Sex Roles (2023) 88:268–289
https://doi.org/10.1007/s11199-023-01354-0

ORIGINAL ARTICLE

People Overestimate Backlash Against Helpers Who Violate Gender Stereotypes: Experimental Examination of a Prosociality Paradox

Ciara Atkinson[1] · **Gillian Sandstrom**[2] · **Alyssa Croft**[1]

Accepted: 8 February 2023 / Published online: 11 March 2023
© The Author(s) 2023

Abstract

Men and women typically help others in gender stereotypic ways (gender-consistent helping), but how might people judge helpers who do so in counter-stereotypic ways (gender-inconsistent helping)? Most of the time helpers are viewed favorably, but behaviors that deviate from gender stereotypes tend to elicit social sanctions from others. Thus, gender-inconsistent helping presents a paradox wherein people may *anticipate* facing negative judgments from others despite helping being a positive, prosocial act. Across three experiments (two pre-registered), participants provided their own (Studies 1–3) and normative (Studies 2–3) evaluations of gender-consistent and gender-inconsistent helpers. Taken together, results revealed that participants expected other people to evaluate gender-inconsistent helpers less favorably than gender-consistent helpers (Hypothesis 1), and less favorably than they *actually* did themselves (Hypothesis 2). These findings show that gender-inconsistent helping is less susceptible to backlash than people think, and instead suggest that pluralistic ignorance could be a barrier to gender-inconsistent helping, if people fear that others' judgments of gender-inconsistent helpers are harsher than their own. Our results highlight novel opportunities for addressing persistent occupational gender segregation in prosocial contexts (by confronting pluralistic ignorance), which could subsequently enhance gender equality more broadly.

Keywords Gender roles · Stereotypes · Prosocial behavior · Pluralistic ignorance

Helping, like other prosocial behaviors—actions designed to benefit others—is cross-culturally valued (Klein et al., 2015). Despite the positive nature of helping, previous research shows that our participation in these behaviors is regulated by social norms, with a variety of factors promoting or hindering helping engagement. In particular, numerous studies demonstrate that gender roles shape the way people help others, showing that people usually perform *gender-consistent* helping—assistance in line with gender roles—and shy away from *gender-inconsistent* helping—assistance at odds with gender roles (e.g., Atkinson et al., 2021; Eagly, 2009; Eagly & Crowley, 1986). For instance, research suggests that it is highly male-stereotypic to help someone with yard work or household repairs, so it would be gender-consistent for a man to help in this way and gender-inconsistent for a woman to help in this way (Atkinson et al., 2021). Moreover, the same research indicates that it is highly female-stereotypic to help someone choose what to wear for a job interview or first date, meaning that it would be gender-consistent for a woman to engage in this type of helping but gender-inconsistent for a man to do so (Atkinson et al., 2021).

One specific barrier hindering participation in gender-inconsistent helping (i.e., women doing male-stereoypic helping tasks; men doing female-stereotypic helping tasks) arises from concerns about negative social evaluations and penalties. In a recent study exploring factors that hinder gender-inconsistent helping, Atkinson and colleagues (2021) found that people reported lower intentions to engage in gender-inconsistent helping behaviors than gender-consistent behaviors, in part because they expected other people would not approve of them engaging in helping that violates gender roles. While these concerns for penalties contribute to decreased likelihood of helping in a gender-inconsistent way, it is unclear how these helpers are *actually* perceived by others. Do people truly experience social penalties when helping others in a gender-inconsistent way, despite the reputational

✉ Gillian Sandstrom
g.sandstrom@sussex.ac.uk

[1] Department of Psychology, University of Arizona, Tucson, AZ, USA

[2] Department of Psychology, University of Sussex, Brighton, UK

rewards that helpers generally experience? Alternatively, these helping scenarios may be contexts of *pluralistic ignorance*, whereby people privately evaluate gender-inconsistent helping favorably, but anticipate other group members to penalize these helpers. The present research addressed these questions by comparing people's personal judgements of gender-consistent and gender-inconsistent helpers to their expectations for others' judgements.

People Sometimes Perceive Helpers Negatively

Helping is universally valued, and helpers are generally regarded favorably and accrue reputational rewards, with observers rating them as more trustworthy and higher in status and influence (Klein & Epley, 2014; Klein et al., 2015; Willer, 2009; Willer et al., 2010). However, in practice, things are more complicated. People draw inferences about helpers based on how they choose to help, and who they choose to help.

Helping can signal, or make salient, a status difference between the helper and the help recipient; when people are in a position to help someone, it often means that they have resources which the help recipient does not. Research on assumptive help suggests that when people receive task-specific help from advantaged helpers that they have not asked for, it triggers negative self-evaluations, and resentment towards the helper (Fisher et al., 1982; Nadler & Fisher, 1986; Nadler & Halabi, 2015). When help is provided through a gift of money from someone higher in perceived socioeconomic status, the reaction to help is more mixed. The financial help signals social identity threat, resulting in the help recipient feeling pitied and experiencing self-conscious negative emotions, while still recognizing the generosity and good intentions of the helper (Sandstrom et al., 2019).

Helping can signal not only higher status, but reveal the helper's attitudes towards the help recipient. Just as receiving money from someone higher in socioeconomic status can feel patronizing, receiving help from someone higher in social status can feel patronizing when it is perceived as signaling the help recipient's own inability. Benevolent sexism can sometimes be perceived as more positive than hostile sexism, because it encompasses the elements of sexism that are about cherishing women's purity and protecting femininity. Benevolent sexism is a set of beliefs rooted in the delicately balanced dynamic between heterosexual men and women, which depicts women as dependent on men but also acknowledges men as dependent on women (Glick & Fiske, 1997). People who endorse benevolent sexism tend to help in dependency-oriented ways more than autonomy-oriented ways (i.e., doing the task for the help recipient rather than showing them how to do it themselves), and tend to expect and seek dependency-oriented help (Shnabel et al., 2016). This is true both for men providing women with male-stereotypic help (e.g., on a mathematical/logical task; Shnabel et al., 2016), and women providing men with female-stereotypic help (e.g., cleaning a burned pot; Bareket et al., 2021). Male helpers who provide dependency-oriented help to women are often viewed negatively (Ruiz, 2019), but these perceptions depend on how the woman reacts to the help, and the level of benevolent sexism endorsed by the participant (Becker et al., 2011). In the current study, we manipulate the stereotypicality of the helping behavior (i.e., performing a task that is male- or female-stereotypic) rather than the stereotypicality of the helping context (i.e., who helps whom, and in what way).

Gender Stereotypes and Roles

Gender stereotypes exert a strong influence on behavior. According to the social role theory perspective on gender differences, men and women tend to enact roles and behaviors that are congruent with gender roles in a variety of contexts, including helping behaviors (for a review, see Croft et al., 2021; Eagly, 2009). Gender roles represent shared cultural beliefs about the values, attributes, and behaviors that are associated with men and women. Moreover, gender stereotypes can be both descriptive and prescriptive in nature, not only outlining how men and women typically are, but also dictating how they should or should not be (Prentice & Carranza, 2002). Social role theory contends that gender stereotypes arise from the unequal distribution of men and women into different roles, with women overrepresented in communal roles involving caregiving and interdependence, and men overrepresented in agentic roles involving leadership and independence. Repeated observations of men and women in distinct social roles facilitates gender stereotype development as perceivers infer the traits of an individual from their behaviors (Koenig & Eagly, 2014). For example, because men are observed in leadership roles more than women, people come to believe that men possess the agentic traits, like assertiveness or competitiveness, required to succeed in these roles. These stereotypes become incorporated into the self-concept during socialization, guiding the selection of behaviors—including helping behaviors—congruent with gender roles (e.g., Wood & Eagly, 2012).

Gender Differences in Prosocial Behavior

While there are no overarching gender differences in prosocial participation (i.e., there is not a more "helpful gender") numerous studies show the extent to which men and women help in a particular way depends largely on

whether it is a communal or agentic behavior (Atkinson et al., 2021; Eagly, 2009). Helping contexts drawing on communion, like caregiving or emotional support, typically elicit greater assistance from women, while contexts that call on agency, like independence or assertiveness, tend to elicit greater help from men. In the following section, we summarize prior research on gender differences in prosocial behavior and the role of gender stereotypes in shaping these helping tendencies.

In close, communal relationships, women are more likely to be regular emotional support providers (Burleson & Kunkel, 2006; Eagly et al., 2003), and women provide more ongoing care to members within their household than men by providing care for children and elderly relatives (Cancian & Oliker, 2000). These behaviors also carry over to the workplace with women being more expected to engage in recurring supportive behaviors to aid their colleagues (Farrell & Finkelstein, 2007). In a study examining gender differences in physician communication with patients, for instance, women physicians provided more psychosocial counseling and positive, emotionally focused talk compared to men physicians (Roter et al., 2002).

By contrast, men are overrepresented in helping situations that elicit agency, such as helping strangers, protecting people from harm, and helping that follows chivalrous norms. In a review of the psychological literature on heroism, Becker and Eagly (2004) concluded that men are more likely than women to perform heroic actions in emergency situations, in part because these actions draw on physical strength and personal risk. Men are overrepresented among those performing chivalrous helping behaviors, consisting of courteous and protective behaviors directed at those less powerful (e.g., women). For example, studies show that men are more likely to hold doors for women (Yoder et al., 2002), and assist a women confederate with picking up a dropped item when primed with romantic love in a laboratory study (Lamy et al., 2009).

Gender Stereotypes and Roles in Prosocial Contexts

The Gender Roles Inhibiting Prosociality (GRIP) model draws on both social role theory and the theory of planned behavior to explain the distal and proximal factors that predict helping behavior (Croft et al., 2021). First, in terms of the distal factors, social role theory suggests that there will be stereotypical divisions in helping behavior, which will reflect the stereotypical divisions in other social roles that men and women occupy in Western societies. The fact that women tend to help in certain ways (e.g., nurturing, relational) and men tend to help in other ways (e.g., physical, dominant) will lead people to believe that women *should* help in certain ways and men in others. These broad gender stereotypic beliefs will ultimately become internalized, and people will start to believe that *they should* help in gender-consistent ways, and that *others should* also help in gender-consistent ways.

