 500-companies is female; Catalyst, 2012), whereas men are underrepresented in low status positions that require communal qualities (e.g., only 7% of registered nurses in the US is male; US Department of Health and Human Services, 2010). Backlash research offers a potential explanation for this effect by pointing out that agentic women and communal men face social and economic penalties (termed "backlash for gender atypicality"; Moss-Racusin, Rudman, & Phelan, 2010; Rudman, 1998; Rudman, & Glick, 2001). Backlash limits women's chances of obtaining positions of power (for an overview, see Rudman, Moss-Racusin, Glick, & Phelan, 2012), and straitjackets members of both genders by limiting the behavioral and occupational options that are available to them.

Status Incongruity and Backlash

The Status Incongruity Hypothesis (SIH; Rudman, Phelan, Moss-Racusin, & Nauts, 2012a) suggests that people engage in backlash to protect the gender hierarchy, in which men are awarded more societal status than women (Ridgeway, 2001). Whenever a woman enacts high status behavior or a man enacts low status behavior, this behavior is incongruent with the status of their gender (i.e., it is *status incongruent*). Status incongruent behavior jeopardizes the gender hierarchy, and people engage in backlash as a way of restoring this hierarchy. Thus, women are proscribed from high status behavior such as being dominant, stubborn or demanding, whereas men are proscribed from low status behaviors such as being weak, uncertain, or emotional. In line with the SIH, backlash is exacerbated when people's motivation to protect the status quo is heightened (Chapters 2 and 3; Rudman et al., 2012a), suggesting that people engage in backlash to reduce threats to the status quo.

System justifying motives exacerbate backlash in explicit target ratings (Chapter 2) and affect people's mental representations of the facial appearance of atypical men (Chapter 3). In the present chapter, we will study whether they also affect memory. More specifically, we will investigate if people's memory for stereotype violations becomes better if they are motivated to protect the status quo. Gender stereotypes typically consist of two components: a descriptive and a prescriptive component.

The descriptive component stipulates how men and women are *expected* to behave; the prescriptive component how men and women *ought to* behave. Negative prescriptions containing rules about how men and women ought *not* to behave are termed proscriptions (Burgess, & Borgida, 1999; Prentice, & Carranza, 2002; 2004). Thus, next to specifying how men and women typically behave, proscriptive stereotypes also lay out gender rules that specify the kinds of traits and behaviors men and women are not allowed to portray. Proscriptive stereotype violations threaten the gender status quo and are therefore strongly policed (Chapter 3; Rudman, & Glick, 2008). Because proscriptive stereotypes play such an important role in causing backlash, in the present chapter, we will study the effect of system justifying motives on memory for proscriptive stereotype violations.

If people's motivation to protect the status quo is heightened (either chronically or experimentally induced), we expect that they will have better memory for proscriptive stereotype violations. Put differently, system justifying motives are expected to improve memory for proscriptive stereotype violations, as these behaviors may jeopardize the gender hierarchy. According to System Justification Theory (Jost, Banaji, & Nosek, 2004), people are strongly motivated to maintain societal hierarchies and defend the legitimacy of social arrangements. Even people who may be better off if existing hierarchies would be overturned (e.g., women and disadvantaged minorities; Jost et al., 2004) may be unconsciously motivated to protect the very system that disadvantages them. System justification serves a palliative function, as defending the social system helps people to avert the anxiety, uncertainty, and guilt that they would experience if they were to acknowledge that the system is unjust (Jost, & Hunyady, 2002). Because system justification is such a powerful force, we hypothesize that when system justifying motives are heightened, people will have better memory for behaviors that can potentially disrupt the gender status quo. In other words, system justifiers may be more likely to remember proscriptive stereotype violations (compared to stereotypical behaviors) because these behaviors likely constitute a threat to the status quo.

Memory for Stereotype-Inconsistent Information

There is a vast body of literature on stereotype-consistency-effects in memory. Some studies suggest that memory for stereotype-consistent

information is better than memory for stereotype-inconsistent information (e.g., Fyock, & Stangor, 1994; Martin, & Halverson, 1983; Rothbarts, Evans, & Fulero, 1979; Stangor, & Ruble, 1989a; Stangor, & Ruble, 1989b), whereas other studies suggest that, at least in certain contexts, memory for stereotype-inconsistent information is better than memory for stereotype-consistent information (e.g., Dijksterhuis, & Van Knippenberg, 1995; Dijksterhuis, Van Knippenberg, Kruglanksi, & Schaper, 1996; Macrae, Hewstone, & Griffiths, 1993; Sherman, & Frost, 2000; Stangor, & McMillan, 1992). Proponents of the first view suggest that stereotype-consistent information has an encoding-advantage because it fits existing cognitive schemas (this is termed the *'schematic filtering*-hypothesis: Sherman, & Frost, 2000). In line with the schematic filtering-hypothesis, a large meta-analysis suggests that memory for stereotype-consistent behavior surpasses memory for stereotype-inconsistent behavior (Fyoch, & Stangor, 1994). In the domain of gender stereotypes, research in line with the schematic filtering hypothesis suggests that children preferentially recall information that is consistent with gender stereotypes (Stangor, & Ruble, 1989b). Moreover, 6-year olds may distort stereotype-inconsistent behavior in memory (e.g., by incorrectly recalling that a boy played with a train if, in fact, it was a girl; Martin, & Halverson, 1983). Because stereotype-inconsistent behavior may be forgotten more easily than stereotype-inconsistent information, the preferential recall of stereotype-consistent information plays an important role in stereotype maintenance (Fyoch, & Stangor, 1994).

As an alternative to the schematic filtering-hypothesis, it has been argued that stereotype-inconsistent behavior is remembered better than stereotype-consistent information because unexpected information requires more elaboration (the *elaboration hypothesis*). In line with the elaboration hypothesis, a meta-analysis of 54 studies (Stangor, & McMillan, 1992) suggests that unexpected information is remembered slightly better than expectancy-consistent information, although there are several moderators of this effect (e.g., type of memory task and strength of the expectancy). It may be efficient to ignore unexpected information (Sherman, & Frost, 2000), but there are several situations in which people will nevertheless attend to unexpected information, for example if they are outcome-dependent on the target person (Erber, & Fiske, 1984), or if they are low in

Need for Closure (Dijksterhuis et al., 1996). Moreover, when people have ample time and resources to elaborate on the information provided to them, they may have better memory for stereotype-inconsistent information because they have more opportunities to elaborate on this information (Dijksterhuis, & Van Knippenberg, 1995; Macrae, et al., 1993, but see Sherman, & Frost, 2000 for an alternative view).

In sum, research on stereotype-consistency-effects in memory suggests that, depending on the context, people may have better memory for either stereotype-consistent or stereotype-inconsistent information. People's ability to attend to stereotype-inconsistent information is an important moderator of this effect, and people may attend to stereotype-inconsistent information if they are sufficiently motivated to do so (e.g., Dijksterhuis et al., 1996; Erber, & Fiske, 1984). In line with this reasoning, we conceive that, if people's motivation to protect the gender hierarchy is heightened, they may show better memory for stereotype-inconsistent behaviors that may threaten the status quo (i.e., proscriptive stereotype violations).

In Study 1, we tested if individual differences in motivation to protect the status quo (i.e., ratings on a Gender System Justification Beliefs scale; GSJB; Jost, & Kay, 2005) predicted memory for gender-inconsistent behaviors (compared to gender-consistent behaviors). In Study 2, we tested if a system threat prime (cf. Kay et al., 2009; Rudman et al., 2012a) affected memory for stereotype-inconsistent behavior. In both studies, we used behaviors that constitute a proscriptive stereotype violation (i.e., behaviors that were considered atypical and undesirable for men/women). We expect that, when the motivation to protect the status quo is heightened (either chronically or experimentally induced), perceivers will show better recall for proscriptive stereotype violations (compared to stereotype-consistent behaviors).

Study 1

Method

Overview and design. The goal of Study 1 is to investigate if people with a higher chronic motivation to protect the gender status quo remember relatively more proscriptive stereotype violations (compared to

stereotypical behaviors). The study has a mixed design with stereotype-consistency (stereotype-violating versus stereotypical) as within-subjects variable and the score on a system justification-scale as between-subjects predictor. For exploratory reasons, we included both a recall-measure and a recognition memory-measure, since stereotype-consistency effects may not necessarily be identical for these types of tasks (Stangor, & McMillan, 1992). Moreover, we included two individual difference measures of sexism and gender stereotypes to explore if these measures affected memory for gender-inconsistent behaviors.

Participants. Three-hundred and two people participated in Study 1 through Amazon's Mechanical Turk (MTurk) in exchange for $1. Forty-nine participants were excluded because they failed an instructional manipulation check (Oppenheimer, Meyvis, & Davidenko, 2009) or could not correctly answer a single memory question[1]. The final sample consists of 253 participants (95 males) between the ages of 18 and 81 (average age 37).

Procedure. After giving informed consent, participants were instructed that they would see several sentences, one by one, accompanied by a picture of a person. Their task was to read the sentences carefully and try to form an impression of the person. Next, they were exposed to twelve pictures, each accompanied by a behavioral sentence. Each trial (consisting of one picture plus a behavioral sentence) was presented on screen for 7 s, with 500 ms between trials. After presentation of the trials, participants completed a short filler task consisting of fifteen anagrams of American cities. After the filler task, participants again saw all the pictures that had previously been coupled to the sentences, and they were instructed to write down everything they still knew about the person in the picture (free recall-task). Next, participants were asked to couple all behaviors to the pictures (recognition-task) and to complete three individual difference measures, namely a gender-warmth/power IAT, Gender System Justification Beliefs Questionnaire (GSJB; Jost, & Kay, 2005), and the Ambivalent Sexism Inventory (ASI; Glick, & Fiske, 1996). The IAT and ASI were included to

[1] Criteria for excluding participants were determined a priori. The results of Study 1 show the same pattern of significant and non-significant findings regardless of whether these participants were included in the data analysis or not.

explore if these frequently used tasks predicted memory for proscriptive stereotype violations. Because the ASI and gender-warmth/power-IAT do not pertain to people's motivation to protect the status quo, the SIH does not entail clear predictions for these measures, and we merely included them for exploratory purposes.

Materials

Memory task. In the memory task, participants were exposed to twelve behavioral sentences: four filler sentences and eight critical sentences. Trial order was randomized, with exception of the first and last trial, which was always a filler sentence.

The eight critical trials consisted of four behaviors that were proscribed for males and four behaviors that were proscribed for females. These behaviors were randomly coupled to a male or female name. Thus, four of the male proscriptions were coupled to a female picture (making it a stereotype-consistent trial), the other four were coupled to a male name and picture (making it a stereotype-inconsistent trial). Likewise, four of the female proscriptions were coupled to a male name and picture (making it a stereotype-consistent trial), the other four were coupled to a female name and picture (making it a stereotype-inconsistent trial). We followed this counterbalancing procedure to ensure that differences in recall between the conditions (stereotype-consistent and stereotype-inconsistent) could not be due to idiosyncrasies of the behavior, as a behavior that was stereotype-consistent for one participant would be stereotype-inconsistent for another participant, and vice versa.

The behaviors that were used as stimuli in the present research were all proscriptions: they were expectancy-violating as well as norm-violating. In a pretest, four different groups of MTurk-participants (*N*s 37 to 44 per group) indicated how desirable they thought the behaviors are for men, how desirable they are for women, how typical they are for men, or how typical they are for women. Differences in desirability and typicality for men and women were established by calculating the effect size of the male-female difference (Cohen's *d*). To ensure that the behavioral stimuli constituted strong proscriptive stereotype violations, we selected behaviors that had large differences in perceived typicality and desirability.

Table 1. *Sentences used in Study 1 and effect sizes (Cohen's d) for the male-female differences in typicality (typ. d) and desirability (des. d).*

Male proscriptions	**typ. *d***	**des. *d***
He/she complained when he/she broke a nail.	-2.41	-0.73
He/she started to scream when he/she saw a mouse.	-2.63	-0.44
He/she burst into tears when something did not go the way he/she wanted to.	-1.50	-0.40
He/she was extremely nervous when he/she had to give a talk.	0.04	-0.35
Average	**-1.63**	**-0.48**
Female proscriptions	**typ. *d***	**des. *d***
He/she hit the table with his/her fist.	2.14	0.55
He/she boasted about the large number of sex partners he/she has had.	2.33	0.67
He/she burped in a pub.	2.57	0.39
He/she made a rude, insulting remark about his/her colleague's work.	0.50	0.20
Average	**1.89**	**0.45**
Filler sentences		
She put on a coat because it was cold outside.		
He went to the store to shop for groceries.		
She played a game with some friends.		
He walks to the train station every morning.		

Note. Positive effect sizes indicate that the behavior is deemed more typical or desirable for men than women, negative effect sizes indicate that the behavior is deemed more typical or desirable for women than men. By convention, Cohen's *d*s of 0.20, 0.50, and 0.80 correspond to small, medium and large effect sizes (Cohen, 1988).

Table 1 contains the stimuli used in the present study, as well as the difference ratings for desirability and typicality (Cohen's *d*) for men versus women (which were derived from the pretest). On average, the male proscriptions were considered more typical and desirable for women than men (Mean Cohen's *d*s -1.63 and -0.48, respectively), whereas the female proscriptions were considered more typical and desirable for men than women (Mean Cohen's *d*s 1.89 and 0.45, respectively)[2]. Thus, the stimuli we selected were proscriptions, in the sense that they were considered atypical and undesirable for men or women (cf. Rudman et al., 2012a). Because the male and female behaviors differed in length and were not exactly matched in terms of their perceived typicality and desirability, the present study did not aim to test for memory differences for male and female proscriptions: rather, we aimed to look at the effect of stereotype-consistency more generally.

The critical sentences were randomly coupled to pictures taken from the Radboud Faces Database (RaFD; Langner et al., 2010). We selected frontal images of four males and four females with neutral facial expressions.[3] The filler sentences were also coupled to male and female pictures from the Radboud Faces Database, but this coupling was not random, and the filler sentences were not analyzed.

Recall tasks. In the free recall-task, participants saw the pictures that had been coupled to the behavioral sentences and were asked to write down everything they still knew about the person in the picture. Two independent coders (who were blind to the hypothesis of the study, as well as to participants' scores on the individual difference variables) coded the free recall task. Each sentence was divided into three equal parts by the experimenter, and coders awarded one point for each part that was remembered correctly. They used a liberal coding scheme, in which not only the exact wording was counted as correct, but also the gist of that part of the sentence (e.g., if participants recalled "the desk" instead of "the table" in the sentence "hit the table with his/her fist", this was coded as

[2] By convention, small, medium and large effect sizes correspond to a Cohen's *d* of 0.20, 0.50 and 0.80, respectively (Cohen, 1988).

[3] The female pictures were from models 1, 12, 19 and 37; the male pictures were from model 7, 30, 49 and 71 from the Radboud Faces Database.

correct). Inter rater reliability was good (a =.87), and any differences between the coders were resolved through discussion.

Next to the free recall-task, participants completed a recognition-task. In this task, participants saw all of the behavioral sentences they had seen before, one by one. However, this time, they had to click on the picture of the person who had previously been associated with the behavior. It turned out that this task was much too easy for participants: only 12% of participants made more than two of such errors (out of 12 possible answers) and over 50% of participants made no errors at al. Due to this floor effect, there was so little variance in the data that we chose not to analyze the recognition task.

Individual difference measures. We incorporated the following three individual difference measures in our experiment: a gender-power/warmth IAT, GSJB-scale and the ASI. We will discuss each of these measures in turn.

Gender-power/warmth IAT. The gender-power/warmth IAT measures implicit associations between men/women and power/warmth. The IAT consisted of male names (*Mark, David, Bob, Jason* and *Matthew*), female names (*Sarah, Amy, Barbara, Michelle* and *Laura*), words related to power (*authority, assert, strong, dominant* and *command*) and words related to warmth (*kind, nice, gentle, sweet* and *caring*). The IAT was administered and analyzed in line with the recommendations made by Greenwald, Nosek and Banaji (2003). On average, participants associated warmth with women/power with men, M (d-score) = 0.34, SD = 0.27.

Gender System Justification Beliefs questionnaire (GSJB). The GSJB questionnaire (Jost, & Kay, 2005) is a scale that measures people's motivation to protect and maintain the status quo. This scale consisted of seven items (e.g., "In general, relations between men and women are just and fair" and "Society is set up so that men and women usually get what they deserve") and had reasonable reliability (a = .76; M = 3.67, SD = 0.76). Participants responded to these questions on a 6-point Likert-scale (anchors: 1 "strongly disagree", 6 "strongly agree").

Ambivalent Sexism Inventory. The ASI (Glick, & Fiske, 1996) is a commonly used measure of sexism that consists of 22 items (α = .90; M = 3.11, SD = 0.76). The scale contains items that measure people's antipathy against powerful women (example items: "women are too easily

offended", "women seek to gain power by getting control over men"). It also contains items that measure endorsement of chivalrous attitudes towards traditional women (example items: "a good woman should be put on a pedestal by her man" and "women should be cherished and protected by men"). Participants responded to these questions on a 6-point Likert-scale (anchors: 1 "strongly disagree", 6 "strongly agree").

Results

Free Recall Task. Before turning to our main analysis, we first tested if there was an overall effect of stereotype-consistency on memory. To test this, we conducted a repeated measures analysis with trial type (stereotype-violating versus stereotypical) as within subjects-variable. There was a significant effect of trial type, $F(1, 252) = 11.28$, $p < .001$, $\eta_p^2 = .04$. On average, participants remembered 0.44 elements (out of 3; $SD = 0.36$) of each stereotype-violating behavior and 0.35 elements (out of 3; $SD = 0.31$) of each stereotypical behavior. Thus, participants showed better memory for stereotype-inconsistent compared to stereotype-consistent behaviors (a selective memory-effect).[4]

---

[4]The present study was not designed to test differences between male and female proscriptions, as the proscriptive stereotype violations we selected were not matched in length and in terms of how desirable/ typical they were (see Table 1). If, for exploratory reasons, we look at the difference between memory for proscriptive stereotype violations (versus stereotype-congruent behaviors) for male versus female targets, there is no significant difference, $F(1, 252) = 2.21$, $p = .14$. However, GSJB seems to be a slightly better predictor of selective memory for male proscriptions ($\beta = .20$, $p < .001$) than for female proscriptions ($\beta = .08$, $p = .20$).

Based on previous backlash research (for an overview, see Rudman, Moss-Racusin, Glick, & Phelan, 2012), we did not expect effects of participant sex on memory for stereotype violating versus stereotype-consistent behavior. To test for effects of participant sex, we conducted a regression analysis in which we included participant sex and the interaction between participant sex and GSJB as additional predictors (next to GSJB). In line with our expectations, there was no significant effect of participant sex ($\beta = -0.06$, $p = .39$), and no significant effect of participant sex x GSJB ($\beta = 0.02$, $p = .27$) on selective memory for stereotype-violating behaviors.