According to the GRIP model, once gender stereotypes about helping are internalized, they influence more proximal predictors of helping behavior, as described by the theory of planned behavior: people's attitudes about helping, their perceptions of the norms about helping, and their belief about their own ability to help in certain ways. Indeed, in a recent empirical analysis, Atkinson and colleagues (2021) found that people generally favor helping opportunities that are congruent with gender roles, and perceptions of the subjective norms surrounding the behavior explained this process in a mediation analysis. In other words, people are reluctant to engage in gender-inconsistent helping, in part, because they believe that other people will judge them negatively for doing so. While these findings show that concerns about the consequences for violating gender roles restrict engagement in gender-inconsistent helping behaviors, findings are mixed on how gender-consistent and gender-inconsistent helpers are actually perceived by others. What follows is a brief discussion of prior research investigating perceptions of helping, particularly in gendered contexts.

How People Perceive Helpers: Rewards or Backlash?

Previous research suggests opposing predictions for how gender role violations in helping contexts are perceived. First, it is important to note that when helping is not gendered, or at least the gendered nature of the helping is not salient, helpers are generally perceived positively (Klein & Epley, 2014; Klein et al., 2015; Willer, 2009; Willer et al., 2010).

While empirical investigations of perceptions of *gendered* helping are sparse, some evidence suggests that helpers who violate gender roles may be evaluated *more* favorably than their gender congruent counterparts. For example, in a study examining evaluations of a person providing stereotypically masculine assistance in a high-risk emergency scenario, participants rated a woman more favorably than a man (Taynor & Deaux, 1973). Similarly, a study examining perceptions of workers who engaged in stereotypically feminine, organizational citizenship behaviors (e.g., helping a co-worker after hours at personal cost) found that participants evaluated men more favorably than women (Heilman & Chen, 2005).

However, there is also an extensive literature in non-helping contexts demonstrating that people who deviate from traditional gender roles incur social and economic penalties (e.g., Moss-Racusin et al., 2010; Rudman & Fairchild, 2004; Rudman et al., 2013). These *backlash effects* can range from negative interpersonal evaluations to overt

forms of discrimination, such as not being considered for a promotion. For example, "pushy" and assertive career-oriented women violate communal prescriptions, and accrue numerous social and economic penalties, including being viewed as less hirable and deserving of promotions (Phelan et al., 2008), and are more likely to face sabotage from their peers (Rudman & Fairchild, 2004). Working mothers also face penalties, being viewed as selfish and less effective mothers compared to stay-at-home mothers who prioritize caring roles (Brescoll & Uhlmann, 2005; Okimoto & Heilman, 2012). Gender-violating men also incur backlash when they fail to live up to agency and dominance prescriptions. People think men who excel in communal and stereotypically feminine roles should be paid less and are viewed as less favorable workers (Heilman & Wallen, 2010; Moss-Racusin & Johnson, 2016; Moss-Racusin et al., 2010). Family-oriented men who request family leave to provide care for children and relatives are also viewed as weak and more likely to encounter workplace penalties, like being demoted (Rudman & Mescher, 2013). Taken together, these findings paint an unclear picture for how gender-inconsistent helpers will be evaluated.

Pluralistic Ignorance and Perceptions of Gendered Helping

In the present research we propose that gender-inconsistent helping scenarios are contexts of *pluralistic ignorance*. Pluralistic ignorance refers to situations where an individual privately rejects a social norm, but erroneously believes that other people endorse the norm (Prentice & Miller, 1996). This discrepancy plays a powerful role in maintaining social norms and stereotypes and explains why people may engage in behaviors that are at odds with their personal beliefs (i.e., their own acceptance of the norms; Prentice & Miller, 1993). The existence of pluralistic ignorance has been demonstrated in a variety of contexts. People do not help in emergency situations because they think others will intervene (Latané & Darley, 1970), college students believe they are more uncomfortable with drinking practices than their peers (Prentice & Miller, 1993), and workers believe their colleagues endorse dominance and competitiveness more than they actually do (Munsch et al., 2018; Van Grootel et al., 2018).

Evaluations of gender-inconsistent helping (e.g., a man taking a home-cooked meal to a sick friend or a woman helping someone jumpstart their car) may be particularly susceptible to pluralistic ignorance. In line with previous research showing that helpers are generally evaluated positively, people may privately evaluate helpers favorably, regardless of the degree to which the behavior falls in line with gender stereotypes. By contrast, people's own observations and experiences with gender role violations may lead them to conclude that people who violate gender roles would face social sanctions and penalties (Rudman & Fairchild, 2004), and thus may lead people to anticipate that gender-inconsistent helpers will be penalized too. If both of these statements are accurate, it would result in a discrepancy between people's own evaluations of gender-consistent and gender-inconsistent helpers—which would be generally positive—and their estimations for how others would judge these helpers—which would be less favorable toward gender-inconsistent helpers.

Some support for this idea comes from another domain: asking questions after academic talks. A recent study found that people stereotype certain question-asking behaviors as feminine (thanking the speaker, complimenting the talk, greeting the speaker, referring to the speaker by title/surname), and others as masculine (asking a question that is really more than one question, asking a long-winded question, referring to their own research; Sandstrom et al., 2022). Women in this study expected to be judged, and were judged, as less likeable for asking a male- vs. female-stereotypic question, and men expected to be judged, and were judged, as lower in status for asking a female- vs. male-stereotypic question. However, there was little evidence for backlash, or fear of backlash, because men also expected to be judged, and were judged, as less likeable for asking a male- vs. female-stereotypic question, and women expected to be judged, and were judged, as lower in status for asking a female- vs. male-stereotypic question. In other words, own and normative judgments of question-askers seemed to be driven more by the perceived positivity of the question-asking behavior, rather than by whether the question was gender-consistent or -inconsistent.

Thus, in the current research, we predict that people will expect others to judge gender-inconsistent helpers negatively, because they are aware that gender violations are often subject to backlash. However, given that helping is generally perceived positively, we predict that people will privately evaluate gender-inconsistent helpers relatively favorably.

The Current Research

The current research examined perceptions of gendered helping, considering the role of pluralistic ignorance in evaluations. As a preliminary analysis, Study 1 measured people's evaluations of gender-consistent and gender-inconsistent helpers on dimensions of performance and deservingness of reward. Building on Study 1, we compared people's own evaluations of helpers to their estimations for how other people would judge these helpers in Studies 2 and 3. Across all studies, we predicted that people would expect others to judge gender-inconsistent helpers less favorably than gender-consistent

helpers (Hypothesis 1). Importantly, in Studies 2 and 3, we further hypothesized that people would expect others to judge gender-inconsistent helpers less favorably than they did in their own evaluations (Hypothesis 2). All materials, data, and syntax are available on OSF (https://osf.io/pruab), and the hypotheses, study design, and analysis plan for Studies 2 and 3 were formally pre-registered prior to data collection.

Study 1

We conducted a preliminary experiment to investigate people's own judgements of helpers. Participants were randomly assigned to read a vignette about a hypothetical man or a woman helping in a gender-consistent or gender-inconsistent way, then evaluated the helper's performance and deservingness of rewards. As an exploratory analysis, we investigated the extent to which people's judgments were moderated by sexist attitudes, measured using the ambivalent sexism inventory. As this study represents our first empirical examination of people's perceptions of gendered helping, we did not pre-register confirmatory hypotheses.

Method

Participants and Procedure

Eight hundred and twenty-three adults were recruited from MTurk as participants in this study. We aimed to recruit 800 participants because 1) we anticipated a small effect size (basing our estimate on the effect size found by Taynor & Deaux, 1973), and 2) we wanted to ensure that an adequate number of people rated each of the ten helping behaviors. Participants received $0.42 in exchange for their participation. To be eligible for this study, participants were required to (a) be at least 18 years of age, (b) be able to read and write in English, and (c) pass an honesty check. Participants completed the honesty check at the end of the survey, which asked them to disclose how closely they read their assigned helping vignette, using three options: (a) I didn't read it at all; (b) I skimmed it; and (c) I read it closely. Twenty participants did not answer this question, 25 participants selected "I didn't read it at all" and 100 participants selected "I skimmed it". Data from these 145 participants (17.6%) were excluded from analyses.

The final sample contained 678 participants. Participants identified as women (50.3%; $n = 341$), men (49.4%; $n = 334$), and non-binary (0.3%; $n = 3$). For a full description of participant demographics, see Table 1. A sensitivity analysis conducted in G*Power (Faul et al., 2007) indicated that with $\alpha = .05$ and $1–\beta = .80$, our final sample of 678 participants was sufficient to detect a between-subjects ANOVA with main effects and an interaction effect size of at least $f = .11$, equivalent to $d = .22$.

Prior to data collection, an institutional review board responsible for the ethical treatment of human participants reviewed and approved this research study, and all the studies reported in this paper. Eligible participants were invited to complete an online study examining perceptions of helping behavior. Participants were randomly assigned to read one scenario describing a man or a woman engaging in a stereotypically masculine or stereotypically feminine helping behavior. The ten helping behaviors were developed and pilot tested in prior research (for more details about the pilot test, see Atkinson et al., 2021) and are depicted in Table 2. After reading the helping scenario, participants completed a manipulation check in which they reported how expected it was for the helper to perform the described behavior. Then, participants reported their impressions of the helper's performance and deservingness of material and reputational rewards. Finally, participants reported their gender role attitudes and demographics. At the end of the survey, participants were fully debriefed on the true purpose of the study and received monetary compensation in exchange for their participation.

Measures

Manipulation Check: Helping Expectedness

After the experimental manipulation, participants reported how expected it was for the helping target to perform the behavior depicted in the scenario. Using a 4-item scale adapted from prior research (Bettencourt et al., 1997), participants evaluated the helper on the following items: (a) How *surprising* is it for a man/woman to help in this way? (reverse-scored) (b) How *unexpected* is it for a man/woman to help in this way? (reverse-scored) (c) How *typical* is it for a man/woman to help in this way? and (d) To what extent does this man/woman engaging in this behavior fit common stereotypes for men/women? All items were scored using a 7-point semantic differential scale, with lower values indicating the behavior was less expected and higher values indicating the behavior was more expected (e.g., 1 = *very atypical* to 7 = *very typical*). The four items were averaged to create an expectedness scale ($\alpha = .78$).

Table 1 Sample Demographics in Studies 1–3

	Study 1	Study 2	Study 3
Age	*M* = 37.33 *SD* = 12.02	*M* = 36.18 *SD* = 10.91	
18–24			9 (5.5%)
25–34			18 (10.9%)
35–44			29 (17.6%)
45–54			24 (14.5%)
55 – 64			40 (24.2%)
65–74			33 (20%)
75–84			11 (6.7%)
85 or older			1 (0.6%)
Ethnicity			
Aboriginal or Native	2 (0.3%)	2 (0.8%)	0 (0%)
Black or African American	67 (9.9%)	16 (6.3%)	17 (10.3%)
East Asian	15 (2.2%)	13 (5.1%)	5 (3%)
Hispanic or Latino	25 (3.7%)	9 (3.5%)	2 (1.2%)
Middle Eastern	1 (0.1%)	1 (0.4%)	1 (0.6%)
Southeast Asian	15 (2.2%)	7 (2.7%)	0 (0%)
White	533 (78.8%)	199 (77.7%)	137 (83%)
Other or mixed	18 (2.7%)	9 (3.5%)	2 (1.2%)
Sexual Orientation			
Bisexual	61 (9.0%)	16 (6.3%)	13 (7.9%)
Gay or lesbian	16 (2.4%)	9 (3.5%)	4 (2.4%)
Heterosexual	588 (87.0%)	228 (89.1%)	146 (88.5%)
Other	16 (0.4%)	2 (0.8%)	1 (0.6%)
Prefer not to identify	61 (1.0%)	0 (0%)	0 (0%)
Political Orientation			
Very liberal	83 (12.3%)		20 (12.1%)
Liberal	160 (23.7%)		32 (19.4%)
Somewhat liberal	82 (12.1%)		0 (0%)
Neutral	123 (18.2%)		39 (23.6%)
Somewhat conservative	83 (12.3%)		22 (13.3%)
Conservative	81 (12.0%)		0 (0%)
Very conservative	56 (8.3%)		50 (30.3%)
Prefer not to identify	7 (1.0%)		2 (1.2%)
Education			
Some high school or less	2 (0.3%)	1 (0.4%)	4 (2.4%)
High school diploma or equivalent	71 (10.5%)	31 (12.1%)	39 (23.6%)
Some college or university	152 (22.5%)	68 (26.6%)	39 (23.6%)
Vocational/technical certification	43 (6.4%)	15 (5.9%)	17 (10.3%)
University degree	250 (37.0%)	92 (35.9%)	37 (22.4%)
Some graduate/professional school	37 (5.5%)	10 (3.9%)	3 (1.8%)
Graduate/professional degree	118 (17.5%)	39 (15.2%)	26 (15.8%)
Prefer not to identify	3 (0.4%)	0 (0%)	0 (0%)
Income			
Under $20,000	82 (12.1%)	39 (15.2%)	29 (17.6%)
$20,0001-$50,000	236 (34.9%)	93 (36.3%)	52 (31.5%)
$50,001-$80,000	192 (28.4%)	78 (30.5%)	40 (24.2%)
$80,001-$100,000	75 (11.1%)	25 (9.8%)	17 (10.3%)
$100,001-$150,000	67 (9.9%)	15 (5.9%)	14 (8.5%)
$150,000 +	24 (3.6%)	6 (2.3%)	13 (7.9%)

Table 2 Helping Stimuli Used in Studies 1–3

Masculine Helping		Feminine Helping	
Helping someone jump-start a car	*	Helping someone choose what to wear (e.g., to a job interview, on a date, to a party)	*
Helping someone move into a new place	*	Calling someone who seems like they need to be cheered-up	*
Helping someone carry groceries, a heavy box, etc	*	Taking a home-cooked meal to a sick friend/neighbor	*
Teaching someone how to use their new electronic device/tool/appliance		Keeping someone company during a doctor's appointment (i.e., providing moral support)	
Doing yard work or household repairs for someone else (e.g., raking leaves, shoveling snow, mowing, fixing a leaky faucet)		Visiting folks at a retirement home or hospital	

All behaviors listed were used in Studies 1 and 2. Behaviors with an * were used in Study 3

Performance Evaluations

Participants reported their impressions of the helper's performance using a measure created for this study. The 5-item scale contained the following items: (a) This man's/woman's assistance was effective; (b) The person this man/woman assisted was satisfied; (c) This man/woman provided valuable assistance; (d) This man's/woman's assistance was worthwhile; and (e) This man/woman was able to meet this receiver's needs. All items were rated using a 7-point Likert scale ranging from 1 (*strongly disagree*) to 7 (*strongly agree*) and were averaged to create a performance evaluations scale ($\alpha = .94$).

Deservingness of Rewards

Participants rated the extent to which the helper deserved a series of rewards in exchange for their assistance. To measure deservingness of rewards, participants indicated how deserving the helper was for six rewards using the following items created for this study: "In response to the help provided, to what extent do you think the man/woman you read about is deserving of…" (a) a thank you card; (b) a gift; (c) gratitude from the person they assisted; (d) a reciprocal favor in the future; (e) recognition from their community; and (f) a reward. These six items were rated using a 7-point Likert scale ranging from 1 (*strongly disagree*) to 7 (*strongly agree*) and were aggregated into a single scale ($\alpha = .83$).

As an additional measure of rewards, we adapted a scale from prior research (Willer, 2009) to examine deservingness of reputational rewards. Participants were asked to rate the helper on six traits (trustworthy, cooperative, influential, a strong leader, status, and popular), and their agreement with three statements (e.g., has a large circle of friends; is deserving of respect from peers; and works well with others). These nine items were rated using a 7-point Likert scale ranging from 1 (*strongly disagree*) to 7 (*strongly agree*) and were averaged to create a single scale ($\alpha = .89$).

Demographics

Participants reported their gender, age, race/ethnicity, sexual orientation, political orientation, educational background, and income.

Additional Measures

While not reported in the manuscript, participants also made attributions for the helper's behavior on dimensions of effort, ability, luck and task ease. If we had found differences in people's perceptions of helpers, we planned to look at these aspects of the helping behavior as possible mediators. Given that we found no differences in perceptions, we saw no point in examining them (and indeed there were no main effects or interactions involving effort, ability or luck). We also included a measure of hostile and benevolent sexism and tested the extent to which people's perceptions of helpers were moderated by participant gender and endorsement of hostile and benevolent sexism. A full summary of the results of these exploratory analyses can be found in the Study 1 Supplementary Online Materials document on the project OSF page: https://osf.io/pruab.

Analytic Strategy

To examine evaluations of helpers, we conducted a series of 2 (help type: masculine helping vs. feminine helping) × 2 (helper gender: man helper vs. woman helper) between-subjects analyses of variance (ANOVAs) on expectedness ratings, performance evaluations, and deservingness of rewards. The means and standard deviations for all interaction effects are illustrated in Table 3.

Table 3 Means and Standard Deviations for Help Type x Helper Gender Interaction Effects in Study 1

	Masculine Helping		Feminine Helping	
	Man Helper	Woman Helper	Man Helper	Woman Helper
	M (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
Manipulation Check	4.96 (1.17)	4.04 (1.23)	3.79 (1.07)	5.02 (1.18)
Performance	6.24 (0.83)	6.25 (0.85)	6.22 (0.76)	6.11 (0.93)
Material Rewards	4.69 (1.32)	4.69 (1.27)	4.59 (1.19)	4.69 (1.34)
Reputational Rewards	5.32 (0.82)	5.40 (0.86)	5.50 (0.80)	5.39 (0.89)

Study 1 Results

Manipulation Check: Helping Expectedness

In line with the goals of the experimental manipulation, we observed a significant help type x helper gender interaction, $F(1, 674) = 144.759$, $p < .001$, $\eta_p^2 = .177$. Pairwise comparisons revealed that participants expected men to perform (a) more masculine helping than feminine helping ($p < .001$), and (b) more masculine helping than women ($p < .001$). Following a similar pattern, participants expected women to perform (a) more feminine helping than masculine helping ($p < .001$), and (b) more feminine helping than men ($p < .001$). In other words, participants found helping scenarios depicting gender-consistent helping more expected than the scenarios depicting gender-inconsistent helping. The main effects of help type, $F(1, 674) = 1.102$, $p = .294$, $\eta_p^2 = .002$, and helper gender, $F(1, 674) = 3.009$, $p = .083$, $\eta_p^2 = .004$, were not statistically significant.

Performance Evaluations

The main effects of help type, $F(1, 674) = 1.582$, $p = .209$, and helper gender, $F(1, 674) = .587$, $p = .444$, $\eta_p^2 = .001$, were not statistically significant. The help type x helper gender interaction was also not statistically significant, $F(1, 674) = .790$, $p = .374$, $\eta_p^2 = .001$. In summary, these findings suggest that people evaluated men and women helpers similarly, and their evaluations did not differ depending on whether the helping was in line with gender roles.

Deservingness of Rewards

Material Rewards

The main effects of help type, $F(1, 673) = .606$, $p = .606$, $\eta_p^2 < .001$, and helper gender, $F(1, 673) = .232$, $p = .630$, $\eta_p^2 < .001$, were not statistically significant. The help type x helper gender interaction was also non-significant, $F(1, 673) = .636$, $p = .636$, $\eta_p^2 < .001$.

Reputational Rewards

The main effects of help type, $F(1, 671) = 1.756$, $p = .186$, $\eta_p^2 = .003$, and helper gender, $F(1, 671) = .057$, $p = .812$, $\eta_p^2 < .001$, were not statistically significant. The help type x helper gender interaction was also non-significant, $F(1, 671) = 2.363$, $p = .125$, $\eta_p^2 = .004$.

Taken together, people think men and women are deserving of an equal number of material and reputational rewards, across gender-consistent and gender-inconsistent forms of helping.

Study 1 Discussion

Despite acknowledging that gender-inconsistent helping was atypical, participants were no more likely to reward these helpers in their judgements. Importantly, these findings proved similar for men and women, and regardless of whether the target helped in a gender-consistent or -inconsistent way. Taken together, these results suggest that gender role violations in helping contexts do not evoke the same social penalties and backlash that are evoked by violations in other contexts, such as the workplace. However, it is difficult to draw firm conclusions based on (even well-powered) null effects. In Study 2, we aimed to replicate the finding that people do not personally evaluate gender-inconsistent helpers negatively (or any differently than gender-consistent helpers) and extend this analysis to test if they believe other people in society have less favorable perceptions of these helpers.