Finally, we tested the above-mentioned effects for male and female proscriptions, separately, as it may be the case that male participants had particularly good memory for male proscriptive stereotype violations and that female participants had particularly good memory for female proscriptive stereotype violations (or that these effects interacted with GSJB). We did not find evidence for such effects (βs -1.05 to 0.07, ps .29 to .69).

Next, we turned to our main analysis. We tested if participants with a high chronic motivation to protect the status quo would show better memory for stereotype-violating (compared to stereotype-congruent) behavior. To do so, we calculated a difference score by subtracting scores from typical behaviors from scores on atypical behaviors (so that higher scores reflect better memory for stereotype-violating behaviors, compared to stereotype-consistent behaviors). We performed a regression analysis with the system justification scale as predictor and the difference score as dependent variable. As expected, system justification predicted the difference in memory between stereotype-incongruent and stereotype-congruent sentences, $\beta = .19$, $p = .003$, suggesting that participants with higher motivation to protect the gender status quo had better memory for stereotype-violating behaviors compared to stereotypical behaviors. In line with the Status Incongruity Hypothesis, those who were motivated to protect the status quo seemed to have particularly good memory for gender rule violations. Next, we added the IAT and ASI as separate steps to the regression, but neither the IAT ($\beta = -.04$, $p = .56$) nor ASI ($\beta = .01$, $p = .85$) was a significant predictor of the memory difference-score.

Discussion

In Study 1, we investigated if individual differences in people's motivation to protect the gender status quo predicted memory for proscriptive stereotype violations. In line with our expectations, the results of Study 1 suggest that people who were motivated to protect the gender status quo had better memory for behavior that is norm-violating (e.g., men who complain when breaking a nail, women who hit the table with their fist). Chapter 2 of the present dissertation suggests that people with high GSJB-scores are most likely to engage in backlash against gender deviant women; the present study extends this research by suggesting that people with high GSJB-scores are more likely to remember gender deviant behavior. There were no effects of the IAT or ASI. Although it is always difficult to interpret null results, the SIH predicts that system justifying motives, not (implicit) stereotypes or sexism should predict backlash.

Next to suggesting that system justifying motives enhance memory for proscriptive stereotype violations, Study 1 suggests that people had overall better memory for stereotype-inconsistent behavior. Although many

studies suggest that people show better recall for stereotype-consistent behavior than for stereotype-inconsistent behavior (Fyoch, & Stangor, 1994), this was not the case in the present study. This points to the possibility that the proscriptive component of gender stereotypes influences memory for stereotype-inconsistent behavior, and that stereotype-inconsistency cannot simply be equated with expectancy-inconsistency (cf. Stangor, & McMillan, 1994). Screaming at mice or complaining when breaking a nail, for example, is not merely unexpected for men, it is also considered undesirable, and this proscriptive component may influence memory for these stereotype-inconsistent behaviors.

The data of Study 1 suggest that people who are motivated to protect the status quo have better memory for behavior that violates the gender rules. However, due to its correlational nature, there may be alternative explanations for the findings in Study 1. For example, the system justification scale was administered *after* participants had completed the recall task. Therefore it is possible (though not plausible) that the number of stereotype violating behaviors people recalled influenced their GSJB-scores. To rule out such alternative explanations, in Study 2, system justification beliefs were experimentally manipulated using a system threat prime.

Study 2

Overview and Design

Participants in Study 2 were exposed to a system threat prime: they read an alleged newspaper article about the decline of the American economy (system threat-condition; cf. Chapter 3 of the present dissertation; Kay et al., 2009; Rudman et al., 2012a) or bird watching (control condition). Prior research suggests that such a system threat prime can increase participants' motivation to protect the gender status quo (Kay et al., 2009). After this system threat-manipulation, participants were exposed to behaviors that were stereotype-consistent or stereotype-inconsistent (i.e., proscriptive stereotype violations), and their recall for these behaviors was measured using a free recall task. Study 2 had a mixed design with system threat prime (system threat versus control) as between-subjects-variable and stereotype-consistency (stereotype-consistent versus stereotype-inconsistent) as within subjects variable. We expected that, after a system threat manipulation,

participants would recall more stereotype-violating behaviors (compared to stereotype-consistent behaviors).

Participants. Three-hundred-and-sixty-six Americans participated through Amazon's MTurk in exchange for $ 0.90. Eighty-five participants were excluded because they had already participated in Study 1 (which largely used the same study materials as Study 2), because they did not pass the instructional manipulation check, because they experienced computer malfunction (i.e., had to reboot their computer halfway through the memory task), did not correctly recall even a single item, or because they could not indicate what the article they had read (as part of the system threat manipulation) was about.[5] This resulted in a final sample of 281 participants (133 males) between the ages of 18 and 74 (average age 35).

Procedure. The procedure of Study 2 was highly similar to the procedure of Study 1, with several important differences. First of all, before commencing the memory task, participants read an alleged newspaper article about the decline (system threat-condition) of the American economy (cf. Chapter 2 of the present dissertation; Kay et al., 2009; Rudman et al., 2012a) or about bird watching (control condition). We refrained from using a system affirmation condition (cf. Rudman at al., 2012a) but instead used a more neutral control condition, because the results of a pilot study suggested that MTurk-participants may have trouble believing that America's economy is on the rise. The system threat prime was not directly related to gender, but previous research has suggested that a threat to an existing societal structure increases the need to defend other societal structures, including the gender status quo (Kay et al., 2009; Rudman et al., 2012a).

[5] Criteria for excluding participants were determined a priori. The pattern of results changes when all participants are retained in the study: the omnibus interaction between stereotype-consistency and system threat is not significant when all participants are included ($F < 1$). Within the system threat condition and control condition, however, the pattern of results remains the same. Within the control-condition, there is no effect of stereotype-consistency on memory ($F < 1$). Within the system threat-condition, stereotype violating behaviors are remembered better than stereotypical behaviors, $F(1, 189) = 9.45$, $p = .002$, M(stereotype violating) = 1.30, M(stereotypical) = 1.09.

Participants were told that the study consisted of two unrelated parts and that, in this first part, they would read a newspaper article to pretest materials for an upcoming study. Different fonts were used throughout the experiment to bolster the impression that the system threat-manipulation was not part of the same study as the memory task. After completing the system threat manipulation, participants again completed a memory task, an anagram task, and a free recall task. We did not include a recognition task in Study 2.

Similar to Study 1, the sentences in Study 2 were divided into three parts, and a coder awarded one point for each part of the sentence that a participant remembered correctly (so that every participant had an average score of 0 to 3 points per sentence). This coder was blind to the study's hypotheses and to the between subjects-condition. We used a different, more liberal coding scheme, so that the average number of points participants received in Study 2 was higher than in Study 1. Instead of having a second coder code all the sentences, we created a computer script to validate the coder's ratings. This computer script awarded points for every word participants remembered correctly (or a close synonym). The coder's ratings and computerized ratings were highly correlated (r = 0.97). In the analyses reported below, we used the ratings of the human coder.

Materials

System threat manipulation. In the system threat-condition, participants read an alleged newspaper article about the decline of the American economy. In the control-condition, participants read an alleged newspaper article about the increasing popularity of bird watching (see Appendix 1). After reading one of these two articles, participants were given three minutes to write about why the author's position was justified. Finally, to bolster our cover story, they indicated how well-written, compelling, interesting and understandable they thought the article was.[6]

[6] The articles on bird watching and the decline of America's economy were rated as equally clear, $F(1, 279) = 1.03, p = .31$, and understandable, $F < 1$. However, the article on America's decline was rated as more interesting than the bird watching-article, $F(1,279) = 16.42, p < .001$, M(America) = 5.49 (SD = 1.44), M(bird watching) = 4.76 (SD = 1.57). It was also rated as more compelling, $F(1, 279) = 27.81, p < .001$, M(America) = 5.10 (SD = 1.57), M(bird watching) = 4.11 (SD =1.57).

Memory task. The memory task was identical to the task used in Study 1, but with slightly different stimuli. Table 2 contains an overview of the stimuli that were used in Study 2. Although we largely used identical behaviors in Study 1 and Study 2, we replaced the behaviors with small differences in desirability between men and women by behaviors with larger differences, so that all stimuli in Study 2 were clearly proscribed.

Table 2. *Sentences used in Study 2 and effect sizes (Cohen's d) for the male-female differences in typicality (typ. d) and desirability (des. d).*

Male proscriptions	**typ. *d***	**des. *d***
He/she complained when he/she broke a nail.	-2.41	-0.73
He/she started to scream when he/she saw a mouse.	-2.63	-0.44
He/she blushes whenever someone talks to him/her.	-1.05	-1.13
He/she burst into tears when his/her boss criticized him/her.	-1.82	-0.53
Average	**-1.98**	**-0.71**
Female proscriptions	**typ. *d***	**des. *d***
He/she hit the table with his/her fist.	2.14	0.55
He/she boasted about the large number of sex partners he/she has had.	2.33	0.67
He/she burped in a pub.	2.57	0.39
He/she yelled at the referee in a sports match.	1.94	0.40
Average	**2.25**	**0.50**
Filler sentences		
She put on a coat because it was cold outside.		
He went to the store to shop for groceries.		
She played a game with some friends.		
He walks to the train station every morning.		

Note. Positive effect sizes indicate that the behavior is deemed more typical or desirable for men than women, negative effect sizes indicate that the behavior is deemed more typical or desirable for women than men. By convention, Cohen's *d*s of 0.20, 0.50, and 0.80 correspond to small, medium and large effect sizes (Cohen, 1988).

Results

We conducted a repeated measures analysis with stereotype-consistency (stereotype-consistent versus stereotype-inconsistent) as within subjects-variable and system threat (system threat versus control) as between subjects-variable. There was a main effect of stereotype consistency, $F(1, 279) = 5.80$, $p = .02$, Cohen's $d = 0.17$. On average, participants remembered more elements of stereotype violating sentences ($M = 1.35$, $SD = 0.80$) than of stereotypical ones ($M = 1.21$ $SD = 0.82$).

Table 3. *Mean number of elements participants remembered of stereotype violating and stereotypical sentences in Study 2, per condition.*

	stereotypical behavior	**stereotype-violating behavior**
control condition	1.26 *(SD = 0.85)*	1.27 *(SD = 0.80)*
system threat condition	1.17 *(SD = 0.80)*	1.42 *(SD = 0.80)*

This main effect of stereotype consistency was qualified by an interaction with system threat. We expected that, after a system threat prime, participants would have better memory for stereotype-inconsistent (compared to stereotype-consistent) behaviors. In line with this expectation, there was a significant interaction between system threat (system threat versus control) and stereotype-consistency (stereotype-consistent versus stereotype-inconsistent), $F(1, 279) = 4.71$, $p = .03$. In the control condition, there was no effect of stereotype-consistency, $F < 1$. As depicted in Table 3, participants in the control condition remembered an equal number of elements of stereotype-violating sentences ($M = 1.27$, $SD = 0.80$) and stereotypical sentences ($M = 1.26$, $SD = 0.85$). In the system threat condition, there was the expected effect of stereotype-consistency on memory, $F(1,147) = 11.44$, $p = .001$, Cohen's $d = 0.31$. On average, participants remembered more elements from stereotype-inconsistent

sentences (M = 1.42, SD = 0.80) than from stereotype-consistent (M = 1.17, SD = 0.80). [7]

Discussion

The results of Study 2 suggest that participants have better memory for norm-violating behaviors after a system threat prime. Participants recalled more elements of stereotype-violating behaviors (compared to stereotype-consistent behaviors) if they had been threatened with the decline of the American economy, but not if they had read an article on bird watching. When the motivation to protect the status quo had been temporarily heightened, participants had better memory for behaviors that are a potential threat to the status quo.

Although there was an overall effect of stereotype-consistency on memory in Study 1, there was no such effect in the control condition of Study 2. The difference between these findings may be due to the article participants read in the control-condition in Study 2. Because this article contains references to "America's natural beauty", it may have affirmed participants' belief in the greatness of their country, and thereby, in the social system. Although speculative, it is possible that our manipulation was not fully neutral, but may have worked as a system affirmation-prime (cf. Kay et al., 2009).

Interestingly, the system threat-manipulation that we employed improved recall for proscriptive stereotype violations even though the manipulation was not specifically geared towards increasing the motivation to protect the *gender* status quo. In line with previous research (Kay et al., 2009; Rudman et al., 2012a), threatening participants with the decline of the American economy increased their motivation to protect the gender

[7] The present study was not designed to test differences between male and female proscriptions, as the proscriptive stereotype violations we selected were not matched in terms of length and how desirable/ typical they were (see Table 2). If, for exploratory reasons, we look at the difference between memory for proscriptive stereotype violations (versus stereotype-consistent behaviors) for male versus female targets, there is no significant difference, $F < 1$.

We did not expect effects of participant sex on memory for stereotype violating versus stereotypical behavior, and indeed there were no significant effects of participant sex, $F < 1$.

hierarchy. This seems in line with other research suggesting that a threat to the sociopolitical system can increase people's motivation to protect the gender hierarchy (Kay et al., 2009), as well as with general theorizing suggesting that people can ward of threats to one psychological domain by reaffirming meaning in other domains (a phenomenon called *fluid compensation*; Heine, Proulx, & Vohs, 2006).

General Discussion

The present research suggests that when people's motivation to protect the status quo is heightened (either experimentally induced or dispositionally), they are more likely to recall proscriptive stereotype violations (e.g., the behavior of a man who screams at a mouse or a woman who hits the table with her fist). In Study 1, individual differences in people's motivation to protect the gender status quo predicted memory for proscriptive stereotype-violating behavior (compared to stereotype-consistent behavior). In Study 2, people who had been threatened with the decline of the American economy showed better memory for proscriptive stereotype-violating behavior (compared to stereotype-consistent behavior). In line with the Status Incongruity Hypothesis, these results suggests that system justifiers have better memory for behaviors that may jeopardize the status quo (i.e., proscriptive stereotype violations).

There are several processes through which the motivation to protect the gender status quo may influence memory for proscriptive stereotype violations, and more research is needed to study through which process this effect occurs. One possibility is that proscriptive stereotype violations are highly goal-relevant for people who are motivated to protect the status quo. Because of this, system justifiers may be more likely to elaborate on them, which may lead to better encoding of these behaviors in memory. However, because this process was not tested directly, it is unclear if the enhanced memory for proscriptive stereotype violations that was observed in the present research is a result of better encoding or of a recall advantage (cf. Sherman, & Frost, 2000).

The present results have clear implications for backlash research. As implied by the epithet "forgive and forget", people's memory for behavior is strongly related to their propensity to penalize or forgive others, as people cannot sanction others for transgressions that they do not

remember. By meticulously remembering how gender deviants behaved, people who are motivated to protect the status quo have the possibility to sanction gender deviants for this behavior later on. Instead of forgiving and forgetting gender deviant behavior, people who are motivated to protect the gender hierarchy are more likely to remember these behaviors. In so doing, they may be more likely to engage in backlash against gender deviants. Because fear of backlash obstructs stereotype disconfirmation (as people are afraid to show atypical behavior for fear of being sanctioned; Moss-Racusin, & Rudman, 2010; Prentice, & Carranza, 2004; Rudman, & Fairchild, 2004), the memory effects observed in the present research may play a role in stereotype maintenance. Interestingly, prior research suggests that enhanced memory for stereotype-*consistent* behavior contributes to stereotype maintenance. The present research poses that, perhaps, enhanced memory for stereotype-*inconsistent* behavior may also contribute to stereotype maintenance to the extent that it contributes to backlash.

Although more research is needed to clarify the role of memory processes in backlash and stereotype maintenance, the present research suggests that people who are motivated to protect the gender hierarchy are not likely to forget those who transgress the gender rules. In the present study, people who were chronically motivated to protect the gender status quo showed better memory for behaviors that may jeopardize the status quo, and so did people who had been threatened by the decline of the American economy. System justifying motives, then, curbed people's propensity to "forgive and forget", encouraging them to keep an eye out for any behavior that may threaten their world view.

Appendix 1

Alleged newspaper articles used in Study 2. In the system threat-condition, participants read the following article, entitled "America in Decline":

> *"These days, many people in the United States feel disappointed with the nation's condition. Whether it stems from the economic meltdown and persistent high rates of unemployment, fatigue from fighting protracted wars in the Middle East that have cost America dearly in blood and treasure, or general anxieties regarding global and technological changes that the government seems unable to leverage to their advantage, Americans are deeply dissatisfied. Many citizens feel that the country has reached a low point in terms of social, economic, and political factors. It seems that many countries in the world are enjoying better economic and political conditions than the U.S. In recent nationwide polls, more Americans than ever before expressed a willingness to leave the United States and emigrate to other nations."*

In the control-condition, participants read the following article, entitled "Bird watching":

> *"These days, many people in the United States enjoy bird watching as a recreational activity. Whether this popularity stems from the opportunities bird watching provides to be outside and get in touch with nature or from demographic trends, an increasing number of Americans is deeply committed to observe birds in their natural habitat. Many people feel that bird watching plays an important role in wildlife preservation because bird watchers help keep track of bird populations in the U.S. Moreover, bird watching can be lots of fun because it provides great opportunities to get some fresh air and enjoy America's natural beauty. In recent nationwide polls, more Americans than ever before expressed an interest in watching birds and other wildlife."*

CHAPTER 5

Spontaneous Backlash for Gender Atypicality

Acknowledgements

We would like to thank Inge Huijsmans and Veronique Louhenapessy for their help with data collection, and Laurie Rudman for her help with the design of Study 3. This research was funded with the generous support of the APA Geis Memorial Award for Dissertation Research.

Gender atypical targets (such as agentic women and communal men) are at risk for social and economic penalties (termed *backlash effects*; Rudman, & Phelan, 2008). Although backlash emerges in (seemingly) deliberate inferences, such as hiring decisions and leadership evaluations, the present line of research aims to examine if there may be a more subtle, spontaneous form of backlash as well. In the first chapter of the present dissertation, we started out studying backlash with very deliberate, explicit measures (e.g., liking), and gradually moved to study more spontaneous forms of backlash (e.g., memory effects). In the present chapter, we will study if gender stereotypes can occur unintentionally, by spontaneously affecting the trait inferences people form of behavior. Specifically, we wonder if people may engage in "spontaneous backlash" by forming more extreme inferences of norm-violating behavior.

Research on Spontaneous Trait Inferences (STIs) suggests that perceivers form impressions of others without intention, and with relatively little effort (Uleman, Newman, & Moskowitz, 1996). For example, when participants read "The professor wins a science quiz", "smart" is spontaneously activated, as indicated by perceivers taking a relatively long time to decide that "smart" was not part of the original sentence. STI-formation can be biased by the stereotypic expectancies perceivers hold about a target, in that stereotype-inconsistent trait inferences are inhibited (Wigboldus, Dijksterhuis, & van Knippenberg, 2003; Wigboldus, Sherman, Franzese, & Van Knippenberg, 2004; Yan, Wang, & Zhang, 2012; the *inhibition-effect*). For example, when perceivers read "the garbage man won the science quiz" they are less likely to activate "smart", compared to when they read "the professor won the science quiz". According to the authors (Wigboldus et al. 2003; 2004), activation of a category (such as garbage men) may inhibit the accessibility of stereotype-incongruent traits (such as intelligence), so that perceivers are temporarily less likely to spontaneously form a dispositional inference based on the incongruent behavior. More recently, other researchers have argued that perceivers form a situational inference for stereotype-incongruent behavior. In this view, instead of inferring that a garbage men who wins a science quiz is smart, perceivers infer that the science quiz must have been easy (Ramos, Garcia-Marques, Hamilton, Ferreira, & Van Acker, 2012).