Another limitation of Study 1 is that participants were asked to judge the helpers' performance but were not asked to evaluate the helpers themselves. Given that backlash can incur social penalties, we expected to find in Study 2 that not only do gender-consistent helpers provide equally beneficial help, but also are perceived equally positively.

Study 2

The goal of Study 2 was to examine potential discrepancies between people's own evaluations of gendered helpers and their normative evaluations. Drawing from theory and prior research on pluralistic ignorance, we hypothesized people would expect others to judge gender-inconsistent helpers less favorably than gender-consistent helpers (Hypothesis 1), and less favorably than they do themselves (Hypothesis 2). We also tested the extent to which these effects were moderated by sexist attitudes, measured using the ambivalent sexism inventory. The research questions that inform our hypotheses, the study design, and the analysis plan for Study 2 were formally pre-registered prior to data collection and all materials, data, and syntax are available on the project OSF page: https://osf.io/pruab.

Method

Participants and Procedure

Three hundred and five adults were recruited from MTurk as participants in this study. Participants received $1.25 in exchange for their participation. Our target sample size was 300 adults, determined by financial constraints for participant payment, and we reasoned that switching to a within-subjects design should require a sample size about half as large as in our between-subjects design in Study 1. We obtained data from an additional five participants who completed the study and failed to enter a completion code for payment. To be eligible for this study, participants were required to (a) identify as a man or a woman, (b) be at least 18 years of age, (c) be able to read and write in English, and (d) pass all comprehension checks. Three participants (1.0%) were excluded because they did not identify as a man or a woman. A total of 46 (15.1%) additional participants failed at least one of the three checks. Two comprehension checks appeared after the instructions and asked participants to correctly identify the evaluation condition they were assigned to: (a) I understand I am to provide my *own* perceptions; or (b) I understand I am to provide how *others in society* might perceive this individual. Thirty-eight participants (12.5%) answered at least one of these questions incorrectly and were excluded. A third check occurred at the end of the survey and asked participants to select, from a list, the helping behavior described in the assigned helping vignette. Eight (2.6%) participants selected an incorrect behavior and were excluded.

The final sample contained 256 participants (52% men; $n = 133$). For a full description of participant demographics, see Table 1. A sensitivity analysis conducted in G*Power (Faul et al., 2007) suggested that with $\alpha = .05$ and $1–\beta = .80$, our final sample of 256 was sufficient to detect a mixed between-within ANOVA with main effects and an interaction effect size of at least $f = .09$, equivalent to $d = .18$.

Like Study 1, participants were randomly assigned to read a vignette describing a person performing either a stereotypically masculine or stereotypically feminine helping behavior. On dimensions of favorability, warmth, and competence, participants reported their own evaluations of the helper and their normative evaluations of the helper (i.e., their estimations for how others in society would judge the helper). The order in which participants provided their own and normative evaluations was randomized and counterbalanced. After rating the helper, participants reported their gender role attitudes and demographics. Participants were fully debriefed on the true purpose of the study at the end of the survey and received monetary compensation in exchange for their participation.

Independent Variables

Rating Type

Participants reported their own and normative evaluations of helpers on all measures. When reporting their own evaluations, participants were asked to report how they would personally evaluate the helping target ("I would judge this man/woman…" 1 = *extremely unfavorably* to 7 = *extremely favorably*). When reporting normative evaluations, participants were asked to estimate how they expected other people in society to evaluate the helping target ("I think others in society would judge this man/woman …" 1 = *extremely unfavorably* to 7 = *extremely favorably*). The order in which participants provided their own and normative evaluations did not affect analyses involving our primary measures (see the project OSF page for a full summary of analyses testing for order effects: https://osf.io/pruab).

Help Type

We used the same helping behaviors as Study 1. Participants were randomly assigned to read about one helper performing a stereotypically masculine or stereotypically feminine behavior. For the sake of parsimony (and due to financial constraints), given that we found no main effects of helper gender or participant gender in Study 1, we did not manipulate helper gender in the design. Instead, the gender of the helping target was matched to the gender of the participant, such that men read about a man engaging in a helping behavior and women read about a woman engaging in a helping behavior.

We coded the scenarios describing a woman helping in a stereotypically feminine way and a man helping in a stereotypically masculine way as gender-consistent ($n = 130$). The scenarios describing a man helping in a stereotypically feminine way and a woman helping in a stereotypically masculine way were coded as gender-inconsistent ($n = 126$).

Measures

Manipulation Check: Helping Expectedness

Like Study 1, participants completed the same manipulation check assessing the expectedness of the helping scenario. We adjusted the reference group for each item depending on the rating type condition. For example, when providing their own evaluations, participants were asked to, "Consider how unexpected it is for this man/woman to help in this way." When providing their normative evaluations, participants were asked to, "Consider how unexpected you think others in society would think it is for this man/woman to help in this way." Due to a programming error, the first three items were scored using an 8-point semantic differential scale (e.g., 1 = *very atypical*, 8 = *very typical*), while the fourth item was scored using a 4-point Likert scale (1 = *does not fit stereotypes for men*, 4 = *fits stereotypes for men*). To create a composite score, we standardized responses on each item and calculated average scores for own evaluations ($\alpha = .80$) and normative evaluations ($\alpha = .87$).

Favorability

To evaluate expectations of helpers experiencing backlash, participants reported their favorability judgements using a scale adapted from prior research (Bettencourt et al., 1997). The semantic-differential scale asked participants to describe the helper using five word pairs: (a) likable/unlikable, (b) favorable/unfavorable, (c) positive/negative, (d) good/bad, and (e) high status/low status. In the own evaluations condition, each word pair was preceded by the statement, "I would judge this man/woman…", and in the normative evaluations condition, each word pair was preceded by the statement, "I think others in society would judge this man/woman…" All items were rated on a 6-point semantic-differential scale, with higher values indicating more favorable evaluations (e.g., 1 = *extremely bad*, 6 = *extremely good*). The five items were averaged to create a favorability index for own evaluations ($\alpha = .86$) and normative evaluations ($\alpha = .87$).

Warmth and Competence

As a further measure of how positively participants evaluated helpers, we asked participants to rate the helpers' *warmth* (warm/cold, friendly/unfriendly) and *competence* (competent/incompetent, capable/incapable) using a scale adapted from Cuddy et al. (2007). Each word pair was measured on a 5-point semantic-differential scale (e.g., 1 = *extremely cold*, 5 = *extremely warm*), with higher values corresponding to elevated warmth or competence. Like the previous measures, the reference group in the question text varied by rating type condition. The two warmth items were aggregated into a warmth index for own evaluations ($r = .82$) and normative evaluations ($r = .87$). Similarly, the two competence items were averaged to create a competence score for own evaluations ($r = .87$) and normative evaluations ($r = .89$).

Demographics

Participants reported their gender, age, race/ethnicity, sexual orientation, education, and income.

Additional Measures

As in Study 1, we assessed endorsement of hostile and benevolent sexism and ran exploratory analyses to test whether people's own and normative perceptions of helpers were moderated by endorsement of hostile and benevolent sexism. Although we did not find consistent evidence for moderation by benevolent sexism, analyses revealed that hostile sexism moderated participants' own evaluations of helper favorability. In general, people at both high and low levels of hostile sexism evaluated gender-consistent helpers more favorably than gender-inconsistent helpers. People at higher levels of hostile sexism evaluated gender-inconsistent helpers less favorably than those at lower levels of hostile sexism, but there were no significant differences in normative evaluations of gender-consistent helpers at high or low levels of hostile sexism. See the Study 2 Supplementary Online Materials document on the project OSF page for a full report of these analyses: https://osf.io/pruab.

Analytic Strategy

To test if people would expect others to judge gender-inconsistent helpers more negatively than gender-consistent helpers (Hypothesis 1), and more negatively than they do in their own evaluations (Hypothesis 2), we conducted a series of 2 (rating type: own evaluations vs. normative evaluations) × 2 (help type: gender-consistent vs. gender-inconsistent) mixed model ANOVAs. Rating type was a within-subjects factor, and help type was a between-subjects factor. The means and standard deviations for all interaction effects are illustrated in Table 4, and the ANOVA results are in Table 5.

Table 4 Means and Standard Deviations for Rating Type x Help Type Interaction Effects in Study 2

	Own Evaluation		Normative Evaluation	
	Gender-Consistent	Gender-Inconsistent	Gender-Consistent	Gender-Inconsistent
	M (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
Manipulation Check	.26 (.69)	-.27 (.79)	.40 (.69)	-.42 (.80)
Favorability	5.00 (.75)	5.01 (.70)	4.90 (.77)	4.63 (.88)
Warmth	4.37 (.63)	4.37 (.66)	4.33 (.68)	4.11 (.82)
Competence	4.20 (.69)	4.21 (.75)	4.18 (.72)	3.83 (.89)

Study 2 Results

Manipulation Check: Helping Expectedness

In line with the manipulation goals and Study 1 findings, we observed a significant main effect of help type on expectedness (see Table 5). People expected gender-consistent helping ($M = .33$, $SD = .69$) more than gender-inconsistent helping ($M = -.34$, $SD = .79$), $d = .90$, 95% CI [.65, 1.16]. Although we did not pre-register an interaction effect, we observed a significant rating type x help type interaction. Pairwise comparisons revealed that people expected gender-inconsistent helping less than gender-consistent helping in their own evaluations ($p < .001$, $d = .72$, 95% CI [.47, .98]), but even more so in their normative evaluations ($p < .001$, $d = 1.10$, 95% CI [.83, 1.36]). People also estimated that others would (a) expect gender-inconsistent helping less than they did themselves ($p = .01$, $d = .18$, 95% CI [.01, .36]), and (b) expect gender-consistent helping more than they did themselves ($p = .01$, $d = .20$, 95% CI [.03, .37]).

Table 5 Results from Rating Type x Help Type ANOVAs on Expectedness, Favorability, Warmth and Competence in Study 2

	F	df	*p*	η_p^2
Manipulation Check (Expectedness)				
Help type	63.59	1, 252	< .001	.20
Rating type	0.01	1, 252	.91	< .001
Help type x Rating type	13.87	1, 252	< .001	.05
Favorability				
Help type	2.79	1, 254	.10	.01
Rating type	27.94	1, 254	< .001	.10
Help type x Rating type	7.64	1, 254	< .01	.03
Warmth				
Help type	2.13	1, 254	.15	.01
Rating type	13.91	1, 254	< .001	.05
Help type x Rating type	7.63	1, 254	< .01	.03
Competence				
Help type	4.27	1, 254	.04	.02
Rating type	16.32	1, 254	< .001	.06
Help type x Rating type	13.86	1, 254	< .001	.05

Favorability, Warmth, and Competence

For each of these evaluations, any main effects were qualified by a significant rating type x help type interaction. Congruent with Hypothesis 1, pairwise comparisons revealed that people expected others to evaluate gender-inconsistent helpers less favorably, $p = .01$, $d = .33$, 95% CI [.08, .57], as lower in warmth, $p = .02$, $d = .29$, 95% CI [.05, .54], and as less competent, $p = .001$, $d = .43$, 95% CI [.18, .68], than gender-consistent helpers. In contrast, there was no difference in how people rated gender-inconsistent and -consistent helpers in their own evaluations of favorability, $p = .851$, $d = .01$, 95% CI [-.26, .23], warmth, $p = .959$, $d = .01$, 95% CI [-.25, .25], or competence $p = .91$, $d = -.01$, 95% CI [-.26, .23].

In line with Hypothesis 2, people expected others to evaluate gender-inconsistent helpers less favorably, $p < .001$, $d = .48$, 95% CI [.30, .66], as less warm, $p < .001$, $d = .35$, 95% CI [.17, .53], and as less competent, $p < .001$, $d = .46$, 95% CI [.28, .65], than they did in their own evaluations. By contrast, people did not expect others to judge gender-consistent helpers differently than they did themselves for favorability, $p = .073$, $d = .13$, 95% CI [-.04, .30], warmth, $p = .491$, $d = .06$, 95% CI [-.11, .23], or competence, $p = .82$, $d = .03$, 95% CI [-.14, .20].

In summary, these findings illustrate that like favorability evaluations, participants expect others to perceive gender-inconsistent helpers as lower in warmth and competence compared to (a) gender-consistent helpers, and (b) how they themselves see the helpers. People do not personally evaluate gender-consistent helpers negatively, but they anticipate that other people will do so.

Study 2 Discussion

Despite evaluating gender-inconsistent helpers no differently than gender-consistent helpers in their own evaluations, participants expected others to judge them less favorably—and lower in warmth and competence—than gender-consistent helpers. Supporting our prediction for a discrepancy between own and normative evaluations of helpers, Study 2 findings also reveal that people expect others to judge gender-inconsistent helpers less favorably than they

do themselves. Taken together, these results provide evidence in favor of the idea that people might not privately judge gender violations in helping contexts negatively, but they do anticipate other people in society to exhibit this backlash. Based on our earlier reasoning, this pattern of judgments fits with a pluralistic ignorance explanation.

Although the current study provided an initial test for the idea that own and normative evaluations of helpers will differ, there are several design limitations that we must note. First, participants provided their own and normative evaluations of a single gender-consistent or gender-inconsistent helper, potentially decreasing the generalizability of these findings. To remedy this design limitation, in Study 3 we employed a fully within-subjects approach and asked participants to report their evaluations of multiple gender-consistent and gender-inconsistent helpers. Further, unlike in Study 1, we matched the gender of the helper to the gender of the participant. Although this decision felt justified by the fact that we did not find any main effects of helper gender in Study 1 (and indeed was also driven by financial constraints), this design decision prevented us from testing for moderation by helper gender in normative evaluations of helping targets. Given prior research on backlash and theories of precarious masculinity (Vandello et al., 2008), it is possible that normative evaluations for gender-inconsistent men and women helpers will differ from one another. To test this possibility, in Study 3 participants evaluated men and women helpers performing gender-consistent and gender-inconsistent forms of helping.

Study 3

As in Study 2, we hypothesized that people would expect others to judge gender-inconsistent helpers less favorably than gender-consistent helpers (Hypothesis 1), and less favorably than they do themselves (Hypothesis 2). Utilizing a fully-within subjects design, Study 3 enabled us to investigate if there are gender differences in own and normative evaluations of gender-consistent and gender-inconsistent helpers. Participants read about various men and women engaging in gender-consistent and gender-inconsistent helping behaviors, then reported their own and normative evaluations of each individual. The hypotheses, study design, and analysis plan for Study 3 were formally pre-registered prior to data collection and all materials, data, and syntax are available on the project OSF page: https://osf.io/pruab.

Method

Participants and Procedure

One hundred and sixty-five adults (50.3% women; $n = 83$) were recruited from Qualtrics Panels as participants in this study. Our sample size was determined by financial constraints for participant payment, related to grant funding. To participate in the study, participants were required to (a) be at least 18 years of age or older and (b) be able to read and write in English. All participants met the eligibility criteria, and none were excluded from analyses. For a full description of participant demographics, see Table 1. A sensitivity analysis conducted in G*Power (Faul et al., 2007) suggested that with $\alpha = .05$ and $1–\beta = .80$, our sample of 165 was sufficient to detect a within-subjects ANOVA with main effects and an interaction effect size of at least $f = .11$, equivalent to $d = .22$.

In this fully within-subjects study, participants read about both gender-consistent and gender-inconsistent helpers and reported both their own and normative evaluations of each helper on dimensions of favorability and deservingness of rewards. The order in which participants provided their own and normative evaluation ratings was counterbalanced and these ratings were completed at least 4 days apart to reduce potential carryover effects. To illustrate, half of the participants provided their own evaluations of gender-consistent and gender-inconsistent helpers at the first timepoint, then were invited to provide their normative evaluations of the same helpers at the second timepoint. By contrast, the other half of participants provided their normative evaluations of gender-consistent and gender-inconsistent helpers at the first timepoint, then were invited to provide their own evaluations of the same helpers at the second timepoint. After reporting evaluations of helpers, participants provided demographic information, were fully debriefed on the true purpose of the study, and received monetary compensation in exchange for their participation.

Independent Variables

Rating Type

Like Study 2, participants provided their own and normative evaluations of helpers on all measures. The order in which participants completed own and normative evaluations did not affect analyses involving our primary measures (see the project OSF page for a full summary of analyses testing for order effects: https://osf.io/pruab).

Help Type and Helper Gender

Participants viewed a subset of the helping behaviors used in Studies 1 and 2 (see Table 2). New to Study 3, help type was a within-subjects variable, so participants viewed both gender-consistent and gender-inconsistent helping behaviors (see Appendix for examples). In total, participants read 12 helping vignettes: (a) three describing a woman helper performing masculine helping, (b) three describing a man helper

performing masculine helping, (c) three describing a woman helper performing feminine helping, and (d) three describing a man helper performing feminine helping. In each helping vignette, in addition to the written description, participants viewed a line-drawing depicting the helper engaging in the action. The helping illustrations were included to make the helping scenarios more concrete for participants, and they are posted on the project OSF page: https://osf.io/pruab.

Measures

Manipulation Check: Helping Expectedness

Participants rated the expectedness of the helping scenarios using a shortened version of the scale used in Studies 1 and 2 (Bettencourt et al., 1997). The scale contained the following items: (a) "Compared to men/women in general, how surprising do you think it is for this man/woman to help in this way?" and (b) "Compared to men/women in general, how unexpected do you think it is for this man/woman to help in this way?" Both items were scored using a 7-point semantic-differential scale (e.g., 1 = *extremely unexpected*, 7 = *extremely expected*). Responses to each item were aggregated to create average scores for own evaluations ($r = .95$) and normative evaluations ($r = .94$).

Favorability

The favorability of the helping target was assessed using two items from the favorability scale in Study 2 (Bettencourt et al., 1997). Using a 7-point semantic-differential scale, participants described the helper using two word pairs: likeable/unlikable and negative/positive, with higher values indicating more favorable evaluations of helpers. Responses on each item were aggregated to create an average favorability score for own evaluations ($r = .98$) and normative evaluations ($r = .97$).

Deservingness of Rewards

As in Study 1, participants rated the extent to which the helper deserved a series of rewards in exchange for their assistance. Using a shortened version of the scale from Study 1, participants reported deservingness of two material rewards (a reward; a gift), as well as two reputational rewards (a strong leader; popular). All items were scored on 7-point Likert scales ranging from 1 (*strongly disagree*) to 7 (*strongly agree*), with higher values corresponding to more rewards. Responses to the two material reward items were aggregated to create averages for own evaluations ($r = .99$) and normative evaluations ($r = .99$). Likewise, responses on the two reputational rewards items were aggregated to create averages for own evaluations ($r = .98$) and normative evaluations ($r = .98$).

Additional Measures

We included an exploratory measure of participants' intentions to engage in gender-consistent and gender-inconsistent helping behaviors. Participants also provided their own and normative implicit evaluations of helpers using a modified implicit association task (IAT). We do not report the results pertaining to these measures in the main text (see General Discussion for our reasoning), but rather report the results on the project OSF page, along with a full description of these materials: https://osf.io/pruab.

Analytic Strategy

To test the hypotheses that people would expect others to judge gender-inconsistent helpers more negatively than gender-consistent helpers (Hypothesis 1), and more negatively than they do themselves, in their own evaluations (Hypothesis 2), we conducted a series of 2 (rating type: own evaluations vs. normative evaluations) × 2 (help type: masculine vs. feminine) × 2 (helper gender: man vs. woman) within-subjects ANOVAs on our primary outcome variables. Results from all 3-way interactions are depicted in Table 6, and descriptive statistics are in Table 7.

We conducted exploratory analyses to examine if our results were moderated by participant gender. To do this, we first collapsed across helper gender (i.e., focusing on whether the helping was gender-consistent or -inconsistent, as in Study 2). We then conducted a series of exploratory 2 (rating type: own evaluation vs. normative evaluation) × 2 (condition: gender-consistent vs. gender-inconsistent) × 2 (participant gender: man vs. woman) mixed model ANOVAs. While we found that women evaluated helpers more favorably than men and men rated helpers as more deserving of material rewards (regardless of the help type or rating type), participant gender did not interact with rating type or condition to predict any of our primary outcome variables. See the project OSF page for a full report of these analyses: https://osf.io/pruab.