The stereotypes people hold about garbage men and professors are descriptive: they contain clear expectations about how garbage men and professors are typically expected to behave. As such, descriptive stereotypes stipulate *expectancies* about how group members typically behave. Gender stereotypes are often prescriptive in nature: apart from specifying how men and women are expected to behave, they specify how men and women should behave (stereotypes specifying how men and women should *not* behave are termed proscriptions; Prentice, & Carranza, 2002; 2004). As such, prescriptive stereotypes stipulate *norms* about how group members ought (not) behave. We predict that descriptive and proscriptive stereotypes differentially affect the STI-formation process, such that descriptive stereotype violations are inhibited, but proscriptive stereotype violations become stronger.

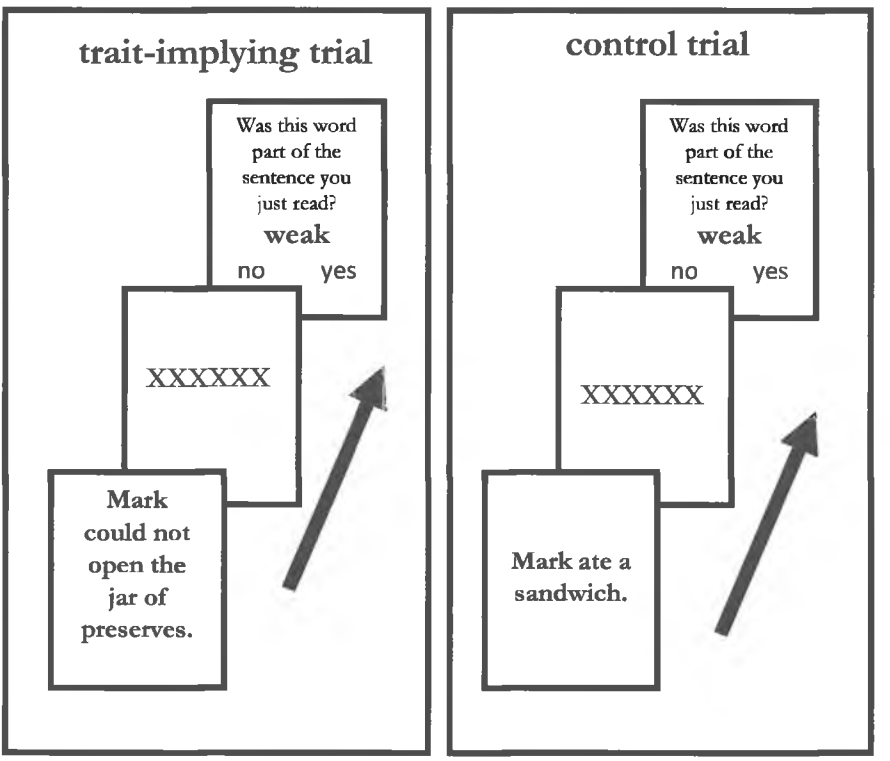
If behavior is not only descriptive (*expectancy-violating*) but also proscriptive (*norm-violating*), spontaneous trait inferences of counterstereotypical behaviors may not necessarily be inhibited. Proscriptive stereotype violations (e.g., dominant behavior for females and weak behavior for males) can jeopardize the gender hierarchy and thereby threaten people's needs for certainty and stability (Jost, & Hyunyady, 2002). Because people are strongly motivated to penalize those who engage in proscriptive stereotype violations (Rudman, Phelan, Moss-Racusin, & Nauts, 2012a), we believe that it is plausible that people will not readily discount them. People may disregard a garbage man who wins a science quiz, but we wonder if they will likewise disregard a man who screams at a mouse.

We expect to find the classic inhibition-effect (Wigboldus et al., 2003; 2004) for descriptive stereotype violations. However, for proscriptive stereotype violations, we wonder if people may spontaneously form stronger inferences of proscriptive stereotype violations (an *amplified STI-effect*). The reason for this expectation is that, on an explicit level, inferences of proscriptive stereotype violations are more extreme than inferences of stereotypical behaviors: for example, a woman who behaves a little dominantly is regarded as much more dominant than a man who shows the same behavior (the *dominance penalty*; Rudman et al., 2012a). Moreover, according to attribution theory, perceivers form stronger trait inferences of norm-violating behavior because it is more distinctive, and therefore more

diagnostic, than normative behavior (Kelley, & Michela, 1980). Although Spontaneous Trait Inferences are formed without intention, these inferences are not necessarily implicit, unconscious, or automatic (Uleman, Newman, & Moskowitz, 1996). People can form trait inferences spontaneously, without intending to form an inference of someone's behavior, and without being asked to do so. However, this does not mean that people are necessarily unaware of having formed such an inference, or cannot deliberate on these inferences. As such, attributional processes may be relevant to the formation of trait inferences that are formed spontaneously, and the desirability of a behavior may affect the strength of trait inferences even if people are not explicitly asked to form an impression of someone's behavior.

The present chapter consists of three studies that were set up to study such an amplified STI-effect. All three studies employed a Probe Recognition Task (PRT; McKoon, & Ratcliff, 1986) to investigate the effect of stereotype violation on STI-formation. As depicted in Figure 1, perceivers in a PRT are exposed to trait-implying sentences or control (neutral) sentences, and have to indicate whether a probe word was part of the sentence they just read or not. Longer latencies for trait-implying compared to control sentences indicate that perceivers need more time to indicate that a trait-word was not part of the sentence if the trait is implied by the sentence, suggesting they formed STIs. In Study 1, we investigated if stereotype-consistency influenced the strength of STIs for proscriptive stereotype violations. In Study 2, we investigated if stereotype-consistency differentially influenced STI-strength for behaviors that were proscriptive (norm-violating) or descriptive (expectancy-violating) and added measures of sexism (Glick, & Fiske, 1996) and system justification beliefs (Jost, & Kay, 2005). Study 3 contained a replication of Study 2, but with category labels instead of names, and with a mixed design instead of a within subjects-design.

Figure 1: Probe Recognition Task (PRT)



Study 1

In Study 1, we investigated if perceivers would form stronger Spontaneous Trait Inferences of proscriptive stereotype violations compared to stereotypical behaviors. Participants completed a Probe Recognition Task that included stereotype-inconsistent behaviors (which were all proscriptive stereotype violations) as well as stereotypical behaviors, combined with control trials that were matched to the stereotype-inconsistent and stereotype-consistent trials. Thus, the study had a 2 (consistency: stereotype-inconsistent versus stereotype-consistent) by 2 (trial type: trait-implying sentence versus neutral control sentence) within subjects-design. We expected that participants would form stronger STIs of stereotype-inconsistent compared to stereotype-consistent behaviors, as apparent in

longer response latencies for stereotype-inconsistent than stereotype-consistent trials (corrected for latencies on control trials).

Method

Participants. 118 Radboud University students participated in exchange for partial course credit or a €5 gift certificate. Nine participants were removed from the dataset for the following reasons: because they were not a native speaker (three participants), because their average response latency was more than three standard deviations higher than the average latency (five participants), or because they had an error rate higher than 20% (one participant).[1] This resulted in a final sample of 109 participants (48 men) between the ages of 18 and 35 (mean age 22).

Overview and design. The PRT used in the present study consisted of three types of trials: trait-implying trials, control trials, and filler trials. Every participant was exposed to two types of trait-implying trials: stereotype-consistent trials (e.g., "Anne was unable to open a jar of preserves", implying "weak") and stereotype-inconsistent trials (e.g., "Mark was unable to open a jar of preserves", implying "weak"). All trait-implying trials were proscriptions: the behaviors in these trials violated stereotypic expectancies of how men and women are expected to behave (expectancy violation), as well as norms of how men and women should behave (norm violation). The present study contained only proscriptive stereotype violations, and no descriptive stereotype violations.

In addition to the trait-implying trials, the task included neutral control trails. These trails contained sentences that were not trait-implying and were followed by the same probe words as the trait-implying sentences (e.g., "Peter ate a sandwich" followed by the probe word "weak"). Moreover, the experiment contained several types of filler trials, which were included to ensure that the correct answer to the question "was this word part of the sentence you just read?" was not always "no".

To test our hypothesis, we analyzed only trait-implying trials and control trials; filler trials were not analyzed. We expected that participants

[1] These exclusion criteria were established a priori, before any analyses were conducted; including these participants in the analyses does not alter the pattern of results.

would need more time to correctly indicate that the probe word was not part of the sentence when the trait had been implied by the sentence, indicating that they formed a spontaneous trait inference. In other words, we expected that perceivers would respond slower to the probe word "weak" after reading that "Mark was unable to open a jar of preserves" than after reading "Peter ate a sandwich", suggesting that they spontaneously inferred that Mark is weak. This difference between trait-implying and control sentences constitutes the basic STI-effect. More relevant to the current study, we expected that people would form stronger STIs (i.e., the STI-effect would be larger) for behaviors that constitute a proscriptive stereotype violation (i.e., for stereotype-inconsistent trials) than for behaviors that do not constitute a proscriptive stereotype violation (i.e., for stereotype-consistent trials). This effect should be mirrored in an interaction between trial type (trait-implying versus control) and consistency (stereotype-consistent versus stereotype-inconsistent).

Procedure. Upon arriving in the lab, participants were seated in individual cubicles and were instructed by the experimenter. They were told that several sentences would be presented on screen, one by one, and that each sentence would be followed by a probe word. Participants' task was to indicate whether this word was part of the sentence they had just read or not by pressing the "a" or "l" key on their keyboard. They were instructed to do this as quickly as possible, but without making too many errors. Sentences were presented on screen for 3300 ms (with 200 ms between trials). After a short practice block (eight trials), participants completed two blocks of 44 trials each. After the first block, participants took a short break, followed by three more practice trials and the second block of the PRT. Finally, participants responded to several demographic questions and were debriefed and thanked for their participation.

Stimuli. The experiment consisted of 88 trials of the following three types: trait-implying trials (16), control trials (16) and filler trials (56). Trial order was fully randomized. The 16 trait-implying trials consisted of eight behaviors that were proscribed for males and eight behaviors that were proscribed for females. These behaviors were randomly coupled to a male or female name. Thus, four of the male proscriptions were coupled to a female name (making it a stereotype-consistent trial, e.g., "Anne was unable to open a jar of preserves"), the other four were coupled to a male

name (making it a stereotype-inconsistent trial, e.g., "Benjamin had to wipe away a tear at a friend's wedding."). Likewise, four of the female proscriptions were coupled to a female name (making it a stereotype-inconsistent trial, e.g., "Carmen gave the player of the other soccer-team a head butt"), the other four were coupled to a male name (making it a stereotype-consistent trial, e.g., "Michael told everyone that he had the highest grade in class."). We included both proscriptive stereotype violations for males and proscriptive stereotype violations for females in the task to ensure that stereotype-congruency was not confounded with the behaviors that were used.

To ensure that trait-implying stimuli were, indeed, clearly trait-implying, a group of participants (N = 17) read the sentences and wrote down which trait came to mind when they read the sentence (cf. Wigboldus et al., 2003; 2004). Only sentences for which at least 70% of participants spontaneously mentioned the intended trait (or a close synonym) were included as trait-implying stimuli. We selected sentences that implied traits related to dominance and weakness, as dominance is proscribed for women and weakness is proscribed for men (Rudman et al., 2012a; Prentice, & Carranza, 2002). Appendix 1 contains the behaviors that were used as trait-implying stimuli. In addition to the 16 trait-implying trials, the experiment consisted of 16 control trials. Control trials consisted of behaviors for which most participants in the pretest indicated that they could not come up with a trait that matched the behavior. Examples of control sentences are "ate a sandwich" and "put a stamp on an envelope and put it in the mailbox". Each trait-implying sentence was matched by a control sentence with the exact same gender and probe word. For example, the trait-implying sentence "Mark was unable to open a jar of preserves" (probe word "weak") could be matched by a control sentence with a male target and the same probe word (e.g., "Peter ate a sandwich", probe word "weak"). Thus, trait-implying trials and control trials were matched in terms of target gender and probe so that each trait-implying sentence had its own control sentence. To make sure that control sentences were not always matched to the same trait-implying sentences, we randomized which control behavior was matched to which trait-implying behavior. Thus, whereas the sentence "was unable to open a jar of preserves" may have been coupled to "ate a sandwich" for one participant, it was coupled to

another control behavior (e.g., "put a stamp on an envelope and put it in the mailbox") for another participant.

Matching trait-implying stimuli to control stimuli is important for two reasons. First of all, by using the same probe words for a trait-implying sentence and its control sentence, any differences between words (e.g., in word length or word frequency) could not influence differences in response latencies to probes between trait-implying trials and control trials. After all, the probes were the same. Second, by using names of the same gender for a trait-implying sentence and its control sentence, we controlled for semantic priming-effects. When people read a male or female name, this may already lead to the activation of stereotype-consistent traits and the inhibition of stereotype-inconsistent traits (Dijksterhuis, & Van Knippenberg, 1997). For example, "weak" is unexpected for males, so the baseline level of activation of "weak" may be inhibited when people just read a male name. By matching every trait-implying trial to its own control trial (with the same gender and probe word), we controlled for semantic priming-effects that may occur regardless of the behavioral information that was presented.

In addition to trait-implying trials and control trials, the experiment contained 56 filler trials. These trials were not analyzed: their sole purpose was to balance the number of yes-and no-responses in the task. The filler trials consisted of sentences that did or did not imply traits, with traits, verbs and other words (e.g., objects) as probes. Filler sentences were carefully crafted to make it difficult for participants to respond strategically. Because sentences could be followed by a word that was part of the sentence or by a word that was not part of the sentence, and because the correct answer to every type of probe (traits, verbs and other words) could be either "yes" or "no", participants could not predict what the correct answer should be based on the type of sentence or probe.

Results

Data preparation. Before conducting the analyses, all incorrect trials (2% of trials) were removed from the dataset, as well as latencies faster than 200 ms and slower than 2000 ms (1.3 % of trials). This is in line with previous work using the PRT (Wigboldus et al., 2003). We performed inverse transformation ($1/x$) to handle the skewed nature of the data, but report untransformed means for ease of interpretation.

Our study had a 2 (stereotype congruency: proscriptive stereotype violation versus no proscriptive stereotype violation) x 2 (trial type: control versus trait-implying) within subjects-design. Before turning to our main analyses, we first tested if participants formed spontaneous trait inferences based on our stimuli (i.e., a basic STI-effect). If participants formed STIs based on our stimuli, they should be slower to indicate that a trait was not part of the sentence if it was preceded by a sentence that implied the trait (trait-implying trial), compared to when it was preceded by a sentence that did not imply the trait (control trial). In line with our expectations, there was a significant main effect of trial type, $F(1,108) = 41.55$, $p < .001$, $\eta_p^2 = .28$. Average latencies were larger for trait-implying trials (716 ms, $SD = 146$) than for control trials (677 ms, $SD = 125$), suggesting that participants formed spontaneous trait inferences. Thus, participants were slower to correctly indicate that a trait (e.g., "weak") was not part of the sentence if the sentence implied the trait (e.g., "Mark could not open the jar of preserves") than if the sentence did not imply the trait (e.g., "Peter ate a sandwich").

Having established that the PRT produced a basic STI-effect, we turned to our main hypothesis[2]. We expected that participants would form amplified STIs of behaviors that are norm-violating, compared to behaviors that are not norm-violating (ie. of stereotype-inconsistent compared to

[2] We were interested in the interaction between stereotype-congruency and trial type (control versus trait-implying) and did not expect differences between male and female norm violations. None of the tasks in the present chapter were designed to test for these effects, and male and female norm violations were not matched, so that any difference could be caused by differences in desirability and typicality of the stimuli, or by differences in word length or word frequency. Taking these limitations into account, we tested if there was an effect of target gender for exploratory purposes. In Study 1, there were no differences in STI-strength between male and female norm violations, $F < 1$.

For all studies, we also performed all analyses with participant sex. We did not expect to find any differences between male and female participants, because prior backlash research has failed to find any effects of participant sex (with the exception of Nauts & Vonk, 2009). Because these analyses result in fifteen or thirty tests of main effects and interactions per study (for Studies 1 and 3 and Study 2, respectively), some effects are bound to be significant purely by chance. However, there was no significant interaction with participant sex that was theoretically meaningful in any of the studies reported in the present chapter (all Fs < 1).

stereotype-consistent trials). This should be reflected in a significant interaction between trial type (trait-implying versus control) and stereotype-consistency (stereotype-consistent versus stereotype-inconsistent). Contrary to our expectations, this interaction was not significant ($F < 1$). The present study did not provide evidence for an effect of stereotype-congruency on STI-strength. Our hypothesis was not supported: we did not find evidence suggesting that inferences of weak and dominant behaviors were influenced by the gender of the person who performed them.[3]

Discussion

The goal of Study 1 was to examine if participants form stronger STIs of stereotype-inconsistent compared to stereotype-consistent behaviors when these behaviors entailed a proscriptive stereotype violation. Although the task used in the present research seemed to have been successful in establishing a basic STI-effect, we did not find evidence for the amplified STI-effect.

The present study did not provide any evidence for an effect of stereotype-congruency on STI-formation. Unfortunately, Fisherian hypothesis testing does not allow researchers to conclude anything based on non-significant results (Cohen, 1990), so all we can conclude from the present study is that we have been unable to find evidence for our hypothesis. Perhaps, STIs of proscriptive stereotype violations are not truly amplified, but the kind of stereotype (descriptive versus proscriptive) may nevertheless moderate the strength of STIs. Put differently, it is possible that stereotypes cannot strengthen STIs, but only inhibit them.

---

[3] Although there was no interaction between stereotype-congruency and trial type, there was a main effect of stereotype-congruency, $F(1, 108) = 4.22$, $p = .04$, M (stereotype-congruent trials) = 701 ms (SD = 132), M(stereotype-incongruent trials) = 692 ms (SD = 138). At first sight, this effect may seem to provide evidence for an effect of stereotype-congruency on STI-formation, but we believe that this is not the case. A main effect of stereotype-congruency can be caused by differences in STI-formation, or by differences in baseline activation levels of traits that are unrelated to STI-formation. A proper amplified STI-effect can only be tested by looking at the interaction between trial type and stereotype-congruency, and this interaction was not significant. Put differently, the main effect of stereotype-congruency suggests that stereotype-incongruent traits (e.g., weakness for males) are inhibited regardless of whether the behavior that is presented is related to this trait or not (which we assume mirrors a semantic priming-effect).

The research by Wigboldus and colleagues (2003; 2004) provides quite some evidence for stereotype-inhibition effects, but not for stereotype-activation effects. Thus, it is be possible that people form such strong inferences of behavior that additional processes cannot further strengthen the STI (i.e., a ceiling effect).

If there was indeed a ceiling effect in Study 1, it may be possible that descriptive and proscriptive stereotype violations nevertheless differentially affect STI-formation. Put differently, it may be the case that descriptive stereotypes, but not proscriptive stereotypes, inhibit STI-formation (replicating Wigboldus et al., 2003; 2004). To test if this is the case, Study 2 included stimuli that are norm-violating (proscriptive stereotype violations), and stimuli that are expectancy-violating, but not norm-violating (descriptive stereotype violations). For descriptive stereotypes, we expected an inhibition of stereotype-inconsistent STIs (Wigboldus et al., 2003; 2004), but we expect that this inhibition-effect would be weaker (or all together absent) for proscriptive stereotype violations.

Study 2

The goal of Study 2 was to investigate the effect of stereotype-consistency on STI-formation for proscriptive and descriptive behaviors. Next to adding descriptive stereotypes, we made some methodological changes to the paradigm used in Study 1. In Study 1, we used behaviors that implied a trait that is known to be proscribed for men or women (as tested by Rudman et al., 2012a and Prentice, & Carranza, 2002), but we did not test if these specific behaviors were proscribed. Moreover, due to cultural differences between the US and The Netherlands, gender norms may differ between these countries, so that traits that are proscribed in the US may not necessarily be proscribed in The Netherlands. In Study 2, we pretested the individual behaviors to make sure that each and every one of them was strongly expectancy-violating, and to make sure that the proscriptions (but not the descriptions) were clearly norm-violating. This procedure allowed us to select better stimuli, and to make sure that the behaviors we chose constituted a norm-violation for our Dutch participants.

Another difference between Studies 1 and 2 is that Study 2 contained more trials. An important reason why studies in the field of social

psychology may provide inconclusive results, or do not always replicate, is because they lack statistical power (Asendorpf et al., 2013; Bakker, van Dijk, & Wicherts, 2012; Cohen, 1962), a pitfall we tried to avoid by increasing the number of trials in our study. Finally, we added several individual difference questionnaires that were expected to affect the effects of stereotype-consistency on STI-formation. Chapter 1 of the present dissertation suggests that people who are motivated to protect the gender status quo (i.e., have high scores on a Gender System Justification Beliefs Questionnaire; GSJB, Jost, & Kay, 2005) are more likely to engage in backlash against gender deviant women, and Chapter 4 suggests that people with high GSJB-scores selectively remember gender deviant behaviors. We hypothesized that high system justifiers would be most likely to show an amplified STI-effect for proscriptive (but not descriptive) trials because they are more motivated to protect the gender status quo. We also added a sexism measure (the Ambivalent Sexism Inventory; ASI, Glick, & Fiske, 1996) to study if people with higher sexism-scores would show stronger stereotype-congruency effects on STI-formation. This questionnaire was added for exploratory reasons.

Method

Participants. Ninety-five students from Radboud University Nijmegen participated in exchange for partial course credit or a €7.50 gift certificate. Seven participants were removed from the dataset because they were not a native speaker (five participants), or because their average response latency was more than three standard deviations above the average latency (two participants)[4]. This resulted in a final sample of 88 participants (41 men) between the ages of 18 and 30 (mean age 22).

Overview and design. The PRT used in Study 2 contained three types of trials: trait-implying trials, control trials, and filler trials. There were two differences in the design of Study 1 and Study 2. First, and most importantly, Study 2 contained four instead of two kinds of trait-implying trials.

[4] These exclusion criteria were established a priori, before any analyses were conducted; including these participants in the analyses does not alter the pattern of results.

Whereas Study 1 contained only stereotype-consistent and stereotype -inconsistent proscriptions, Study 2 contained stereotype-consistent and stereotype-inconsistent proscriptions and descriptions. Thus, the present study had a 2 (trialtype: control versus trait-implying) x 2 (stereotypicality: stereotype-consistent versus stereotype-inconsistent) x 2 (kind of stereotype: descriptive versus proscriptive) within subjects-design.

A second difference between Study 1 and Study 2 is that we increased the number of trials to increase the power of our design. Whereas Study 1 consisted of 16 control and 16 trait-implying trials (8 stereotype-consistent, 8 stereotype-inconsistent), Study 2 consisted of 80 control trials and 80 trait-implying trials (20 stereotype-consistent about a proscription, 20 stereotype-inconsistent about a proscription, 20 stereotype-consistent about a description, and 20 stereotype-inconsistent about a description). Because participants' performance may suffer if the task would become too long, we included relatively fewer filler trials in Study 2 compared to Study 1. Instead of including sufficient filler trials to get a 50/50 balance in yes/no-responses, we used sufficient filler trials so that two thirds of the responses yielded a no-response and one third yielded a yes-response.

Procedure. The procedure that was used in Study 2 was identical to the procedure that was used in Study 1, with three exceptions. First of all, participants completed a total of 300 trials instead of 88 trials. Second, we removed the break halfway through the experiment because several participants in Study 1 had indicated disliking the break (because it interrupted the flow of the experiment). Third, participants completed two additional measures after the PRT: the Gender System Justification Beliefs Questionnaire (GSJB-scale; Jost, & Kay, 2005), and the Ambivalent Sexism Inventory (ASI; Glick, & Fiske, 1996).

Stimuli.

Probe Recognition Task. The PRT used in the present study consisted of 300 trials of the following three types: trait-implying trials (80), control trials (80) and filler trials (140). Trial order was fully randomized. The 80 trait-implying trials consisted of 20 trials that were norm-violating for males, 20 that were norm-violating for females, 20 that were expectancy-violating for males, and 20 that were expectancy-violating for females. There were 40 trait-implying behaviors in total (see Table 2); each behavior was used both with a male and with a female name. Unlike Study

1, in Study 2 stimuli were used more than once (so that the same stimulus would appear both as a stereotype-consistent trial and as a stereotype-inconsistent trial within the same experiment).

In a pilot study, behaviors were pretested to ensure that they were sufficiently trait-implying. Only traits for which at least 70% of the pilot study participants (N = 17) spontaneously mentioned the intended trait (or a close synonym) were included. Moreover, four different groups of participants (N = 27 per group) indicated how desirable they thought the behaviors are for men, how desirable they are for women, how typical they are for men, or how typical they are for women. Differences in desirability and typicality for men and women were established by calculating the effect size of the male-female difference (Cohen's *d*). As norm- violating stimuli (proscriptive stereotypes), we selected those behaviors that had large differences in perceived typicality and desirability. As expectancy-violating stimuli (descriptive stereotypes), we selected those behaviors that had large differences in perceived typicality, but small differences in desirability. There were few behaviors that were not norm-violating in an absolute sense (because typicality and desirability were highly correlated in our sample of behaviors), so we selected behaviors that were extremely norm-violating (with very large effect sizes) as proscriptions, and behaviors that were only slightly norm-violating (with small to medium effect sizes) as descriptions.

Appendix 1 contains the trait-implying stimuli used in the present study, as well as differences in desirability and typicality (Cohen's *d*) for men versus women. On average, the male norm violations and expectancy violations were considered more typical of women than men (Mean Cohen's *d*s -1.27 and -1.14, respectively), whereas the female norm violations and expectancy violations were considered much more typical of men than women (Mean Cohen's *d*s 0.89 and 1.20, respectively). Thus, the behaviors that were used in Study 2 were all clear stereotypes: the behaviors were considered typical for one gender, but not for the other.

To distinguish between descriptive stereotypes and proscriptive stereotypes, we calculated differences in the perceived desirability of these behaviors for men and women. For male descriptions, the effect size for differences between men and women was medium (Cohen's d = -0.43), for female descriptions, it was small (Cohen's d = 0.24). For male and female proscriptions, the effect size for differences between men and women was

large (Cohen's *d*s 1.99 and -2.19, respectively). Thus, the stimuli that were selected as proscriptions were extremely norm-violating, whereas the stimuli that were selected as descriptions were only slightly norm-violating.

Next to the trait-implying trials, the PRT contained control trials and filler trials. The control trials were matched to the trait-implying trials in the same way as in Study 1. The filler trials were again carefully crafted to make it difficult for participants to respond strategically. Compared to Study 1, there were relatively fewer filler trials, so that the correct response to two thirds of the trials was "no" and the correct response to one third of the trials was "yes".

Individual difference measures. Two individual difference measures were included in the present study, namely the Gender System Justification Beliefs Questionnaire (GSJB-scale; Jost, & Kay, 2005) and the Ambivalent Sexism Inventory (ASI; Glick, & Fiske, 1996; Dutch translation Glick et al., 2000). The GSJB is an 7-item scale (a = .59) that measures individual differences in people's motivation to protect the gender status quo. The GSJB contains items like "Society is set up so that men and women usually get what they deserve" and "The division of labor in families generally operates as it should", measured on a 6-point Likert scale. People with higher scores on the GSJB-scale tend to believe that the gender status quo is just and should be protected (Jost, & Kay, 2005). Because high GSJB-ers are more likely to form extreme inferences of norm-violating behavior (Rudman et al., 2012a), we expect that they will form stronger STIs of norm-violating behavior.

The ASI is a 22-item scale (a = .88) which consists of two subscales. The hostile sexism subscale measures antipathy against powerful women and consists of items like "women are too easily offended" and "women seek to gain power by getting control over men" (a = .86). The benevolent sexism subscale measures endorsement of chivalrous attitudes towards traditional women and consists of items like "a good woman should be put on a pedestal by her man" and "women should be cherished and protected by men" (a = .85), measured on a 6-point Likert scale. Both subscales measure complementary sexist ideologies and are therefore highly correlated (Glick, & Fiske, 2001).

Results

As in Study 1, we removed all incorrect trials (4.8 % of trials) and latencies faster than 200 ms or slower than 2000 ms (0.92 % of trials). We performed inverse transformation ($1/x$) on the data to remove skew (except for the difference scores, which were not skewed), but report untransformed means for the ease of interpretation.

We performed a repeated measures ANOVA with trial type (control vs. trait-implying), stereotype-consistency (stereotype-consistent versus stereotype-inconsistent) and kind of stereotype (proscriptive versus descriptive) as within subject-variables. Before turning to our main analyses, we first tested if participants formed STIs by testing for a main effect of trial type. If participants formed STIs based on our stimuli, they should be slower to indicate that a trait was not part of the sentence if it was preceded by a sentence that implied the trait (trait-implying trial), compared to when it was preceded by a sentence that did not imply the trait (control trial). In line with this expectation, there was a significant main effect of trial type, $F(1,87) = 19.72$, $p < .001$, $\eta_p^2 = .19$. Average latencies were larger for trait-implying trials (725 ms, $SD = 136$) than control trials (695 ms, $SD = 129$), suggesting that participants formed Spontaneous Trait Inferences.

After having established that the PRT produced a basic STI-effect, we turned to our main hypothesis. We expected that stereotype-congruency would differentially affect STI-formation for descriptive and proscriptive stereotypes (i.e., an interaction between trail type, stereotype-congruency and kind of stereotype). Specifically, we expected that STI-formation would be inhibited for descriptive stereotypes, replicating the inhibition of STIs found in previous studies (Wigboldus et al., 2003; 2004). For proscriptive stereotypes, we expected the opposite effect (providing support for the amplified STI-effect), or no effect at all (replicating Study 1).

There was a marginally significant three-way interaction between trial type, stereotype congruency, and kind of stereotype, $F(1, 82) = 3.18$, $p = .078$. $\eta_p^2 = 0.04$. For ease of interpretation, we calculated difference scores between trait-implying and control trials for each cell of the design, so that larger difference scores indicate larger STI-effects. Table 3 contains these difference scores (trait-implying minus control). As depicted in Table

3, there is an effect of stereotype-consistency on STI-formation for proscriptive, but not for descriptive stereotypes.

Table 3. *Difference scores between trait-implying trials and control trials (in ms) for proscriptive and descriptive stereotype-incongruent and stereotype-congruent trials. Larger difference scores reflect stronger Spontaneous Trait Inferences.*

	stereotype-incongruent	stereotype-congruent
proscriptions	23[a] (*SD* = 83)	49[b] (*SD* = 82)
descriptions	19[a] (*SD* = 68)	27[a] (*SD* = 77)

Note. Means with different subscript differ from each other at $p < .08$.

Contrary to our predictions, and contrary to the findings of Study 1, we find that participants form stronger STIs of stereotype-consistent than stereotype-inconsistent behaviors if the behaviors are norm-violating. If the stereotype is proscriptive, the STI-effect is larger for stereotype-consistent than stereotype-inconsistent trials (the mean difference in latency between trait-implying and control trials is 49 and 23 ms, respectively; $F(1,82) = 7.17$, $p = .009$, $\eta_p^2 = 0.08$). This is not the case if the behavior is merely descriptive (the mean difference in latency between trait-implying and control trials is 27 and 19 ms for consistent and inconsistent trials, respectively; $F < 1$). Contrary to our expectations, STIs seem to be inhibited, not amplified, when behavior is norm-violating. When behavior is merely expectancy-violating, we did not find evidence suggesting that STIs are inhibited or amplified. No other main effects or interactions were significant (all Fs < 1) [5].

[5] As in Study 1, we tested for effects of target sex, but did not expect any effects of target sex. We conducted a 2 (target sex: male versus female) x 2 (stereotype: proscriptive/descriptive stereotype violation for males versus females) x 2 (kind of stereotype: descriptive versus proscriptive) x 2 (trial type: control versus TI) within-subjects analysis. Contrary to our expectations, the 4-way interaction was marginally significant, $F(1,80) = 3.81$, $p = .054$. For female (but not male) targets, there was a marginally significant effect of stereotype-congruency for proscriptive stereotypes (but not for descriptive stereotypes), $F(1,80) = 3.62$, $p = .060$. In our view, this effect should be interpreted carefully, because the male and female norm violations differ in terms of how typical and desirable they are for men and women (as depicted in Table 1).

To explore if individual differences in sexism or system justification beliefs had an effect on STI-formation, we calculated a relative index for the difference in the strength of the STI-effect for proscriptive and descriptive stimuli in the following way. First, we calculated the basic STI-effect per sentence by subtracting latencies for control sentence from the trait-implying sentences. Next, we calculated a difference score by subtracting this STI-effect for stereotype-consistent sentences from the STI-effect for stereotype-inconsistent sentences. Thus, higher difference scores indicate stronger STIs for stereotype-inconsistent relative to stereotype-consistent sentences (i.e., a stronger amplified STI-effect). We conducted separate regression analyses with these difference scores as dependent variable and GSJB/ASI as predictors. System justification beliefs did not have an effect on the difference score for descriptive stereotypes, $\beta = 0.15$, $p = .18$ or proscriptive stereotypes, $\beta = 0.14$, $p = .20$. The Ambivalent Sexism Inventory also did not affect the difference score for descriptive stereotypes, $\beta = 0.15$, $p = .17$, or proscriptive stereotypes, $\beta = 0.02$, $p = .86$. We also conducted regression analysis for the difference scores of consistent and inconsistent descriptive and proscriptive STI-effects (in which control sentences were subtracted from the trait-implying sentences), separately. None of these effects was significant (all βs -.12 to .20, all ps .06 to .99). In sum, there were no effects of individual differences in system justifying motives or sexism on the strength of Spontaneous Trait Inferences. We had expected that there would be an amplified STI-effect for proscriptive stereotypes (i.e., that STIs would be stronger for stereotype-inconsistent than stereotype-consistent trials), and that this effect would be stronger for system justifiers. We did not find evidence for such an amplified STI-effect, or for moderation by system justification motives.

Discussion

Study 2, like Study 1, failed to provide evidence for the hypothesized amplified STI-effect. Whereas Study 1 did not provide any evidence for the effect of stereotype-congruency on STI-formation, Study 2 suggests that stereotype-inconsistent STIs may be inhibited for norm-violating (but not for expectancy-violating) behaviors. This runs contrary to our hypothesis, which was that stereotype-inconsistent STIs would be inhibited for

expectancy-violating (but not for norm-violating) behaviors. However, because the three way interaction between trial type, stereotype-consistency, and kind of stereotype was only marginally significant, these results should be interpreted carefully.

Studies 1 and 2 did not provide evidence for an amplified STI-effect, and only partially replicated earlier STI-research (Wigboldus et al., 2003; 2004). One important difference between the present research and the research by Wigboldus and colleagues (2003; 2004) is that the present research used names (e.g., "Susan", "Jonathan") to denote category-membership, whereas Wigboldus and colleagues used category labels (e.g., "the garbage man", "the professor"). Category labels may lead people to form more category-based inferences, while names may lead to more individualized inferences (Fiske, Neuberg, Beattie, & Milberg, 1987). Because we used names, gender stereotypes may not have had the expected effect on STIs because participants in our study may not have formed strong category-based inferences of men and women. To ensure that perceivers would form strong category-based inferences in Study 3, we used category labels (e.g., "man", "woman") to denote the gender of the target.

Another way in which we tried to improve upon our design is by using a mixed design instead of a within subjects-design in Study 3. We tried to maximize power in Study 2 by using many trials, but participant's level of concentration may have suffered as a result of this, and participants complained about the large number of trials in the study. To alleviate this concern, we used a much shorter and simpler task with fewer trials, in which participants were only exposed to stereotype-congruent or stereotype-incongruent proscriptive stereotypes. This task ran on MTurk with a large number of participants to maximize power.

Study 3

Method

Participants. Two hundred and twenty-five Americans participated through Amazon's Mechanical Turk (MTurk)-website in exchange for $1. Four people were removed from the dataset because more

than 25% of their responses were incorrect. The remaining sample consisted of 221 participants (111 men), aged 18 to 75 (mean age 35)[6].

Overview and design. Like Study 1, the PRT used in Study 3 contained two types of trait-implying trials, namely stereotype-consistent and stereotype-inconsistent proscriptions. No descriptive stereotypes were used so we could keep the task short and simple, and therefore, suitable for an online study. The present study had a 2 (trial type: control versus trait-implying) x 2 (stereotype-consistency: stereotype-consistent versus stereotype-inconsistent) mixed design with trial type as within-subjects variable and stereotype-consistency as between-subjects variable.

Procedure. Unlike Studies 1 and 2, Study 3 was run online using MTurk. After providing informed consent, participants were instructed to ensure they were in a quiet place where they would not be disturbed, and were asked not to take any breaks during the experiment. The remainder of the instructions was identical to the instructions of Study 1 and 2. The PRT consisted of eight practice trials, followed by 32 experimental trials. Sentences were presented on screen for 3500 ms (with 500 ms between trials). After completing the PRT, participants completed the GSJB-questionnaire, ASI, and several demographic questions.

Stimuli.