Study 3 Results

Manipulation Check: Helping Expectedness

We observed a significant main effect of helper gender on expectedness ratings, $F(1, 164) = 28.84$, $p < .001$, $\eta_p^2 = .15$. Overall, people expected women to help ($M = 4.08$, $SD = 1.56$) more than men ($M = 3.82$, $SD = 1.47$), $d = -0.17$, 95% CI [-0.33, -0.03]. The main effect of help type was also statistically significant, $F(1, 164) = 12.95$, $p < .001$, $\eta_p^2 = .07$, illustrating that people expected masculine helping ($M = 4.03$, $SD = 1.26$) more than feminine helping ($M = 3.88$, $SD = 1.56$), $d = 0.11$, 95% CI [-0.05, 0.26].

Table 6 Results from Rating Type x Help Type x Helper Gender ANOVAs on Expectedness, Favorability and Deservingness of Material and Reputational Rewards in Study 3

	F	df	*p*	η_p^2
Expectedness				
Help type	12.95	1, 164	< .001	.07
Helper gender	28.84	1, 164	< .001	.15
Rating type	3.08	1, 164	.08	.02
Help type x Helper gender	172.63	1, 164	< .001	.51
Help type x Rating type	.19	1, 164	.67	< .01
Helper gender x Rating type	.39	1, 164	.54	< .01
Help type x Helper gender x Rating type	1.00	1, 164	.32	.01
Favorability				
Help type	.36	1, 164	.56	< .01
Helper gender	17.19	1, 164	< .001	.10
Rating type	1.69	1, 164	.20	.01
Help type x Helper gender	22.11	1, 164	< .001	.12
Help type x Rating type	4.73	1, 164	.03	.03
Helper gender x Rating type	2.68	1, 164	.10	.02
Help type x Helper gender x Rating type	6.12	1, 164	.01	.04
Material Rewards				
Help type	44.05	1, 163	< .001	.21
Helper gender	10.78	1, 163	< .01	.06
Rating type	1.04	1, 163	.31	.01
Help type x Helper gender	1.79	1, 163	.18	.01
Help type x Rating type	5.10	1, 163	.03	.03
Helper gender x Rating type	.79	1, 163	.38	.01
Help type x Helper gender x Rating type	.98	1, 163	.33	.01
Reputational Rewards				
Help type	< .001	1, 163	.99	< .001
Helper gender	2.78	1, 163	.10	.02
Rating type	.23	1, 163	.64	< .01
Help type x Helper gender	3.23	1, 163	.07	.02
Help type x Rating type	4.02	1, 163	.05	.02
Helper gender x Rating type	.03	1, 163	.87	< .001
Help type x Helper gender x Rating type	.35	1, 163	.56	< .01

Table 7 Means and Standard Deviations for Rating Type x Help Type x Helper Gender Interaction in Study 3

	Own Evaluations				Normative Evaluations			
	Masculine Helping		Feminine Helping		Masculine Helping		Feminine Helping	
	Man Helper	Woman Helper	Man Helper	Woman Helper	Man Helper	Woman Helper	Man Helper	Woman Helper
	M (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
Expectedness	4.55 (1.67)	3.35 (1.43)	3.21 (1.34)	4.67 (1.76)	4.47 (1.65)	3.43 (1.34)	3.06 (1.22)	4.48 (1.72)
Favorability	6.19 (0.82)	6.16 (0.88)	6.14 (0.91)	6.27 (0.79)	6.19 (0.87)	6.13 (0.87)	5.96 (1.01)	6.26 (0.79)
Material Rewards	3.79 (1.97)	3.96 (1.96)	3.60 (2.09)	3.71 (2.08)	3.77 (1.97)	3.93 (1.97)	3.49 (2.06)	3.51 (2.10)
Reput. Rewards	5.18 (1.36)	5.16 (1.35)	5.16 (1.35)	5.28 (1.33)	5.26 (1.36)	5.25 (1.33)	5.16 (1.37)	5.25 (1.33)

These main effects were qualified by a significant help type x helper gender interaction, $F(1, 164) = 172.62$, $p < .001$, $\eta_p^2 = .51$. Congruent with the manipulation goals and the findings in Studies 1 and 2, pairwise comparisons revealed that people found masculine helping more expected when performed by men ($M = 4.51$, $SD = 1.66$) compared to women ($M = 3.54$, $SD = 1.39$), $p < .001$, $d = 0.64$, 95% CI [0.47, 0.80]. Participants also expected feminine helping more when it was performed by women ($M = 4.62$, $SD = 1.74$) compared to men ($M = 3.13$, $SD = 1.28$), $p < .001$, $d = -0.99$, 95% CI [-1.17, -0.80]. Among men helpers, participants expected masculine helping more than feminine helping, $p < .001$, $d = 0.94$, 95% CI [0.75, 1.12]. By contrast, among women helpers, participants expected feminine helping more than masculine helping, $p < .001$, $d = -0.69$, 95% CI [-0.86, -0.52]. Taken together, these findings confirm that gender-inconsistent helping is less expected than gender-consistent helping for men and women, but especially for men as illustrated by the larger effect sizes relative to women.

The main effect of rating type on helping expectedness was not statistically significant, nor were any of the two-way or three-way interactions involving rating type (see Table 6 for a full summary of the results).

Favorability

There was a significant main effect of helper gender on favorability ratings, $F(1, 164) = 17.19$, $p < .001$, $\eta_p^2 = .10$. Overall, helping performed by women ($M = 6.21$, $SD = 0.76$) was evaluated more favorably than helping performed by men ($M = 6.12$, $SD = 0.78$), $p < .001$, $d = -0.11$, 95% CI [-0.27, 0.04]. This main effect was qualified by a statistically significant help type x helper gender interaction, $F(1, 164) = 22.11$, $p < .001$, $\eta_p^2 = .12$, and a statistically significant help type x rating type interaction, which, in turn, were qualified by a statistically significant 3-way interaction, $F(1, 164) = 6.12$, $p = .01$, $\eta_p^2 = .04$ (see Fig. 1).

Congruent with Hypothesis 1, people expected others to judge men and women who helped in a gender-inconsistent way less favorably than men and women who helped in a gender-consistent way. Specifically, people expected others to judge *men helping in feminine ways* (i.e., gender-inconsistent helping; $M = 5.96$, $SD = 1.01$) less favorably than *men helping in masculine ways* (i.e., gender-consistent helping; $M = 6.19$, $SD = 0.87$), $p < .001$, $d = 0.24$, 95% CI [0.09, 0.40]. Likewise, people expected others to judge *women helping in masculine ways* (i.e., gender-inconsistent helping; $M = 6.13$, $SD = 0.87$) less favorably than *women helping in feminine ways* (i.e.,

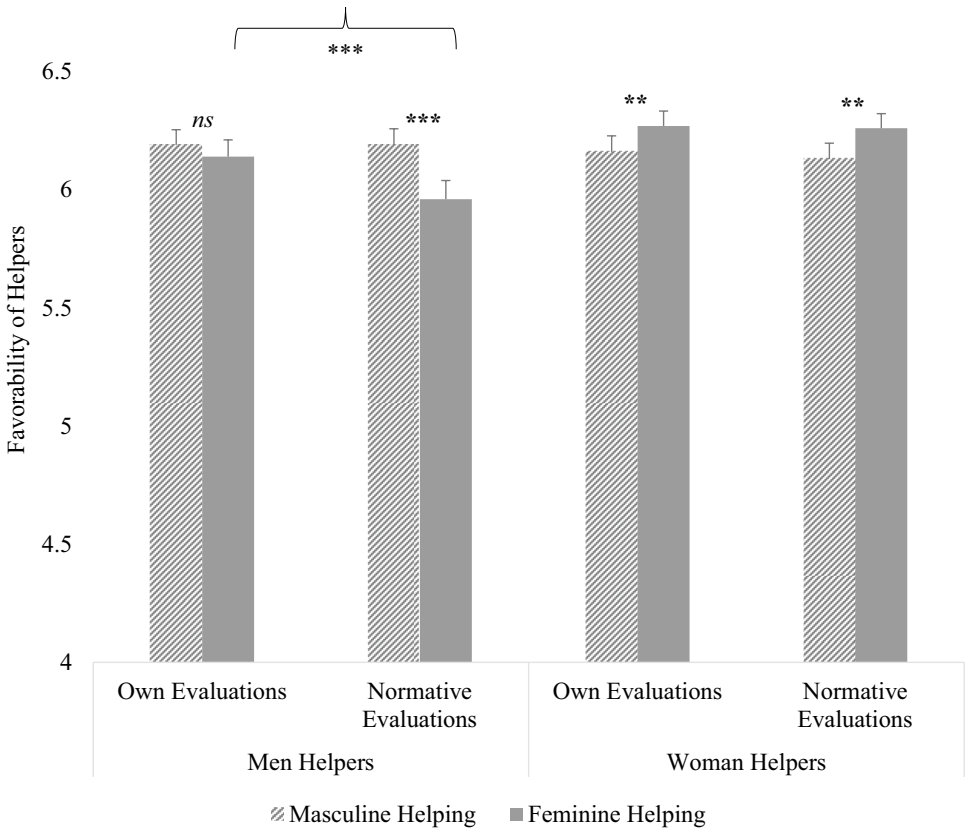

Fig. 1 Favorability Ratings of Helpers as a Function of Help Type, Helper Gender, and Rating Type in Study 3

Note. * $p < .05$. ** $p < .01$. *** $p < .001$

gender-consistent helping; $M = 6.26$, $SD = 0.79$), $p < .01$, $d = -0.16$, 95% CI [-0.31, < .01]. People's own evaluations of women helpers mirrored their normative evaluations in that they evaluated women who helped in masculine ways ($M = 6.16$, $SD = 0.88$) less favorably than women who helped in feminine ways ($M = 6.27$, $SD = 0.79$), $p < .01$, $d = -0.13$, 95% CI [-0.28, 0.02], but there were no significant differences in people's own evaluations of men helping in masculine ($M = 6.19$, $SD = 0.82$) and feminine ways ($M = 6.14$, $SD = 0.91$), $p = .15$, $d = 0.06$, 95% CI [-0.10, 0.21].