Probe Recognition Task. The PRT used in the present study consisted of 33 trials of the following three types: trait-implying trials (6), control trials (6) and filler trials (21). Trial order was fully randomized. The 6 trait-implying trials consisted of 3 trials that were norm-violating for males and 3 that were norm-violating for females. The behaviors were coupled to the category labels "the man" and "the woman". For participants in the stereotype-consistent condition, all dominant behaviors were coupled to the label "the man" and all weak behaviors were coupled to the label "the woman". For participants in the stereotype-inconsistent condition, all dominant behaviors were coupled to the label "the woman" and all weak behaviors were coupled to the label "the man". Similar to Study 1, each behavior was used only once. The control trials were again matched to the trait-implying trials, and filler trials were again carefully

[6] Exclusion criteria were established a priori, before any analyses were conducted; including these participants in the analyses does not alter the pattern of results.

crafted to make sure participants could not respond strategically. There were sufficient fillers to ensure a 50/50-balance in yes/no-responses across the task.

Because the sample of participants used in the present study was American, not Dutch, we pretested a new set of behaviors in the same way as in Study 2. Thirty-one MTurkers wrote down which traits came to mind when reading the behavioral sentences, and only sentences for which at least 70% of participants spontaneously wrote down the intended trait (or a close synonym) were selected. Next, four groups of participants (*N*s 17 to 24) indicated how desirable or typical these behaviors are for men or women, and we selected behaviors that differed strongly in terms of their perceived typicality and desirability for males and females. Appendix 1 contains the trait-implying sentences that were used in Study 3, as well as the differences in desirability and typicality between men and women for each of the selected stimuli (Cohen's *d*). As depicted in Appendix 1, stimuli that were selected as male norm violations were considered as less typical for men than women (Cohen's d = -1.25), and were considered as less desirable for men than women (Cohen's d = -1.60). Female norm violations were considered as less typical for women than men (Cohen's d = 1.77), and were considered less desirable for women than men (Cohen's d = 0.84). Thus, the stimuli selected for Study 3 were strongly proscriptive in nature: they violated expectancies about how men or women are expected to behave, as well as norms about how men or women should behave.

Like in Study 2, the GSJB-scale (a = .80) and ASI (a = .92) were used as individual difference measures in Study 3. Because the experiment was run in English, the original versions of the questionnaires were used.

Results

Latencies for incorrect responses (6.8%), as well as latencies smaller than 200 ms and larger than 2000 ms (0.6 %) were removed from the dataset. We performed inverse transformation ($1/x$) to deal with the skewed nature of the reaction time data, but report untransformed means for ease of interpretation.

Before turning to our main analyses, we first tested if participants formed STIs. In line with our expectations, there was a significant main effect of trial type, $F(1,220) = 227.48$, $p < .001$, $\eta_p^2 = .51$. Average latencies

were larger for trait-implying trials (788 ms, SD = 245) than control trials (682 ms, SD = 181), suggesting that participants formed STIs.

After having established that the PRT produced a basic STI-effect, we turned to our main hypothesis. We expected that there would be an effect of stereotype-consistency on STI-formation, which would be reflected in a significant interaction between trial type (trait-implying versus control) and stereotype-consistency. This interaction was not significant, F < 1.[7] Participants formed STIs of trait-implying behaviors, regardless of whether they were stereotype-consistent or stereotype-inconsistent.

To establish if individual differences in system justifying motives (scores on the Gender System Justification Beliefs scale) and sexism (scores on the Ambivalent Sexism Inventory) influenced the strength of the STI-effect, we conducted separate regression analyses with the difference between trait-implying and control trials as dependent variable. This difference score was calculated by subtracting the latency of control trials from the latency of trait-implying trials and reflects the strength of the STI-effect. There was no effect of GSJB on the difference score in the stereotypical condition, β = .08, p = .56, or in the counterstereotypical condition, β = .08, p = .52. There was also no effect of ASI on the difference score in the stereotypical condition, β = -0.03, p = .84, or in the counterstereotypical condition, β = .01, p = .95. In sum, individual differences in system justifying motives or sexism did not predict the strength of STIs. We expected that participants would form relatively stronger STIs of stereotype-inconsistent than stereotype-consistent sentences (an amplified STI-effect) but did not find evidence for such an effect. We also expected that this effect would be stronger for system justifiers, but did not find evidence for an effect of system justification motives on the strength of the STI-effect.

Discussion

[7] As in studies 1 and 2, we tested if there was an effect of target sex by conducting a 2 (target gender: male versus female) x 2 (trial type: control versus TI) within subjects analysis with stereotype congruency (stereotype-congruent versus stereotype-incongruent) as between-subjects factor. The three-way-interaction was not significant (F < 1), suggesting that there was no effect of target sex on stereotype-congruency effects for STIs.

Like Study 1 and 2, Study 3 did not provide evidence for an amplified STI-effect. Although the present study had a different design and used category labels instead of names (to make sure participants categorized the targets on the basis of their gender), we did not find evidence for an effect of stereotype-consistency on STI-formation. In so doing, Study 3 did not provide evidence for an amplified STI-effect, but also did not replicate the inhibition-effect for proscriptive stereotype violations that we found in Study 2.

General Discussion

Dominant women and weak men face adverse social and economic consequences, but can backlash occur at a spontaneous level? The goal of the present chapter was to investigate if gender stereotypes influence STI-formation. We expected that spontaneous trait inferences would be amplified for proscriptive stereotype violations (compared to stereotypical behaviors). This hypothesis was tested in three studies, using different stimuli, in both Dutch student samples and an American sample. These studies did not provide evidence for the effect of gender stereotypes on STIs. There was some evidence for an inhibition-effect for proscriptive stereotypes in Study 2 (although the omnibus interaction between kind of stereotype, consistency and type of trial was not significant), but this effect was not present in Studies 1 and 3. We also did not find evidence for the expected effect of system justifying beliefs on the amplified STI-effect. However, given that we did not find any evidence for an amplified STI-effect, it may be unsurprising that the effect was not moderated by system justifying beliefs.

The results reported in the present chapter are inconclusive, and there are many reasons why they may not have provided evidence for the hypothesized amplified STI-effect. As our studies largely produced null results, we can only speculate about these reasons. One explanation is that effects such as the dominance penalty occur only in intentional trait inferences, not in spontaneous trait inferences. Perhaps, backlash is a deliberate form of penalization of gender deviants, and people engage in backlash only if there is justification for it (as suggested by Rudman, & Fairchild, 2004). We expected that attributional processes would affect the formation of spontaneous trait inferences, but it is possible that these

processes only occur for intentional, explicit inferences. In Chapter 3 of this dissertation, we found evidence for an amplification of inferences using other tasks that measure spontaneous inferences, namely a Reverse Correlation Image Classification Task and Draw-a-Face Task. Arguably, however, the Probe Recognition Task used in the present chapter may measure STIs at a more indirect level because participants have to read sentences and respond to trait probes under time pressure. The task may therefore be less sensitive to the influence of certain attributional processes. We assumed that STIs are not necessarily implicit or automatic, but the STIs as measured with a Probe Recognition Task may have been more automatic than we assumed, and therefore may not have been influenced by attributional processes.

Alternatively, our failure to find evidence for stereotype-congruency effects may have been due to the methods we used, and the fact that we did not find evidence for the inhibition of stereotype-incongruent STIs for descriptive stereotypes (while other authors clearly did, e.g., Wigboldus et al., 2003; 2004) seems to be in line with a more methodological explanation for our null results. We will discuss several of these methodological explanations in turn. This is by no means and exhaustive list of possible methodological problems, as the number of post hoc explanations as to why our study produces null results is vast, giving us plenty of opportunities to CARK (Critique After the Results are Known; Nosek, & Lakens, in press). We nevertheless think it is useful to list some methodological concerns that could possibly benefit future research.

A first methodological concern is that our stimuli may have been so clearly trait-implying that they overruled any effects of stereotype-congruency. Put differently, if behaviors are extremely trait-implying, they may lead to trait inferences regardless of whom performed the behavior. For example, kicking a small puppy dog will likely lead to a strong negative STI, independent of the characteristics of the actor. In line with this explanation, we found clear STI-effects in all of our studies. Perhaps, stereotyping effects occur only if researchers use behaviors that are moderately trait-implying, not too extreme, and slightly ambiguous, so that perceivers can form situational inferences as well as trait inferences based on these behaviors. Indeed many of the behaviors that have been used in previous research (e.g., Wigboldus et al., 2003) were open to situational

explanations (e.g., "the professor could not answer the question" can imply something about the professor or about the question). In retrospect, we should probably not have selected behaviors that were so strongly trait-implying and that, therefore, left little room for alternative inferences. Future research could investigate if stereotypes may affect STIs most strongly for ambiguous behaviors that are not too strongly trait-implying.

A second methodological explanation is that the Probe Recognition Task may not pick up differences in inferential extremity. It is unclear if latencies in the PRT reflect changes in associative strength and, if so, if changes in inferential extremity are reflected in associative strength. Put differently, it is unclear if the PRT measures how strongly a trait (e.g., weakness) is associated with a person, and if the strength of this association is influenced by the extremity of the inference (e.g., how weak a person is considered to be). Although some research suggests that inferential extremity and STI-strength are related (Crawford, Skowronski, Stiff & Scherer, 2007), we cannot be certain that the PRT measures differences in inferential extremity. Perhaps, perceivers spontaneously form more extreme inferences of proscriptive stereotype violations (e.g., they infer that a man who complains when he breaks a nail is extremely weak), but this may not affect the associative strength between the target and the trait (e.g., the extent to which the man is associated with weakness). It may be useful for future research to find out how inferential extremity and associative strength are related, and how they are reflected in the PRT and other reaction time measures.

A third reason why the present studies may have produced null results is because they controlled for the effects of semantic priming. By matching every critical trial to its own control trial, we could separate STI-effects from effects of stereotypes that occurred in the absence of trait-implying behavior. Indeed, in Study 1, there was a main effect of stereotype-congruency that suggested that participants were quicker to indicate that stereotype-incongruent traits had not been part of the sentence they just read, regardless of whether this sentence had been trait-implying or not. This effect suggests that, when participants read a male name (e.g., "Peter"), stereotype-incongruent traits were inhibited (e.g., "weak"), so they were quicker to indicate that "weak" had not been part of the sentence hey just read. This effect occurred regardless of whether the sentence was trait-

implying (e.g., "could not open the jar of preserves") or not (e.g., "ate a sandwich"), suggesting that stereotypes affected the baseline level of activation of traits. Because the effect occurred in the absence of trait-implying behaviors, in our view, it merely reflects a semantic priming effect (cf. Dijksterhuis, & Van Knippenberg, 1995) unrelated to the inhibition of trait inferences. In the present study, we have put a high bar for stereotyping effects in STI-research by fully controlling for the effects of semantic priming, something that has not been done in all previous studies (e.g., Yan, Wang, & Zhang, 2012).

In sum, the present chapter set out to investigate effects of gender stereotypes on STI-formation, but produced inconclusive results. There are several reasons as to why we may have failed to find evidence for an amplified STI-effect. One possibility is that STIs are formed more automatically than we assumed, and that gender stereotypes influence deliberate, intentional processes, but not STIs. Chapter 3 and 4 of the present dissertation suggest that backlash can be measured with more or less subtle and spontaneous measures, but these measures allow participants more room for deliberation than the PRT. Given that we did not succeed in replicating previous research (Wigboldus et al., 2003; 2004), it is also possible that we failed to find results due to methodological issues. For now, it is difficult to conclude if spontaneous backlash does not exist, or if it exists but we have failed to find it.

Coda

The present chapter failed to provide an answer to our research question. Although our null results do not allow us to draw conclusions about spontaneous backlash, we hope that the questions raised in the present chapter may nevertheless be helpful to future STI-researchers. In closing, we would like to refer to Stanley Milgram, who invested many years on a study that, eventually, produced null results. When reflecting on this study in his book "The individual in a social world", he explains that he loves empirical science precisely because it does not always produce the expected results:

> *"Every experiment is a situation in which the end is unknown: it is tentative, indeterminate, something that may fail. An experiment may only produce a restatement of the obvious, or yield unexpected insights. The indeterminacy of its outcome is part of its excitement."* Milgram (1977).

We refer to this comment because, in a scientific culture in which finding support for a hypothesis has often been equated with success, and $F < 1$ with failure, we find solace in the realization that the indeterminacy of its outcome is exactly what made an experiment worth doing in the first place.

Appendix 1

Table 1. *Trait-implying sentences used in Study 1, and the traits implied by these sentences, in the original Dutch and with English translation (in italics).*

Male Proscriptions

sentence	trait
klaagde dat hij/zij zich niet zo lekker voelt net voor een tentamen. *complained that he/she did not feel well right before an exam.*	zeur *whiney*
gaf op terwijl hij/zij aan het trainen was voor een wedstrijd omdat hij/zij geen zin heeft om in de regen verder te rennen. *gave up while training for a race because he/she did not feel like continuing to run in the rain.*	opgever *spineless*
was bang dat iedereen zijn/haar nieuwe kapsel lelijk zou vinden. *worried that everyone would dislike his/her new haircut.*	onzeker *insecure*
moest huilen bij het zien van een gevoelige scene in een romantische comedy. *had to cry at a tender scene in a romantic comedy.*	emotioneel *emotional*
slaagde er niet in een jampotje open te krijgen. *was unable to open a jar of preserves.*	zwak *weak*
pinkte bij een bruiloft van vrienden meerdere malen een traantje weg. *had to wipe away a tear at a friend's wedding several times.*	sentimenteel *sentimental*
reageerde hysterisch toen hij/zij zag dat het boek dat hij/zij wilde kopen was uitverkocht. *responded hysterically when the book he/she wanted to buy was sold out.*	aansteller *melodramatic*
had lang nodig om te kiezen wat hij/zij in een restaurant wilde eten, en kwam tot twee keer toe op zijn/haar keuze terug. *needed a lot of time to decide what he/she would like to eat at a restaurant, and changed his/her mind about it twice.*	besluiteloos *indecisive*

Female Proscriptions

sentence	trait
nam de leiding zonder rekening te houden met wat anderen willen. *took the lead without taking into account what other people want.*	egoïstisch *self-centered*
eiste bij een overleg alle aandacht op terwijl de voorzitter probeerde zich aan de agenda te houden. *demanded everyone's full attention in a meeting while the chair was trying to stick to the agenda.*	egocentrisch *egocentric*
maakte een bijtende opmerking over hoe dik een vriend/vriendin is waar die vriend/vriendin bij stond. *made a biting remark to others about how fat a friend is in front of his/her friend.*	lomp *insensitive*
gaf de speler van het andere voetbalteam een kopstoot. *gave a player of the other soccer team a head-butt.*	agressief *aggressive*
nam bewust de rol van de voorzitter over door bij een vergadering direct zelf het woord nemen. *deliberately took over from the chair by speaking out first during a meeting.*	dominant *dominant*
maakte zijn/haar teamgenoten duidelijk dat hij/zij de beste speler van de ploeg is. *made clear to his/her teammates that he/she is the best player on the team.*	opschepper *show-off*
maakte zijn/haar studiegenoten duidelijk dat hij/zij de leiding over het groepje op zich wil nemen. *made clear to his/her fellow students that he/she wants to be in charge of the study group.*	bazig *bossy*
vertelde iedereen dat hij/zij het hoogste cijfer heeft gehaald in de cursus. *told everyone that he/she got the highest grade in class.*	arrogant *arrogant*

Table 2. *Trait-implying sentences used in Study 2 (translated from Dutch), the traits implied by these sentences, and effect sizes (Cohen's d) for the male-female differences in typicality (typ. d) and desirability (des. d).*

Male Proscriptions			
sentence	**trait**	**typ. *d***	**des. *d***
begon te gillen toen hij/zij een muis zag. *started to scream when he/she saw a mouse.*	aansteller *melodramatic*	-1.42	-2.86
durfde niet achteruit in te parkeren in de stad. *did not dare to park in reserve in the city.*	angstig *scared*	-1.50	-2.37
raakte in paniek toen er een spin in zijn/haar kamer zat. *panicked when there was a spider in his/her room.*	bang *scared*	-1.57	-2.54
barstte in tranen uit toen iets niet ging zoals hij/zij wilde. *burst into tears when something did not go the way he/she wanted.*	zwak *weak*	-1.37	-2.00
klaagde steen en been toen het een beetje begon te miezeren. *complained incessantly when it started to drizzle.*	zeur *whiney*	-0.87	-1.37
had buikpijn toen hij/zij een presentatie moest geven. *had a stomach ache when he/she had to give a presentation.*	nerveus *nervous*	-1.16	-2.76
begon te huilen toen iemand hem/haar kritiek gaf. *started to cry when someone criticized him/her.*	emotioneel *emotional*	-1.43	-2.33
raakt altijd de weg kwijt. *always gets lost.*	onhandig *inept*	-0.79	-2.01
botste bij het inparkeren tegen een paaltje aan. *crashed into a pole while parking.*	kluns *clumsy*	-0.97	-1.94
kon niet kiezen wat voor smaak ijsje hij/zij wilde. *could not decide what flavor of ice cream he/she wanted.*	besluiteloos *indecisive*	-1.62	-1.71
average		**-1.27**	**-2.19**

Female proscriptions			
sentence	**trait**	**typ. *d***	**des.*d***
gaf een kopstoot aan de speler van het andere voetbalteam. *gave the player of the other soccer team a head butt.*	agressief *aggressive*	0.74	1.92
schreeuwde tegen de scheidsrechter bij een sportwedstrijd. *yelled at the referee during a sports match.*	driftig *hot-headed*	1.20	2.18
sloeg met zijn/haar vuist op tafel. *hit the table with his/her fist.*	boos *angry*	0.85	1.40
zette keihard in bij de onderhandelingen, en gaf geen strobreed toe. *negotiated toughly, without giving in.*	dominant *dominant*	1.10	1.64
schepte op over de prestaties van zijn/haar sportteam. *boasted about the performance of his/her sports team.*	arrogant *arrogant*	0.76	1.75
vertelde een botte, beledigende mop. *cracked a rude, insulting joke.*	lomp *insensitive*	0.57	2.76
liet een wind in gezelschap. *farted in public.*	onbeschoft *rude*	0.89	1.96
liet in de kroeg een boer. *burped in public.*	onbeleefd *rude*	1.36	2.12
schepte op over het aantal sekspartners dat hij/zij heeft gehad. *boasted about the number of sex partners he/she has had.*	losbandig *promiscious*	0.78	2.24
maakte een lompe opmerking over het werk van een collega. *made an insensitive remark about a colleague's work.*	gemeen *mean*	0.68	1.89
average		**0.89**	**1.99**

Male Descriptions			
sentence	**trait**	**typ.*d***	**des.*d***
begon veel te laat met studeren voor een belangrijk tentamen. *started studying for an exam much too late.*	laks *lazy*	1.36	0.35
leverde een verslag in met spelfouten er in. *handed an essay in with typo's in it.*	slordig *sloppy*	0.82	0.20
werd aangehouden omdat hij/ze te hard had gereden. *was pulled over because he/she was driving too fast.*	roekeloos *reckless*	1.80	0.25
kreeg een snelheidsboete. *got a ticket for speeding.*	onvoorzichtig *incautious*	1.90	0.33
maakte een gat in zijn/haar kleren bij het strijken. *made a hole in his/her clothes while ironing.*	sukkel *nitwit*	1.57	0.32
ging bij het dansen op de tenen van zijn/haar danspartner staan. *stepped on his/her partner's toes when dancing.*	stuntel *clumsy*	0.92	-0.07
boekte een last-minute naar Vietnam en vertrok zonder reizigersvaccinaties. *booked a last-minute to Vietnam and left without getting inoculated.*	impulsief *impulsive*	0.80	0.22
hing rond terwijl hij/ze eigenlijk het huis zou moeten opruimen. *idled when he/she should be cleaning the house.*	lui *lazy*	0.64	0.17
liet zijn/haar spullen overal slingeren. *left his/her belongings laying around everywhere.*	rommelig *untidy*	1.20	0.41
sloeg het advies van zijn/haar collega's in de wind. *ignored his/her colleague's advice.*	eigenwijs *stubborn*	0.94	0.18
average		**1.20**	**0.24**

Female Descriptions			
sentence	**trait**	**typ. *d***	**des. *d***
deed heel aardig tegen iemand, maar kletste achter zijn rug om over hem. *acted very sweetly to someone, but talked behind his back.*	roddelaar *gossipy*	-1.22	-0.45
herschreef een stuk volledig omdat hij/ze het nog net niet helemaal goed genoeg vond. *completely rewrote a paper because he/she thought it was not perfect.*	perfectionistisch *perfectionistic*	-1.36	-0.37
twijfelde of hij/ze wel goed genoeg is om de opleiding af te maken. *doubted if he/she is smart enough to finish his/her degree.*	onzeker *insecure*	-1.21	-0.43
weet niet hoe de minister heet. *did not know the name of the minister.*	onwetend *ignorant*	-0.99	-0.23
vertelde een studiegenoot dat hij onvoldoende serieus is. *told a fellow student that he/she should be more serious.*	bemoeial *busybody*	-1.05	-0.45
kocht iets dat hij/ze niet nodig heeft, alleen omdat anderen zeiden dat hij/ze dat moest doen. *bought something, that he/she didn't need, just because other people told her that she should.*	beïnvloedbaar *gullible*	-0.85	-0.47
veroorzaakte per ongeluk kortsluiting. *accidentally caused the power to short-circuit.*	atechnisch *unpractical*	-1.15	-0.42
klaagt altijd over iemand, maar gedraagt zich poeslief tegen hem. *always complains about someone, but acts sickly sweet when the person is there.*	achterbaks *sneaky*	-1.33	-0.25
begreep niets van de sommen. *did not understand the arithmetic assignments at all.*	dom *stupid*	-0.97	-0.68
deed de verkeerde batterijen in de zaklamp, waardoor het lampje doorbrandde. *put the wrong batteries in the flashlight, causing it to break down.*	stom *stupid*	-1.25	-0.58
average		**-1.14**	**-0.43**

Table 3. *Trait-implying sentences used in Study 3, the traits implied by these sentences, and effect sizes (Cohen's d) for the male-female differences in typicality (typ. d) and desirability (des. d).*

sentence	implied trait	typ. *d*	des. *d*
Male Proscriptions			
The man/ woman panicked when there was a spider in the basement and send his wife/her husband in to kill it.	scared	-1.77	-2.29
The man/woman screamed at a mouse and jumped up against a cupboard, breaking the antique vase.	coward	-1.15	-1.50
The man/woman was unable to hold on to the heavy couch and dropped it on his wife's/her husband's foot.	weak	-0.83	-1.02
average		**-1.25**	**-1.60**
Female Proscriptions			
The man/woman did not cry when his/her mother died, and failed to comfort his/her sister	mean	1.30	0.99
The man/woman threw the remote at the TV so hard that it broke the screen.	violent	1.47	0.92
The man/woman burped loudly, interrupting a friend's heartfelt story.	disgusting	2.53	0.60
average		**1.77**	**0.84**

Note. Positive effect sizes indicate that the behavior is deemed more typical or desirable for men than women, negative effect sizes indicate that the behavior is deemed more typical or desirable for women than men. By convention, Cohen's *d*s of 0.20, 0.50, and 0.80 correspond to small, medium and large effect sizes (Cohen, 1988).

CHAPTER 6

General Discussion

Women who apply for a managerial position face a Catch-22: they have to behave agentically to be regarded as sufficiently competent for the job, but are disliked if they do. Backlash impedes women's ascent up the corporate ladder and makes it difficult for aspiring female leaders to shatter the glass ceiling. Men, too, can suffer negative consequences as a result of backlash. For example, men who are weak, shy, anxious, afraid or nervous may be socially sanctioned. As a result of backlash, men are required to refrain from showing their weaknesses, which restricts their opportunities to get social support when they need it most (e.g., when they feel anxious or scared). Backlash limits men's and women's behavioral options, such that women need to curb their agency to avoid being perceived as overly dominant, while men need to hide their fears and weaknesses to avoid been casted off as overly weak.

Backlash is a major roadblock for reaching gender parity, but what motivates people to penalize gender deviants? In the present dissertation, I aimed to learn more about the motives that underlie backlash. Throughout four empirical chapters, I studied backlash in light of the Status Incongruity Hypothesis (SIH; Rudman, Moss-Racusin, Phelan, & Nauts, 2012a). The SIH suggests that agentic women and communal men threaten the status quo, and that people engage in backlash as a way of alleviating this threat. In the present dissertation, I explored three propositions that follow from the SIH, namely that a) system justifying motives underlie backlash; b) these motives underlie backlash against *both* genders and c) backlash results from the violation of prescriptive stereotypes (not descriptive stereotypes). Although propositions a) and b) are highly related to each other, I will discuss separately the evidence this dissertation provides for the role of system justifying motives in predicting backlash against men, because the present research is the first to empirically test the role of system justifying motives in predicting backlash against men. In the current chapter, I will first discuss the theoretical contributions of the present dissertation before going on to discuss methodological implications, the limitations of the present line of research, and practical implications.

Theoretical Contributions

System Justifying Motives and Backlash

According to the Status Incongruity Hypothesis (SIH), system justifying motives underlie backlash against agentic women and communal men. Thus, people who are motivated to protect the gender status quo (either chronically or experimentally induced) should be more likely to penalize gender deviants, because gender deviants threaten the legitimacy of the status quo. I studied this proposition in four empirical chapters. In line with the Status Incongruity Hypothesis (Rudman et al., 2012a), system justifying motives affected liking of agentic female job applicants (Chapter 2), mental images of scared/nervous men (Chapter 3) and memory for gender deviant behaviors (Chapter 4). Unexpectedly, system justifying motives did not affect spontaneous inferences of gender deviant behaviors (Chapter 5).

The role of system justifying motives was studied using correlational as well as experimental procedures. In Chapters 2 and 4, individual differences in system justifying motives were related to backlash: these studies suggest that people who are chronically motivated to protect the status quo are more likely to engage in backlash. In other studies, system justifying motives were experimentally manipulated (Chapters 3 and 4). In these studies, participants read an alleged newspaper article about the decline of the economy (system threat-condition), an article about the rise of the economy (system affirmation-condition; Chapter 3) or an article about bird watching (control condition; Chapter 4): people in the system threat-condition were more likely to engage in backlash than people in other conditions. Taken together, these studies suggest that people are more likely to engage in backlash if they have been primed with a threat to an existing social structure (i.e., a decline in their country's economic prowess). Interestingly, in line with research suggesting that a threat to an existing social system can be diminished by bolstering an unrelated system (Heine, Proulx, & Vohs, 2006; Kay et al., 2009), the effect of a system threat prime occurred even though this prime was unrelated to the gender status quo.

In sum, in three out of the four empirical chapters of this dissertation, people who were motivated to protect the status quo were more likely to penalize gender deviants. This was the case regardless of

whether system justifying motives were measured as an individual difference variable or manipulated using a system threat prime: in both cases, system justifiers were more likely to engage in backlash. This is in line with the SIH, which predicts that backlash serves to defend male hegemony when societal structures are threatened.

Backlash against Men

In line with the Status Incongruity Hypothesis (SIH), the data in this dissertation suggest that system justifying motives exacerbate backlash against both genders. Previous researchers have mostly focused on studying backlash against women. To the best of my knowledge, the present research is the first to study the motives that underlie backlash against men. Throughout two chapters, system justifying motives predicted backlash against men. When people's motivation to protect the status quo was high, this altered their impression of men who are nervous or afraid (Chapter 2) and their memory for men who cry, blush, scream or complain (Chapter 3). These data suggest that the penalization of agentic or dominant women and communal or weak men stems from the same underlying motive. Agentic women jeopardize the status quo by engaging in high status behaviors that are reserved for leaders and men. Likewise, communal men jeopardize the perceived legitimacy of the status quo by engaging in low status behaviors that are reserved for women. The gender status quo is perceived as fair only to the extent that men are regarded as having superior leadership skills: men who challenge this notion jeopardize the gender hierarchy, and backlash serves to put them "back in their place". In sum: the SIH presents a functional, motivational account of backlash by suggesting that gender deviant men and women are penalized as a way of protecting the gender hierarchy. The present dissertation corroborates that view by suggesting that people engage in backlash against both genders, and that system justifying motives underlie backlash against women as well as men.

Proscriptive Stereotypes

Next to suggesting that system justifying motives underlie backlash against both genders, the Status Incongruity Hypothesis (SIH) suggests that gender atypical behavior should be penalized only to the extent that it poses a threat to the status quo. Proscriptive stereotypes are *gender rules* that are

strongly aligned with status (Rudman et al., 2012a, Study 1): as such, people who violate them pose a threat to the status quo and are likely to be sanctioned. We tested this proposition in two chapters. In Chapter 3, we tested if backlash effects were most pronounced if behavior constituted a violation of proscriptive stereotypes and people were motivated to protect the status quo. Indeed, our results suggested that gender deviant men were imagined as having the most negative facial features if their behavior constituted a proscriptive stereotype violation and participants had been threatened with the decline of the economy. In Chapter 5, we tested if people spontaneously formed more extreme inferences of proscriptive stereotype violations compared to stereotypical behaviors (but not of descriptive stereotype violations), but we did not find evidence for this effect. Taken together, these studies do not provide sufficient evidence to draw conclusions about this proposition of the SIH, since I only found evidence for this proposition in one study. More research is needed to disentangle the descriptive and proscriptive components of gender stereotypes, for example, by exposing people to stereotype violations that are equal in terms of how unexpected they are, but differ in desirability.

Backlash and the SIH

In sum, the results of the present dissertation are relevant with regard to three propositions of the Status Incongruity Hypothesis (SIH). First of all, this dissertation suggests that system justifying motives underlie backlash. Second, it suggests that people engage in backlash against both men and women, and that backlash against both genders stems from this same system justifying motive. Third, the SIH suggests that backlash stems from a violation of proscriptive (not descriptive) stereotypes, but I did not find conclusive evidence for this third proposition of the SIH. Taken together, these results suggest that it is unlikely that backlash merely is the result of perceptual contrast effects, as could be expected based on Role Congruity Theory (Eagly, & Karau, 2002). Instead, backlash serves to put gender deviants back in their place as a way of maintaining the status quo. Although more research is needed to study the role of proscriptive stereotypes in backlash, the results presented in this dissertation suggest that there is a motivational component to backlash.

Methodological Implications

Next to studying backlash in relation to the Status Incongruity Hypothesis (Rudman et al., 2012a), the goal of the present dissertation was to introduce new methodologies to study backlash. In every chapter of this dissertation, a new way of studying backlash was introduced. Table 1 contains an overview of these methods, as well as the most important positive and negative features of these tasks.

Table 1. *Novel backlash-paradigms used in the present dissertation and their main features.*

Chapter	Paradigm	Main task features
2	live job interviews	high ecological validity, but not optimally suited to test backlash against both genders.
3	reverse correlation (RCIC)/Draw-a-Face Task (DaFT)	data-driven measure of spontaneous inferences, but the results of the RCIC and DaFT do not necessarily align.
4	memory task	subtle, indirect measure of a possible precursor of backlash, but the relationship between selective memory and backlash needs to be studied further.
5	probe recognition task	indirect measure of spontaneous trait inferences, but the results of the task are highly inconclusive.

We introduced a paradigm in Chapter 2 in which live job interviews were used to study backlash against agentic women. This task extended the ecological validity of backlash research, but was not optimally suited to study backlash against men. In Chapters 3, 4 and 5, new paradigms were introduced that are more suitable to study backlash against men. These paradigms have several advantages, but they also raise new questions about what constitutes backlash (e.g., do subtle, indirect processes such as biased

memory constitute backlash?) and how it can be measured (e.g., are the RCIC and DaFT sensitive to the same facial information?). These measures have several advantages: for example, by using behavioral sentences, they allow researchers to study a wide range of behaviors in a very controlled setting. Moreover, the chapters presented in this dissertation employed paradigms that were increasingly more subtle, spontaneous, and indirect. Additionally, the mental imaging-paradigms presented in Chapter 3 allow researchers to explore people's impressions of gender deviants in a bottom-up, data driven fashion. Beyond the realm of backlash research, the DaFT may be a promising alternative to a RCIC for researchers who are interested in studying mental images, but are unable to use a RCIC for practical reasons. A RCIC is a long and tedious task for participants, and it may therefore be difficult to use the task in certain populations (e.g., with children or patient groups) and settings (e.g., in field research). Because the DaFT is a short, easy and fun task for participants, it may be a suitable alternative, although more research is needed to validate the task.

Although the present dissertation presents several new toolkits to study backlash, it also leaves many questions unanswered. For example, in Chapter 3, the results of the RCIC and DaFT do not align, and in Chapter 5, the results of the Probe Recognition Paradigm are inconclusive. As such, these studies should be regarded as a proof of concept, and more research is needed to study them further. Nevertheless, the present research suggests that it is possible to venture beyond classic backlash-paradigms and explore backlash against both genders using subtle, indirect, and data-driven methodologies borrowed from the social cognition-literature.

Limitations and Suggestions for Future Research

Although the present research provides new insight into the causes of backlash, it also leaves several questions unanswered. As discussed above, the present dissertation provides a proof of concept for different paradigms, but more research is needed to study why these paradigms sometimes provide inconclusive results. Moreover, future researchers could further investigate the extent to which the findings presented in this dissertation are replicable across different cultural and societal contexts. The research in this dissertation was conducted using Dutch student samples, American student samples, and American MTurk-samples.

Although we did not find evidence for a cultural difference in backlash, a cross-cultural investigation of backlash seems long overdue, since most backlash research has been conducted in the US.

Additionally, future researchers could more directly test if behaviors are status related. In the present research, I assumed that the behavioral stimuli that I selected were status incongruent because they reflected traits that are status incongruent (e.g., "hitting the table with your fist" is indicative of dominance, which is a high status trait), but it would be a good idea to directly test this assumption. Another limitation of the present research is that it only indirectly suggests that engaging in backlash helps people maintain the status quo. Throughout three chapters, people were more likely to engage in backlash if they were motivated to protect the status quo, but this does not provide conclusive evidence for the SIH's assertion that people engage in backlash *because* doing so helps them protect the status quo. Future studies could investigate if backlash effectively lowers people's system justifying motivation, suggesting that engaging in backlash is an effective, goal-directed process.

Finally, the present research was not optimally suited to test for differences in the strength of backlash towards men and women, because the proscriptions for men and women differed from each other in several ways (e.g., in terms of how typical they were). However, when conducting the studies for this dissertation, I subjectively felt that participants had no qualms about engaging in backlash against gender deviant men. People laughed, snorted or frowned upon hearing male (but not female) proscriptions. Interestingly, the results of the pretests we conducted to select stimuli suggest that proscriptions were much stronger for males than they were for females (Cohen's d = |0.51| and = |0.95| for males in American and Dutch samples, versus d = |0.17| and |0.58| for females). In other words, people seemed more willing to indicate that feminine behaviors were out of bounds for men than to indicate that masculine behaviors were out of bounds for women. This effect could be due to idiosyncrasies of the stimuli I selected, but it does raise the question if people may be less motivated to curb prejudice against communal men than they are to curb prejudice against agentic women. Future research could further investigate if, as a result of the women's movement, people

may curb their prejudice towards women, but not their prejudice towards men.

Practical Implications

The Status Incongruity Hypothesis suggests that backlash against men and women stems from the same underlying motive. Because both genders may be victims (and perpetrators) of backlash, casting off backlash as a "women's issue" ignores the negative consequences backlash has for men's well-being. As such, commonly used strategies aimed at shattering the glass ceiling (e.g., including women in selection committees for managerial positions) may not be optimally effective, since both women and men may penalize agentic female leaders. Moreover, presenting women (but not men) as powerless victims of backlash may do little to change extant associations between women and low status characteristics.

Next to suggesting that backlash against men and women is inextricably intertwined, the present dissertation indicates which people are most likely to engage in backlash, and why. Although it may seem plausible to expect that individual differences in sexism or (implicit) gender stereotypes predict backlash, this contention has not been corroborated in the literature (Rudman, & Glick, 2008; but see Rudman, & Glick, 2001). Thus, interventions aimed at diminishing sexism or gender stereotypes are unlikely to limit backlash, although they may have other beneficial effects. Instead, interventions aimed at fulfilling basic psychological motives such as people's needs for certainty may be more successful, because they diminish people's need to protect the system and, with that, their motivation to lash out at gender deviants. Put differently, reaffirming people's faith in existing social systems or reducing their anxiety and uncertainty may be a fruitful strategy to combat backlash against both genders. By shedding new light on the motives that underlie backlash and the processes that may lead up to it, the present research may help policy makers develop more evidence-based interventions to effectively target gender inequality.

References

Alsop, R., Fitzsimons, A., & Lennon, K. (2002). *Theorizing gender: An introduction.* Hoboken, NJ: Wiley-Blackwell Publishing.

Asch, S. E. (1946). Forming impressions of personality. *The Journal of Abnormal and Social Psychology, 41*, 258-290.

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J., Fiedler, K., ... & Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality, 27*, 108-119.

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*, 543-554.

Ballew, C. C., & Todorov, A. (2007). Predicting political elections from rapid and unreflective face judgments. *Proceedings of the National Academy of Sciences, 104*, 17948-17953.

Banaji, M. R., & Hardin, C. D. (1996). Automatic stereotyping. *Psychological Science, 7*, 136-141.

Bar, M., Neta, M., & Linz, H. (2006). Very first impressions. *Emotion, 6*, 269-278.

Bureau of Justice Statistics (2011). Homicide trends in the US: 1980-2008. Retrieved from: http://www.bjs.gov/content/pub/pdf/htus8008.pdf.

Burgess, D., & Borgida, E. (1999). Who women are, who women should be: Descriptive and prescriptive gender stereotyping in sex discrimination. *Psychology, Public Policy, and Law, 5*, 665-692.

Butler, D., & Geis, F. L. (1990). Nonverbal affect responses to male and female leaders: Implications for leadership evaluations. *Journal of Personality and Social Psychology, 58*, 48-59.

Carranza, E. (2004). Is what is good for the goose derogated in the gander? Reactions to masculine women and feminine men. (unpublished doctoral dissertation). Princeton University, Princeton, NJ.

Carré, J. M., & McCormick, C. M. (2008). In your face: facial metrics predict aggressive behaviour in the laboratory and in varsity and professional

hockey players. *Proceedings of the Royal Society B: Biological Sciences, 275*, 2651-2656.