In line with Hypothesis 2, people expected others to judge gender-inconsistent helpers less favorably than they did in their own evaluations, but this relationship looked different by helper gender. Specifically, people expected others ($M = 5.96$, $SD = 1.01$) to judge m*en helping in feminine ways* less favorably than they did in their own evaluations ($M = 6.14$, $SD = 0.91$), $p < .01$, $d = 0.19$, 95% CI [0.03, 0.34]. However, there was no difference between normative ($M = 6.19$, $SD = 0.87$) and own evaluations ($M = 6.19$, $SD = 0.82$) of *men helping in masculine ways*, $p = .91$, $d < 0.01$, 95% CI [-0.15, 0.15]. There were also no differences in normative and own evaluations of w*omen helping in masculine ways*, $p = .50$, $d = 0.03$, 95% CI [-0.12, 0.19], or *women helping in feminine ways*, $p = .89$, $d = 0.01$, 95% CI [-0.14, 0.17] (see Table 7 for descriptive statistics).

The main effects of help type and rating type, as well as the helper gender x rating type interaction were not statistically significant (see Table 6 for a full summary of the results). For a summary of the significant help type x helper gender and help type x rating type interactions, see the project OSF page: https://osf.io/pruab/?view_only=df1e97a302fd48b8bd18980588352489.

Deservingness of Rewards

Material Rewards

There was a significant main effect of helper gender on deservingness of material rewards, $F(1, 163) = 10.78$, $p < .01$, $\eta_p^2 = .03$. Participants rated women helpers ($M = 3.78$, $SD = 2.03$) as more deserving of material rewards than men helpers ($M = 3.66$, $SD = 2.02$), $d = -0.06$, 95% CI [-0.21, 0.09]. There was also a statistically significant main effect of help type, $F(1, 163) = 44.05$, $p < .001$, $\eta_p^2 = .21$. Overall, people thought masculine helping ($M = 3.86$, $SD = 1.84$) was more deserving of material rewards than feminine helping ($M = 3.58$, $SD = 1.95$), $p < .001$, $d = 0.15$, 95% CI [-0.01, 0.30].

These main effects were qualified by a statistically significant help type x rating type interaction, $F(1, 163) = 5.10$, $p = .03$, $\eta_p^2 = .03$. Pairwise comparisons revealed that people rated masculine helping ($M = 3.88$, $SD = 1.92$) as more deserving of material rewards than feminine helping ($M = 3.66$, $SD = 2.05$) in their own evaluations, $p < .001$, $d = 0.11$, 95% CI [-0.04, 0.26]. A similar but slightly larger effect emerged for normative evaluations whereby people expected others to rate masculine helping ($M = 3.85$, $SD = 1.93$) as more deserving of material rewards than feminine helping ($M = 3.50$, $SD = 2.04$), $p < .001$, $d = 0.18$, 95% CI [0.02, 0.33]. There were no statistically significant differences in deservingness of material rewards for masculine helping in own vs. normative evaluations ($p = .77$) or feminine helping in own vs. normative evaluations ($p = .09$).

The main effect of rating type and the help type x helper gender and helper gender x rating type interactions were not statistically significant. In addition, the help type x helper gender x rating type interaction was also not statistically significant (see Table 6 for a full summary of the results). Thus, these results do not indicate that gender-consistent and gender-inconsistent helpers were evaluated any differently on deservingness of material rewards in own and normative evaluations.

Reputational Rewards

As with material rewards, there was a statistically significant help type x rating type interaction for reputational rewards, $F(1, 163) = 4.02$, $p = .05$, $\eta_p^2 = .02$. Despite the significant interaction, pairwise comparisons revealed that none of the simple effects were statistically significant. No main effects or additional interaction effects were statistically significant (see Table 6 for a full description of the results). In summary, these results indicate no differences in deservingness of reputational rewards for gender-consistent and gender-inconsistent helpers in own and normative evaluations.

Study 3 Discussion

As in previous studies, people expected men and women to perform gender-consistent helping more than gender-inconsistent helping. Specifically, people found masculine helping more expected when performed by men compared to women, and feminine helping more expected when performed by women compared to men. Although gender-inconsistent helping was less expected regardless of the helper gender, the magnitude of this difference was larger for men ($|d| = 0.99$) relative to women ($|d| = 0.64$), $z = 2.16$, $p = .03$, a finding congruent with theories of precarious manhood and asymmetry in gender role change. From these findings we can conclude that shifts in gender stereotypes may also carry over to gendered helping: people note the unexpected nature of women's gender-inconsistent helping, but they find men's gender-inconsistent helping even more unexpected.

Congruent with pre-registered hypotheses and Study 2 findings, participants generally expected others to judge gender-inconsistent helpers less favorably than gender-consistent helpers, and less favorably than they did themselves. In line with Hypothesis 1 and replicating Study 2, people

expected others to judge men and women who helped in a gender-inconsistent way less favorably than men and women who helped in a gender-consistent way. This finding supports the idea that people expect others to express backlash toward gender-inconsistent helpers, regardless of the helper gender. It appears that the assumption that others will penalize gender-inconsistent helpers is stronger for men helpers who violate gender roles ($|d|=0.24$) compared to women helpers who violate gender roles ($|d|=0.16$), $z=2.56$, $p=.01$.

While people did not privately judge gender-inconsistent men any differently than gender-consistent men, we unexpectedly observed that people did rate women helping in a masculine way less favorably than women helping in a feminine way, thus suggesting that there may in fact be backlash against women who help in gender-inconsistent ways. One factor explaining why people rated gender-inconsistent women (but not men) less favorably than gender-consistent helpers may stem from differences in the type of gender role violation depicted. The stereotypically masculine and agency-oriented helping behaviors used in the current study, like jumpstarting a car or carrying a box, might have represented a stronger deviation from gender roles for women than the stereotypically feminine and communal oriented helping behaviors, like offering moral support or making someone food, did for men. As illustrated by the role prioritization model (Haines & Stroessner, 2019), backlash for gender role violations occurs when people underprioritize gender typical roles—or put differently, overprioritize gender atypical roles—so it is possible people thought women helping in gender-inconsistent ways prioritized agentic behaviors to a higher degree than they thought gender-inconsistent men prioritized communal behaviors. On the other hand, gender stereotypes have changed over time, and in particular, women are believed to be more competent than in the past (Eagly et al., 2020). This may suggest that the explanation for our unexpected result is to be found elsewhere, and we leave the question for future research.

We found partial support for pre-registered Hypothesis 2. In line with our prediction and conceptually replicating Study 2 findings, people expected others to evaluate men helping in a gender-inconsistent way less favorably than they did themselves. Interestingly, this finding did not carry over to own and normative evaluations of gender-inconsistent women helpers, as we observed no significant differences in these ratings. It is plausible that these results emerged because there is a higher degree of pluralistic ignorance in judgements of men violating gender norms. Many cultures around the world share the belief that manhood is an earned social status that can be lost in the eyes of others, especially when men prioritize stereotypically feminine interests and roles (Vandello et al., 2008; for a review, see Bosson & Vandello, 2011). When men violate masculinity norms, they risk losing their status as a "real man" and are viewed by others as feminine and weak (Moss-Racusin & Johnson, 2016; Moss-Racusin et al., 2010). While individual people may not endorse or agree with these norms, they may expect other people to uphold them because there is usually a great deal of consensus surrounding gender stereotypes. Supporting the existence of pluralistic ignorance in endorsement of masculinity norms, in a recent study Munsch and colleagues (2018) found that working men expected others to endorse masculinity contest norms, like competitiveness and dominance, more than they did themselves. The current findings suggest this tendency to overestimate masculinity norms extends to evaluations of men violating gender roles in helping contexts. It will be important for future research to replicate these findings, as we observed this effect in a single study, and the prediction was not formally pre-registered.

Finally, despite observing that people expect others to judge gender-inconsistent helpers less favorably, these evaluations did not follow a similar pattern in ratings of helper deservingness of material and reputational rewards. In general, we found that people rated women as more deserving of material rewards than men, and targets engaging in masculine forms of helping were more deserving of rewards than those helping in feminine ways. However, these variables did not interact on ratings of deservingness of material or reputational rewards, suggesting no differences between gender-consistent and gender-inconsistent helpers. Taken together, these findings suggest that people might expect others to judge gender-inconsistent helpers less favorably than gender-consistent helpers—and in some cases less favorably than they do themselves—but these negative evaluations do not carry over to withholding tangible rewards.

General Discussion

The goal of the present research was to investigate evaluations of gender-consistent and gender-inconsistent helpers and examine the role of pluralistic ignorance in this process. Participants read about people helping in a gender-consistent or gender-inconsistent way, then provided their own evaluations (Studies 1–3) and their estimations for how other people in society would evaluate the helpers (Studies 2–3). Using a diverse set of measures, we evaluated perceptions of helpers on dimensions of performance, favorability, warmth and competence, and deservingness of reputational and material rewards. If pluralistic ignorance operates in situations of gendered helping, then we expected to find that people would expect others in society to evaluate gender-inconsistent helpers less favorably than gender-consistent helpers (Hypothesis 1), and less favorably than they did in their own evaluations (Hypothesis 2). Across three experimental studies, we found support for these predictions.

Congruent with the goals of our experiments and gender role stereotypes about helping (Croft et al., 2021; Eagly,

2009), gender-consistent forms of helping were more expected than gender-inconsistent forms of helping, in both own and normative evaluations, and regardless of helper gender. Furthermore, in Studies 2 and 3, participants generally expected others to evaluate gender-inconsistent helpers less favorably than gender-consistent helpers, supporting Hypothesis 1. Supporting Hypothesis 2, participants also expected others to evaluate gender-inconsistent helpers less favorably than they did themselves. However, Study 3 findings showed that the difference in own and normative evaluations of gender-inconsistent helpers relies somewhat on the gender of the helper; while people expected others to judge men helping in a gender-inconsistent way less favorably than they did themselves, these differences did not emerge in evaluations of gender-inconsistent women helpers. We did not predict significant differences in own evaluations of gender-consistent and gender-inconsistent helpers and found null effects for these differences in Studies 1 and 2. In Study 3 we unexpectedly found that own evaluations of women helpers mirrored normative evaluations, such that people personally rated gender-inconsistent women helpers less favorably than gender-consistent women helpers.