Catalyst (2012). *Women CEOs of the Fortune 100.* New York: Catalyst.

Centre for Disease Control (2012). Suicide: Facts at a glance. Retrieved from: http://www.cdc.gov/violenceprevention/pdf/Suicide_DataSheet-a.pdf.

Cleary, A. (2012). Suicidal action, emotional expression, and the performance of masculinities. *Social Science, & Medicine, 74*, 498-505.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology, 65*, 145-153.

Cohen, J. (1988). *Statistical power for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*, 1304-1312.

Connell, R. W. (1995) *Masculinities* Cambridge: Polity.

Costrich, N., Feinstein, J., Kidder, L., Marecek, J., & Pascale, L. (1975). When stereotypes hurt: Three studies of penalties for sex-role reversals. *Journal of Experimental Social Psychology, 11*, 520-530.

Crawford, M. T., Skowronski, J. J., Stiff, C., & Scherer, C. R. (2007). Interfering with inferential, but not associative, processes underlying spontaneous trait inference. *Personality and Social Psychology Bulletin, 33*, 677-690.

Crocker, J., & McGraw, K. M. (1984). What's good for the goose is not good for the gander: Solo status as an obstacle to occupational achievement for males and females. *American Behavioral Scientist, 27*, 357-369.

Derlega, V. J., & Chaikin, A. L. (1976). Norms affecting self-disclosure in men and women. *Journal of Consulting and Clinical Psychology, 44*, 376-380.

Diekman, A. B., Goodfriend, W., & Goodwin, S. (2004). Dynamic stereotypes of power: Perceived change and stability in gender hierarchies. *Sex Roles, 50*, 201-215.

Dijksterhuis, A. P., & Van Knippenberg, A. D. (1995). Memory for stereotype-consistent and stereotype-inconsistent information as a function of processing pace. *European Journal of Social Psychology*, *25*, 689-693.

Dijksterhuis, A. P., & Van Knippenberg, A. D. (1996). The knife that cuts both ways: Facilitated and inhibited access to traits as a result of stereotype activation. *Journal of Experimental Social Psychology*, *32*, 271-288.

Dijksterhuis, A. P., Van Knippenberg, A. D., Kruglanski, A. W., & Schaper, C. (1996). Motivated social cognition: Need for closure effects on memory and judgment. *Journal of Experimental Social Psychology*, *32*, 254-270.

Dotsch, R., Wigboldus, D. H. J., Langner, O., & van Knippenberg, A. (2008). Ethnic out-group faces are biased in the prejudiced mind. *Psychological Science*, *19*, 978-980.

Dotsch, R., Wigboldus, D. H. J., & van Knippenberg, A. (2011). Biased allocation of faces to social categories. *Journal of Personality and Social Psychology*, *100*, 1-16.

Dotsch, R., & Todorov, A. (2012). Reverse correlating social face perception. *Social Psychological and Personality Science*, *3*, 562-571.

Dunham, Y., Srinivasan, M., Dotsch, R., & Barner, D. (2014). Religion insulates ingroup evaluations: the development of intergroup attitudes in India. *Developmental Science*, *17*, 311-319.

Eagly, A. H., Johannesen-Schmidt, M. C., & Van Engen, M. L. (2003). Transformational, transactional, and laissez-faire leadership styles: a meta-analysis comparing women and men. *Psychological Bulletin*, *129*, 569-591.

Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review*, *109*, 573-598.

Eagly, A. H., Makhijani, M. G., & Klonsky, B. G. (1992). Gender and the evaluation of leaders: A meta-analysis. *Psychological Bulletin*, *111*, 3-22.

Eagly, A. H., & Mladinic, A. (1994). Are people prejudiced against women? Some answers from research on attitudes, gender stereotypes, and judgments of competence. *European Review of Social Psychology, 5*, 1-35.

Erber, R., & Fiske, S. T. (1984). Outcome dependency and attention to inconsistent information. *Journal of Personality and Social Psychology, 47*, 709-726.

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.

Feinman, S. (1981). Why is cross-sex-role behavior more approved for girls than for boys? A status characteristic approach. *Sex Roles, 7*, 289-300.

Fiske, S. T., Bersoff, D. N., Borgida, E., Deaux, K., & Heilman, M. E. (1991). Social science research on trial: Use of sex stereotyping research in "Price Waterhouse v. Hopkins". *American Psychologist, 46*, 1049-1060.

Fiske, S.T., Neuberg, S.L., Beattie, A.E., & Milberg, S.J. (1987). Category-based and attribute-based reactions to others: Some informational conditions of stereotyping and individuating processes. *Journal of Experimental Social Psychology, 23*, 399-427.

Frosh, S., Phoenix, A., & Pattman, R. (2003). The trouble with boys. *The Psychologist, 16*, 84-87.

Fyock, J., & Stangor, C. (1994). The role of memory biases in stereotype maintenance. *British Journal of Social Psychology, 33*, 331-343.

Gill, M. J. (2004). When information does not deter stereotyping: Prescriptive stereotyping can foster bias under conditions that deter descriptive stereotyping. *Journal of Experimental Social Psychology, 40*, 619-632.

Glick, P., & Fiske, S. T. (1996). The Ambivalent Sexism Inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology, 70*, 491-512.

Glick, P., & Fiske, S. T. (2001). An ambivalent alliance: Hostile and benevolent sexism as complementary justifications for gender inequality. *American Psychologist, 56*, 109-118.

Glick, P., Fiske, S. T., Mladinic, A., Saiz, J. L., Abrams, D., Masser, B., ... & López, W. L. (2000). Beyond prejudice as simple antipathy: Hostile and benevolent sexism across cultures. *Journal of Personality and Social Psychology, 79*, 763-775.

Glick, P., Lameiras, M., Fiske, S. T., Eckes, T., Masser, B., Volpato, C., ... & Wells, R. (2004). Bad but bold: Ambivalent attitudes toward men predict gender inequality in 16 nations. *Journal of Personality and Social Psychology, 86*, 713-728.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology, 85*, 197-216.

Haldeman, D. C. (2000). Gender atypical youth: Clinical and social issues. *School Psychology Review, 29*, 192-200.

Hall, C. C., Goren, A., Chaiken, S., & Todorov, A. (2009). Shallow cues with deep effects: Trait judgments from faces and voting decisions. In E. Borgida, J.L. Sullivan, & C.M. Frederico (eds). *The political psychology of democratic citizenship* (pp 73-99). New York: Oxford University Press.

Heilman, M. E. (2001). Description and prescription: How gender stereotypes prevent women's ascent up the organizational ladder. *Journal of Social Issues, 57*, 657-674.

Heilman, M. E., Block, C. J., & Martell, R. F. (1995). Sex stereotypes: Do they influence perceptions of managers?. *Journal of Social Behavior & Personality, 6,* 237-252.

Heilman, M. E., & Okimoto, T. G. (2007). Why are women penalized for success at male tasks?: The implied communality deficit. *Journal of Applied Psychology, 92*, 81-92.

Heilman, M. E., & Wallen, A. S. (2010). Wimpy and undeserving of respect: Penalties for men's gender-inconsistent success. *Journal of Experimental Social Psychology, 46*, 664-667.

Heine, S. J., Proulx, T., & Vohs, K. D. (2006). The meaning maintenance model: On the coherence of social motivations. *Personality and Social Psychology Review, 10*, 88-110.

Jack, R. E., Garrod, O. G., Yu, H., Caldara, R., & Schyns, P. G. (2012). Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences, 109*, 7241-7244.

Jost, J. T., Banaji, M. R., & Nosek, B. A. (2004). A decade of system justification theory: Accumulated evidence of conscious and unconscious bolstering of the status quo. *Political Psychology, 25*, 881–919.

Jost, J. T, & Hunyady, O. (2003). The psychology of system justification and the palliative function of ideology. *European Review of Social Psychology, 13*, 111-153.

Jost, J. T., & Kay, A. C. (2005). Exposure to benevolent sexism and complementary gender stereotypes: Consequences for specific and diffuse forms of system justification. *Journal of Personality and Social Psychology, 88*, 498-509.

Jost, J. T., Pelham, B. W., Sheldon, O., & Ni Sullivan, B. (2003). Social inequality and the reduction of ideological dissonance on behalf of the system: Evidence of enhanced system justification among the disadvantaged. *European Journal of Social Psychology, 33*, 13-36.

Kanter, R. M. (1977). Some effects of proportions on group life: Skewed sex ratios and responses to token women. *American Journal of Sociology, 82*, 965-990.

Kay, A. C., Gaucher, D., Peach, J. M., Laurin, K., Friesen, J., & Zanna, M. (2009). Inequality, discrimination, and the power of the status quo: Direct evidence for a motivation to see things as they should be. *Journal of Personality and Social Psychology, 97*, 421-434.

Kay, A. C., & Jost, J. T. (2003). Complementary justice: effects of "poor but happy" and "poor but honest" stereotype exemplars on system justification and implicit activation of the justice motive. *Journal of Personality and Social Psychology, 85*, 823-837.

Kelley, H. H., & Michela, J. L. (1980). Attribution theory and research. *Annual Review of Psychology, 31*, 457-501.

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H., Hawk, S. T., & van Knippenberg, A. (2010). Presentation and validation of the Radboud Faces Database. *Cognition and Emotion, 24*, 1377-1388.

Lundqvist, D., & Litton, J. E. (1998). The averaged Karolinska Directed Emotional Faces-AKDEF [CD ROM]. Stockholm: Karolinska Institutet.

MacKinnon, D. P., Fritz, M. S., Williams, J., & Lockwood, C. M. (2007). Distribution of the product confidence limits for the indirect effect: Program PRODCLIN. *Behavior Research Methods, 39*, 384-389.

Macrae, C. N., Hewstone, M., & Griffiths, R. J. (1993). Processing load and memory for stereotype-based information. *European Journal of Social Psychology, 23*, 77-87.

Mangini, M. C., & Biederman, I. (2004). Making the ineffable explicit: Estimating the information employed for face classifications. *Cognitive Science, 28*, 209-226.

Martin, C. L. (1990). Attitudes and expectations about children with nontraditional and traditional gender roles. *Sex Roles, 22*, 151-166.

Martin, C. L., & Halverson Jr, C. F. (1983). The effects of sex-typing schemas on young children's memory. *Child Development, 54,* 563-574.

Mazur, A., Mazur, J., & Keating, C. (1984). Military rank attainment of a West Point class: Effects of cadets' physical features. *American Journal of Sociology, 90*, 125-150.

McKoon, G., & Ratcliff, R. (1986). Inferences about predictable events. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12,* 82-91.

Milgram, S. (1977). *The individual in a social world: Essays and experiments.* Menlo Park, California: Addison-Wesly Publishing Company.

Moss-Racusin, C. A., Phelan, J. E., & Rudman, L. A. (2010). When men break the gender rules: Status incongruity and backlash toward modest men. *Psychology of Men and Masculinity, 11*, 140-151.

Moss-Racusin, C. A., & Rudman, L. A. (2010). Disruptions in women's self-promotion: The backlash avoidance model. *Psychology of Women Quarterly, 34*, 186-202.

Nauts, S., Langner, O., Huijsmans, I., Vonk, R., & Wigboldus, D. H. J. (2014). Forming impressions of personality: A replication and review of Asch's

(1946) evidence for a primacy-of-warmth effect in impression formation. *Social Psychology, 45*, 153-163.

Nauts, S., & Vonk, R. (2009). *Het backlash effect in Nederland: Waarom pittige vrouwen in Nederland niet aan de top komen.* ASPO, Jaarboek Sociale Psychologie. Groningen, The Netherlands: ASPO-pers.

Nauts, S., Rudman, L.A., Langner, O., & Wigboldus, D.H.J. (Unpublished raw data). *Proscriptions for men and women.*

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*, 231-259.

Nosek, B.A., & Lakens, D. (in press). Registered reports: A method to increase the credibility of published results. *Social Psychology.*

Olivola, C. Y., & Todorov, A. (2010). Fooled by first impressions? Reexamining the diagnostic value of appearance-based inferences. *Journal of Experimental Social Psychology, 46*, 315-324.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology, 45*, 867-872.

Phelan, J. E., Moss-Racusin, C. A., & Rudman, L. A. (2008). Competent yet out in the cold: Shifting criteria for hiring reflect backlash toward agentic women. *Psychology of Women Quarterly, 32*, 406-413.

Phoenix, A., Frosh, S., & Pattman, R. (2003). Producing contradictory masculine subject positions: Narratives of threat, homophobia and bullying in 11–14 year old boys. *Journal of Social Issues, 59*, 179-195.

Prentice, D. A., & Carranza, E. (2002). What women and men should be, shouldn't be, are allowed to be, and don't have to be: The contents of prescriptive gender stereotypes. *Psychology of Women Quarterly, 26*, 269–281.

Prentice, D. A., & Carranza, E. (2004). Sustaining cultural beliefs in the face of their violation: The case of gender stereotypes. In M. Schaller, & C.S. Crandall (Eds.), *The psychological foundations of culture* (259-280). Mahwah, NJ: Lawrence Erlbaum Associates.

Proulx, T., Heine, S. J., & Vohs, K. D. (2010). When is the unfamiliar the uncanny? Meaning affirmation after exposure to absurdist literature, humor, and art. *Personality and Social Psychology Bulletin, 10*, 88-111.

Ramos, T., Garcia-Marques, L., Hamilton, D. L., Ferreira, M., & Van Acker, K. (2012). What I infer depends on who you are: The influence of stereotypes on trait and situational spontaneous inferences. *Journal of Experimental Social Psychology, 48*, 1247-1256.

Ridgeway, C. L. (2001). Gender, status, and leadership. *Journal of Social Issues, 57,* 627-655.

Rothbart, M., Evans, M., & Fulero, S. (1979). Recall for confirming events: Memory processes and the maintenance of social stereotypes. *Journal of Experimental Social Psychology, 15*, 343-355.

Rubin, J. Z., Provenzano, F. J., & Luria, Z. (1974). The eye of the beholder: Parents' views on sex of newborns. *American Journal of Orthopsychiatry, 44*, 512-519.

Rudman, L. A. (1998). Self-promotion as a risk factor for women: The costs and benefits of counterstereotypical impression management. *Journal of Personality and Social Psychology, 74*, 629-645.

Rudman, L.A., & Fairchild, K. (2004). Reactions to counterstereotypic behavior: The role of backlash in cultural stereotype maintenance. *Journal of Personality and Social Psychology, 87*, 157–176.

Rudman, L. A., & Glick, P. (1999). Feminized management and backlash toward agentic women: the hidden costs to women of a kinder, gentler image of middle managers. *Journal of Personality and Social Psychology, 77*, 1004-1010.

Rudman, L. A., & Glick, P. (2001). Prescriptive gender stereotypes and backlash toward agentic women. *Journal of Social Issues, 57*, 732–762.

Rudman, L. A., & Glick, P. (2008). *The social psychology of gender: How power and intimacy shape gender relations.* New York: Guilford Press.

Rudman, L. A., & Goodwin, S. A. (2004). Gender differences in automatic in-group bias: Why do women like women more than men like men? *Journal of Personality and Social Psychology, 87*, 494-509.

Rudman, L. A., & Kilianski, S. E. (2000). Implicit and explicit attitudes toward female authority. *Personality and Social Psychology Bulletin, 26*, 1315-1328.

Rudman, L. A., & Mescher, K. (2013). Penalizing men who request a family leave: Is flexibility stigma a femininity stigma?. *Journal of Social Issues, 69*, 322-340.

Rudman, L. A., Moss-Racusin, C. A., Glick, P., & Phelan, J. E. (2012b). Reactions to vanguards: Advances in backlash theory. In P. G. Devine, & E. A. Plant (Eds.), *Advances in Experimental Social Psychology, 45,* pp. 167-227.

Rudman, L.A., Moss-Racusin, C.A., Phelan, J.E., & Nauts, S. (2012a). Status incongruity and backlash effects: Defending the gender hierarchy motivates prejudice against female leaders. *Journal of Experimental Social Psychology, 48*, 165-179.

Rudman, L.A., & Phelan, J.E. (2008). Backlash effects for disconfirming gender stereotypes in organizations. *Research in Organizational Behavior, 28,* 61-79.

Rudman, L.A., & Phelan, J.E. (2010). The effect of priming gender roles on women's implicit gender beliefs and career aspirations. *Social Psychology, 41*, 192-202.

Sandberg, D. E., Meyer-Bahlburg, H. F., Ehrhardt, A. A., & Yager, T. J. (1993). The prevalence of gender-atypical behavior in elementary school children. *Journal of the American Academy of Child & Adolescent Psychiatry, 32*, 306-314.

Sandnabba, N. K., & Ahlberg, C. (1999). Parents' attitudes and expectations about children's cross-gender behavior. *Sex Roles, 40*, 249-263.

Sherman, J. W., & Frost, L. A. (2000). On the encoding of stereotype-relevant information under cognitive load. *Personality and Social Psychology Bulletin, 26*, 26-34.

Stangor, C., & McMillan, D. (1992). Memory for expectancy-congruent and expectancy-incongruent information: A review of the social and social developmental literatures. *Psychological Bulletin, 111*, 42-61.

Stangor, C., & Ruble, D. N. (1989a). Strength of expectancies and memory for social information: What we remember depends on how much we know. *Journal of Experimental Social Psychology, 25*, 18-35.

Stangor, C., & Ruble, D. N. (1989b). Differential influences of gender schemata and gender constancy on children's information processing and behavior. *Social Cognition, 7*, 353-372.

Todorov, A., Baron, S. G., & Oosterhof, N. N. (2008). Evaluating face trustworthiness: A model based approach. *Social Cognitive and Affective Neuroscience, 3*, 119-127.

Todorov, A., Dotsch, R., Wigboldus, D. H. J., & Said, C. P. (2011). Data-driven methods for modeling social perception. *Social and Personality Psychology Compass, 5*, 775-791.

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science, 308*, 1623-1626.

Todorov, A., Olivola, C.Y., Dotsch, R., & Mende-Siedlecki, P. (in prep). Social attributions from faces: Determinants, consequences, accuracy and emotional significance.

Todorov, A., Pakrashi, M., & Oosterhof, N. N. (2009). Evaluating faces on trustworthiness after minimal time exposure. *Social Cognition, 27*, 813-833.

Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences, 12*, 455-460.

Uleman, J.S., Newman, L. S., & Moskowitz, G. B. (1996). People as flexible interpreters: Evidence and issues from spontaneous trait inference. In M. P. Zanna (Ed.), *Advances in Experimental Social Psychology* (Vol. 28, pp. 211-279). San Diego, CA: Academic Press.

United Nations Women (2010). Facts and figures: leadership and political participation. Retrieved from http://www.unwomen.org/en/what-we-do/leadership-and-political-participation/facts-and-figures.

US Department of Health and Human Services (2010). The registered nurse population: Findings from the 2008 national sample survey of registered nurses. Retrieved from: http://bhpr.hrsa.gov/healthworkforce/rnsurveys/rnsurveyfinal.pdf.

Vonk, R. (1998). The slime effect: Suspicion and dislike of likeable behavior toward superiors. *Journal of Personality and Social Psychology, 74*, 849-864.

Wigboldus, D. H. J., Dijksterhuis, A., & van Knippenberg, A. (2003). When stereotypes get in the way: Stereotypes obstruct stereotype-inconsistent trait inferences. *Journal of Personality and Social Psychology, 84,* 470-484.

Wigboldus, D. H. J., Sherman, J.W., Franzese, H.L., & van Knippenberg, A. (2003). Capacity and comprehension: Spontaneous stereotyping under cognitive load. *Social Cognition, 22*, 292-309.

Wollstonecraft, M. (1792/2004). A vindication of the rights of women. Harmondsworth, UK: Penguin Classics.

Woolf, V. (1929/2012) *A room of one's own.* Eastford, CT: Martino Fine Books.

Yan, X., Wang, M., & Zhang, Q. (2012). Effects of gender stereotypes on spontaneous trait inferences and the moderating role of gender schematicity: Evidence from Chinese undergraduates. *Social Cognition, 30,* 220-231.

Young, R., & Sweeting, H. (2004). Adolescent bullying, relationships, psychological well-being, and gender-atypical behavior: A gender diagnosticity approach. *Sex Roles, 50*, 525-537.

Zucker, K. J., & Bradley, S. J. (1995). *Gender identity disorder and psychosexual problems in children and adolescents.* New York: Guilford Press.

English summary

Women striving to obtain a managerial position face a difficult Catch-22: they need to behave agentically in order to prove that they are sufficiently competent for the job, but are disliked if they do (Rudman, 1998). As a result of this *backlash-effect*, women often have to choose between being respected (when behaving agentically) and being liked (when behaving communally): a choice not faced by men. Either way, they are less likely to be hired than their male counterparts (Eagly, & Karau, 2002; Rudman, & Phelan, 2008; Rudman et al., 2012b) and are met with more negative responses by their subordinates even if they are hired (Butler, & Geis, 1990).

Backlash research suggests that people often respond negatively to women who portray stereotypically masculine behaviors (such as agency), but do people also respond negatively to men who portray stereotypically feminine behaviors? Men may be disliked and casted off as weak or psychologically unstable if they engage in stereotypically feminine behaviors such as modesty or self-disclosure (Derlega, & Chaikin, 1976; Moss-Racusin et al., 2010), suggesting that men, too, may suffer from backlash. Unfortunately, there is relatively little research on backlash towards men, and existing theories (e.g., Role Congruity Theory; Eagly, & Karau, 2002) focus on explaining backlash towards women. In the present dissertation, I will discuss a theory that aims to present an integrative view on backlash towards both genders: the Status Incongruity Hypothesis (SIH; Rudman, Moss-Racusin, Phelan, & Nauts, 2012a). Moreover, in every chapter, I introduce a new paradigm for studying backlash, that is more ecologically valid than existing paradigms (Chapter 2), or is more suitable for studying backlash against men (Chapters 3, 4 and 5).

According to the Status Incongruity Hypothesis, men have more societal status than women, and people are (often unconsciously) motivated to protect and maintain this status quo, since doing so serves an important palliative function (e.g., to reduce guilt and anxiety; Jost, & Hyunyadi, 2004). Women who portray high status behaviors (e.g., dominance) and men who portray low status behaviors (e.g., weakness) threaten the status quo, and backlash serves to put them "back in their place" as a way of restoring the gender status quo. Not hiring agentic women or casting weak men off as psychologically unstable serves to penalize this behavior in order to preserve male hegemony.

To uncover whether the motivation to protect the status quo indeed underlies backlash, I tested whether people respond more negatively to gender deviant behavior when they are motivated to protect the status quo. In Chapters 2, 4 and 5, I measured individual differences in people's motivation to protect the status quo (also called *system justifying motives*); in Chapters 3 and 4, I used an experimental manipulation to temporarily strengthen or weaken this motivation. System justifying motives predicted backlash (with the exception of the findings presented in Chapter 5), regardless of whether they were experimentally altered, or measured as an individual difference variable. These findings corroborate the SIH's contention that the motivation to protect the status quo underlies backlash against both genders.

Another prediction that follows from the SIH is that not all gender atypical behaviors should lead to backlash, but only those behaviors that are status incongruent. Put differently: men are allowed to be a little feminine, as long as they do not portray low status behaviors (e.g., being scared or weak), and women are allowed to be a little masculine, as long as they do not portray high status behaviors (e.g., being dominant or aggressive). This hypothesis was tested in Chapters 3 and 5. In line with the SIH, the results of Chapter 3 suggest that men are only penalized for status incongruent behavior, but not for other types of gender atypical behavior. However, in Chapter 5 no support for this hypothesis was found. Due to these null results, more research is needed to test this hypothesis.

In sum, the research presented in this dissertation used different methods to study why people often respond negatively to gender atypical behavior. This research suggests that system justifying motives underlie backlash-effects, corroborating the Status Incongruity Hypothesis. Dominant, agentic or assertive women and weak, nervous or scared men threaten people's worldview. Backlash seems to serve to put them back in their proper place.

Nederlandse samenvatting

Vrouwen die een leidinggevende positie ambiëren bevinden zich in een lastig parket: ze moeten zich assertief en pittig opstellen om duidelijk te maken dat ze voldoende competent zijn voor de baan, maar als ze dat doen, worden ze onaardig gevonden (Rudman, 1998). Als gevolg van dit *backlash-effect* moeten vrouwelijke managers vaak kiezen: als ze zich stereotiep vrouwelijk opstellen worden ze aardig gevonden, maar niet gerespecteerd; als ze zich stereotiep mannelijk opstellen worden ze gerespecteerd, maar niet aardig gevonden. In beide gevallen hebben ze minder kans om aangenomen te worden dan mannen die exact hetzelfde gedrag vertonen (Eagly, & Karau, 2002; Rudman, & Phelan, 2008; Rudman et al., 2012b). Als ze wel aangenomen worden, hebben vrouwen minder kans op een promotie dan hun mannelijke collega's, en worden ze vaker tegengewerkt door hun ondergeschikten (Butler, & Geis, 1990; Heilman, 2001).

Volgens onderzoek naar het backlash-effect reageren mensen negatief op vrouwen die zich typisch mannelijk gedragen, maar hoe zit het met mannen die zich typisch vrouwelijk gedragen? Mannen worden onaardig of psychologisch instabiel gevonden wanneer ze zich bescheiden opstellen of over hun gevoelens praten (Derlega, & Chaikin, 1976; Moss-Racusin et al., 2010), hetgeen suggereert dat ook mannen last kunnen hebben van backlash. Helaas is er relatief weinig onderzoek naar backlash bij mannen en richten bestaande theorieën (bijv. de Rol Congruentie Theorie; Eagly, & Karau, 2002) zich voornamelijk op het verklaren van backlash bij vrouwen. In dit proefschrift wordt de Status Incongruentie Hypothese (SIH; Rudman, Moss-Racusin, Phelan, & Nauts, 2012a) besproken, een theorie die backlash ten opzichte van zowel mannen als vrouwen tracht te verklaren.

Gender atypisch gedrag (mannelijk gedrag bij vrouwen, vrouwelijk gedrag bij mannen) roept negatieve reacties op, maar waarom zijn mensen gemotiveerd om gender atypisch gedrag af te straffen? Volgens de Status Incongruentie Hypothese hebben mannen meer maatschappelijke status dan vrouwen en zijn mensen (vaak onbewust) gemotiveerd om deze status quo te beschermen. Vrouwen die hoge-statusgedrag vertonen (bijv. door zich pittig of dominant op te stellen) en mannen die lage-statusgedrag vertonen (bijv. door zich bescheiden of zwak op te stellen) bedreigen die status quo en backlash heeft als doel om deze mensen "terug op hun plek" te zetten om zo de status quo te beschermen. Het in stand houden van de

status quo lijkt een duidelijke psychologische functie te hebben: het kan mensen helpen om gevoelens van onzekerheid, angst en schuld tegen te gaan, en daarom zijn mensen vaak onbewust gemotiveerd om de status quo te beschermen (Jost, & Hyunyadi, 2004). Door een pittige vrouw niet aan te nemen of een nerveuze man af te doen als psychologisch instabiel kunnen mensen dit gedrag afstraffen, om zo hun wereldbeeld in stand te houden.

Om te onderzoeken of de motivatie om de status quo te beschermen inderdaad ten grondslag ligt aan het backlash-effect is in dit proefschrift in vier empirische hoofdstukken onderzocht of mensen negatiever reageren op gender atypisch gedrag naarmate ze sterker gemotiveerd zijn om de status quo te beschermen. In Hoofdstukken 2, 4 en 5 werd gekeken naar individuele verschillen in de motivatie die mensen hebben om de status quo te beschermen; in Hoofdstukken 3 en 4 werd deze motivatie tijdelijk verhoogd met behulp van een experimentele manipulatie. Zowel individuele verschillen in de motivatie om de status quo te beschermen als een tijdelijk geïnduceerde motivatie voorspelden backlash (met uitzondering van in Hoofdstuk 5). Dit is in lijn met de Status Incongruentie Hypothese, die voorspelt dat backlash voortkomt uit de motivatie om de bestaande gender hiërarchie te beschermen.

Een andere voorspelling die op basis van de Status Incongruentie Hypothese (SIH) kan worden gedaan is dat niet iedere vorm van atypisch gedrag tot backlash leidt: gedrag wordt alleen afgestraft als het status incongruent is. Met andere woorden: mannen mogen zich best een beetje vrouwelijk gedragen, maar ze mogen geen lage-statusgedrag vertonen (bijv. bang zijn, bescheiden zijn), en vrouwen mogen zich best een beetje mannelijk gedragen, maar ze mogen geen hoge-statusgedrag vertonen (bijv. dominant zijn, arrogant zijn). Deze hypothese werd in Hoofdstukken 3 en 5 getoetst. Hoofdstuk 3 suggereert in lijn met de SIH dat mannen alleen afgestraft worden voor lage-statusgedrag en niet voor andere vormen van vrouwelijk gedrag. Echter, Hoofdstuk 5 liet geen effecten zien in lijn met deze hypothese. Er is daarom meer onderzoek nodig om deze hypothese te toetsen.

Om de Status Incongruentie Hypothese te toetsen is in ieder hoofdstuk van dit proefschrift een nieuwe methode geïntroduceerd om backlash te meten. In bestaand onderzoek wordt veelal gebruik gemaakt van een video van een sollicitatiegesprek, maar deze methode heeft een

aantal beperkingen. In Hoofdstuk 2 werd gebruik gemaakt van een live sollicitatiegesprek om de situatie in een echt sollicitatiegesprek beter na te bootsen en zo de ecologische validiteit van backlash-onderzoek te verhogen. In Hoofdstukken 3,4 en 5 werden paradigma's ontwikkeld die beter geschikt zijn om backlash ten opzichte van vrouwen én mannen te onderzoeken, aangezien het bestaande video-paradigma specifiek is ontwikkeld om backlash ten opzichte van vrouwelijke sollicitanten te bestuderen. Daarnaast was het doel van deze nieuwe paradigma's om backlash op een meer subtiele, indirecte wijze te meten. In Hoofdstuk 3 werden daartoe pictorale taken geïntroduceerd waarmee in kaart kan worden gebracht hoe mensen denken dat een atypische man eruit ziet (de Reverse Correlation Image Classification Task; Dotsch et al., 2008, en de Draw-a-Face-Task; Nauts et al., in prep). In Hoofdstuk 4 werd onderzocht of mensen, wanneer ze gemotiveerd zijn om de status quo te beschermen, atypisch gedrag beter onthouden. In Hoofdstuk 5 werd bestudeerd of mensen *spontaan* een sterkere inferentie vormen van atypisch gedrag dan van typisch gedrag, maar ik kon hier geen evidentie voor vinden.

Samenvattend werden er in dit proefschrift verschillende methodes gebruikt om te onderzoeken waarom mensen negatief reageren op gender atypisch gedrag. In lijn met de Status Incongruentie Hypothese lijken mensen gemotiveerd om status incongruent gedrag af te straffen om zo de status quo te beschermen. Vrouwen die pittig, assertief of dominant zijn en mannen die bang, zenuwachtig of zwak zijn bedreigen ons wereldbeeld. Backlash lijkt te dienen om deze mensen terug op hun plek te zetten.

Acknowledgements

In A.A. Milne's "Winnie-the-Pooh Goes Visiting and Gets into a Tight Spot", Winnie-the-Pooh visits Rabbit, but he gets stuck in the doorway on his way out. He wiggles his paws with all his might, but the harder he tries, the more he gets stuck in the rabbit hole. Luckily, Pooh is never alone. His friends keep him company and eventually, all the animals in the forest line up to help Pooh out. Together, they pull and pull and pull, and with a loud "plop", Pooh comes flying out of the rabbit hole, sporting a big smile.

During my PhD-project, I sometimes felt stuck, just like Winnie-the-Pooh. I tried as hard as I could to get out of the rabbit hole and into the Wonderland beyond, but the harder I wiggled my paw, the more stuck I seemed to get. Luckily, everyone lined up to help me out.

When Pooh gets stuck, Christopher Robin immediately comes to his rescue: calmly assessing the situation, reassuring Pooh, and making sure that Pooh gets his paws back on solid ground. Optimistic and patient, witty and wise, he is always there to save the day, even though Pooh can be a Bear of Little Brain sometimes.

Daniël, thank you for coming to my rescue; for rooting for me, null result after null result, SNAFU after SNAFU, with unwavering enthusiasm and relentless optimism. Thank you for being such a true, no-shortcuts-scientist amidst all turmoil, for your razor sharp mind and your uncanny ability to spot methodological shortcomings. With you, doing research is never about ego's or impact factors: it's about scientific curiosity, cool methods, and rigorous counterbalancing. I hope that you know how grateful I am for everything you have done for me, and how immensely proud I am to be your student.

Oliver, my co-promotor: thank you for your smart, meticulous questions and comments, the 3 AM programming fixes, but most of all: thank you for being a friend. Figuring out how to do content-analysis with some pasta salad is exactly my idea of having a good time. You were an invaluable part of our team, and I had the best time getting to the bottom of that primacy-of-warmth Woozle-effect together.

Laurie, were should I even begin? I keep being flabbergasted by how much you know, by how outlandishly smart you are, by how much I learn from you in every conversation that we have. You are brilliant and funny, and I would sell my soul to the devil to be able to write as well as you do. Thank you (and Bob!) for all the great conversations, bike-

borrowing, grilled portobello's, blasphemy and tap-dancing guys in glittery vests. It's been a blast!

Pooh and Rabbit don't always agree on things: can Rabbit use Pooh's legs as a towel rack? Is Pooh's wedged situation due to the size of Rabbit's door, or to the size of Pooh's belly? Despite their differences of opinion, Pooh is grateful for everything he has learned while being stuck in the rabbit hole. Thank you, Roos for initiating the project. I learned a lot from working with you, and I really appreciate that you gave me room to roam when I needed to.

Then, there is Reine. Kind, caring Reine always makes sure that everyone is doing all right. If it weren't for her, Pooh (that Silly Old Bear!) would constantly be bumping into things. Reine, thanks for taking such great care of me, for all the conversations, insights, dropjes, cards and onesies. You are an exemplary scientist, and I could not have wished for a better roomie and friend!

Hannah: we no longer live in the same Forest, but I have fond memories of illegally sharing my office in the early days, walking in Memphis, playing secretaries (with a glass of Hefeweizen), and so many other things. Thank you for your friendship over the years, and for being my paranymph.

Pooh's good friend Piglet is small but brave, with a darn sharp mind behind a cute façade. Wieteke & Isabelle: thank you for all your support and gossip, and for being more than meets the eye.

When Pooh is a Wedged Bear in Great Tightness, there are lots of animals who come by to tell him stories and keep him happy. Our weekly lab meetings did the same for me, and I would like to thank everyone who was part of them. Johan, thanks for your creative ideas and your clever, constructive comments. Ron, thanks for always helping everyone, for being so responsible and kind. Thijs, you are like a ridiculously smart version of Tigger: playful, but with a whiz-kid-core. Thanks for keeping things interesting.

There were plenty of others who kept me smiling, like my In-Mind gals, Eefje, Jellie and Inge, my summerschool buddies and ASPO dissertation committee-friends. Special thanks go out to Marijn, for all the cake and ice cream we enjoyed when we both had Pooh-sized bellies. Tom

& Tirza: thanks for sharing your ideas with me, as well as a roof. Mike: your thoughtful advice made all the difference.

The BSI is blessed with a wonderful support staff. Marjo, Marijke and Madelon: thanks for running the department in such a nice way. Our great lab manager, Ronny: thank you for all the coffee and conversation. Trudy: I very much appreciate how you patiently explained the "boekhouding" for the Geis award to me (again, and again, and again, and…). Meta, I realize that handling the finances for my many studies was a lot of work, and I am grateful that you always did so with a smile. Ron and Cor: thanks for making everyone feel at home in the Spinoza building.

I would also like to thank Bart, Denise, Floor, Joel, and all my new colleagues at Utrecht's self-regulation lab for creating such a nice, collaborative atmosphere, and for doing such great research. If Pooh would have read your papers, he would never have eaten so much honey and condensed milk!

After a day of philosophizing about Jagulars, Woozles and Heffalumps, Pooh likes to unwind with his friends. I would like to thank all my friends and family for the nice conversations, camping trips and beers we enjoyed together. Mason & Ria, my amazing parents-in-law who take such great care of us. My awesome friends, Chrissie, Maartje, Marjan, Mehdi, Mirella, Myrthe, Nicolette & Wenneke: you are the best. You all know that my brevity here is by no means a reflection of how important you are to me.

My fondness of the Bear of Little Brain can be traced back to the many times my mother read me the stories of Alice, Alibaba, Pooh and Pippi. From creative crafts-ideas and ingenious birthday parties to a purple kayak and Owl-style discussions about anything: special thanks go out to my parents.

In Milne's original book, the picture that accompanies "Winnie-the-Pooh Goes Visiting and Gets Into a Tight Spot" shows Pooh with all the animals that are pulling him out of the rabbit hole. The picture that I have at home shows Pooh with just two animals: Kanga and little baby Roo. Kanga is talking to Pooh while baby Roo is showing him some wildflowers. Winnie-the-Pooh looks at them, smiling from ear to ear. He may be a Wedged Bear in great Tightness, with legs that are used as a towel rack, but as long as Kanga and Roo are in the picture, he is perfectly happy.

Curriculum Vitae

Sanne Nauts was born on January 24th 1985 in Celle, Germany. She completed her primary education at Dutch schools in Blomberg, Stolzenau and Rickelrath and her secondary education in Roermond (Stedelijk Lyceum). In 2003, Sanne started studying psychology at Radboud University Nijmegen. She graduated from the researchmaster Behavioural Science in 2009 (cum laude) and started her PhD-project later that year. Sanne currently works as a postdoctoral researcher at Utrecht University, as part of an interdisciplinary research project on bedtime procrastination headed by prof. dr. Denise de Ridder (department of health psychology) and dr. Joel Anderson (department of philosophy).