Despite finding evidence for a discrepancy in own and normative evaluations of favorability, warmth, and competence ratings, we failed to find similar evidence for differences in ratings of deservingness of rewards in Studies 1 and 3. In practice, our results showed relatively high rankings for deservingness of material and reputational rewards and demonstrated no significant differences between gender-consistent and gender-inconsistent helpers in either rating type. This lack of effect is particularly interesting in light of the reduced warmth ratings in normative evaluations of gender-inconsistent helpers, relative to gender-consistent helpers. Given that helping is itself a behavior that is likely to paint a picture of a helper as kind and generous (i.e., warm), the fact that this picture can be so easily marred by a gender stereotype violation is a testament to the magnitude of impact gendered expectations can have on how people are (assumed to be) judged by others. Importantly, however, this "warmth penalty" does not seem to extend to a helper's apparent deservingness of more tangible or intangible rewards for a service provided.

Our results suggest that gendered helping may be a context that evokes pluralistic ignorance; we found evidence that people think others will negatively judge helpers who help in gender-inconsistent ways (especially male helpers enacting female-stereotyped helping), but they do not judge these helpers negatively themselves. The fact that the positivity of the behavior seems to override the usual negativity of gender-role violations is consistent with recent work on question-asking, in which positive (generally: polite) question-asking behaviors were judged positively even when enacting these behaviors was counter to gender norms (Sandstrom et al., 2022). Further research could investigate the generalizability of this effect for other universally positive, female-stereotyped behaviors, and test whether there are universally positive, male-stereotyped behaviors for which women avoid social sanctions.

Although, at face value, helping seems unequivocally positive, various lines of research have uncovered factors that make the reality more complicated. The current research documents a new factor that adds to the complexity of helping: whether the helping behavior aligns with stereotypes for the helper's gender. Past research has examined cross-gender situations (i.e., where the gender of the helper is different to the gender of the help recipient) involving dependency-oriented help (NOTE: in the current research, we did not specify the gender of the help recipient). These studies have often used helping behaviors that have stereotypes aligned with the helper's gender (e.g., a man helping a woman with a math test, Shnabel et al., 2016; a woman helping a man with cleaning a burned pot; Bareket et al., 2021). It would be interesting to combine these two lines of inquiry, and manipulate both the gender-stereotypicality of the helping behavior and the gender of the help recipient. How would a male helper and a female help recipient be viewed if he helped her by cleaning a burned pot vs. showing her how to clean a burned pot? Would people expect the male helper to be judged even more negatively, since he is both providing dependency-oriented help, and helping in a gender-inconsistent way?

Our results provide support for the Gender Roles Inhibiting Prosociality (GRIP) model. This model draws on social role theory to understand the distal mechanisms by which gender stereotypes affect helping behavior, and the theory of planned behavior to understand the proximal mechanisms affecting individual-level decisions about whether to help in gendered contexts (Croft et al., 2021). Previous empirical research had provided support for the proximal mechanisms, finding that people's attitudes towards gendered helping, their perceptions of the norms about gendered helping, and their feelings of competence in carrying out gendered helping predicted their intentions to engage in gender-inconsistent help (Atkinson et al., 2021). The current paper provides support for the distal mechanisms, demonstrating that gender stereotypes do affect people's perceptions of the norms (i.e., their perceptions of how other people will judge those who help in gender-inconsistent ways). Future work could build on these findings by manipulating people's perceptions of these norms and exposing them to the (true) discrepancy between actual and normative perceptions of gender-consistent and -inconsistent helpers. Such exposure could shape helping interventions informed by the GRIP model (Croft et al., 2021).

Limitations and Future Research Directions

One alternative explanation for our findings is that, although people may not have had apprehensions about predicting that

others have biases, they may have been hesitant to admit their personal biases against gender-inconsistent helpers. We took steps to minimize this concern, as participant responses were confidential and anonymous. Previous research illustrates that people are more willing to report less socially desirable beliefs when they know their responses will be anonymous (Booth-Kewley et al., 1992; Gordon, 1987). We also found in Study 3 that participants expressed bias against gender-inconsistent women helpers in their own evaluations, suggesting they were comfortable expressing their personal biases.

To address the issue of social desirability, we included an own vs. normative IAT in Study 3, using the same images of helpers that we included in the explicit ratings. The results were largely (but not entirely) consistent with participants' explicit evaluations (see project OSF page): in their implicit ratings, people judged gender-consistent helpers more favorably than gender-inconsistent helpers and thought others would do the same (Hypothesis 1), though they did not expect others to judge gender-inconsistent helpers less favorably than they did personally (Hypothesis 2). Given the limited validation of the own/normative IAT, and the more general debate over the validity of using the IAT as a measure of a person's "true" feelings, we feel more research is needed in order to rule out the possibility of bias due to socially desirable responding.

The current findings are strengthened by our use of multiple types of everyday helping behaviors, increasing the generalizability of our results (though we used the same relatively small set of everyday, low-skill helping behaviors across studies). At the same time, it will be important for future research to investigate the extent to which pluralistic ignorance exists in evaluations of gender role violations in other helping contexts, like rescuing or protecting others in a high-risk emergency, or aiding in the workplace. Although we found that people expected others to judge gender-inconsistent helping less favorably, we might expect them to reward gender-inconsistent helping in other contexts. Illustrating a different form of bias than studied in the current research, perceivers occasionally reward gender role incongruent behavior, especially when the behavior is particularly unexpected or when they set lower standards for an actor's success (Bettencourt et al., 1997; Biernat & Manis, 1994). For example, when women aided in a high-risk emergency scenario, they were evaluated more favorably than identically described men (Taynor & Deaux, 1973). Additionally, men who provided social support for their colleagues received more favorable evaluations than similarly described women (Heilman & Chen, 2005). It will be important for future research to expand on the types of helping and the degree to which helpers violate gender roles, as this may not only shift the evaluation of their behavior, but also the direction of the discrepancy between their own and normative evaluations.

Using vignettes in the current study had several advantages, such as providing the means to carefully manipulate different factors and allowing us to test a variety of helping behaviors. However, judgments of hypothetical helpers are likely to differ in many ways from judgments of helpers who are seen actually performing those helping behaviors. For example, an observer could also infer how comfortable or uncomfortable the helper felt performing the behavior, and to what extent the help recipient appeared to be grateful for the help. More work is needed, with increasing ecological validity, to learn how these findings translate to real-world helping.

Practice Implications

One factor undermining interest and intentions to perform gender-inconsistent helping stems from concerns that other people will judge this behavior negatively (Atkinson et al., 2021; Croft et al., 2021). This makes sense, given that humans are social beings and desire to be connected with and approved by others (Baumeister & Leary, 1995). The current findings suggest that this concern for penalties may be overblown, and people generally expect others to judge these helpers less favorably than they *actually* do. It is possible that exposing people to information that others will not judge them as negatively as they think when they violate gender roles could ease some of their hesitations about engaging in these behaviors. We hope the current findings will inspire future research on this and other possible interventions to encourage more gender-inconsistent helping. After all, if people are restricting the ways in which they help, then they are being less helpful than they might be.

Finding ways to encourage more gender-inconsistent helping has the potential to shift gender stereotypes and occupational role segregation more broadly (for a review, see Croft et al., 2021). For instance, if people could feel less worried about negative judgments from others, they might be more interested in joining helping-related fields that are heavily segregated, like nursing or firefighting. These kinds of prosocial, care-oriented careers remain heavily gender segregated (Croft et al., 2015). As people engage in gender-inconsistent helping roles more frequently, this could also increase their comfort and self-efficacy to behave counter-stereotypically in a larger variety of contexts regulated by gendered expectations. Furthermore, frequent observations of people engaging in gender-inconsistent helping roles could shift gender-stereotypic assumptions about the attributes and roles ascribed to people based on their group membership.

Conclusion

Prosocial behavior is yet another domain regulated by traditional gender roles, and these role restrictions perpetuate traditional gender role expectations. Our research finds that

people expect others to judge gender-inconsistent helpers less favorably than they actually do, and this overestimation for penalties could lead people to avoid gender-inconsistent helping opportunities, perpetuating a cycle of gender-segregated prosociality. On a more positive note, our research also highlights that this fear is unfounded; people may fear they will face penalties for helping in a gender-inconsistent way but concerns about this backlash are exaggerated. If we can correct people's inaccurate perceptions of others' beliefs, we might bring about more gender-inconsistent helping.

Appendix

Study 3: Four Sample Helping Vignettes

Directions: In this study, we are interested in learning more about people's perceptions of prosocial behavior (i.e., people helping people). We will ask you to read hypothetical situations describing a person helping another individual. After each scenario you will be asked to answer questions regarding how you perceive the helper. There are no right or wrong answers to these questions. We are merely interested in your perceptions. [*Each vignette seen/rated separately.*]

Gender-Consistent Helping

Imagine you observed a **man** helping someone by jump-starting their car

Imagine you observed a **woman** taking a home-cooked meal to a sick friend or neighbor

Gender-Inconsistent Helping

Imagine you observed a **woman** helping someone by jump-starting their car

Imagine you observed a **man** taking a home-cooked meal to a sick friend or neighbor

Funding This work was funded by a Visiting Fellowship from the British Academy [number: VF1\102397] awarded to Alyssa Croft.

Compliance with Ethical Standards

Conflicts of Interest The authors declare no conflicts of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Atkinson, C., Buie, H., Sandstrom, G., Aknin, L. B., & Croft, A. (2021). Testing the GRIP: An empirical examination of the gender roles inhibiting prosociality model. *Sex Roles, 85*, 440–462. https://doi.org/10.1007/s11199-021-01229-2

Bareket, O., Shnabel, N., Kende, A., Knab, N., & Bar-Anan, Y. (2021). Need some help, honey? Dependency-oriented helping relations between women and men in the domestic sphere. *Journal of Personality and Social Psychology, 120*(5), 1175–1203. https://doi.org/10.1037/pspi0000292

Baumeister, R. F., & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin, 117*(3), 497–529. https://doi.org/10.1037/0033-2909.117.3.497

Becker, J. C., Glick, P., Ilic, M., & Bohner, G. (2011). Damned if she does, damned if she doesn't: Consequences of accepting versus confronting patronizing help for the female target and male actor. *European Journal of Social Psychology, 41*(6), 761–773. https://doi.org/10.1002/ejsp.823

Becker, S. W., & Eagly, A. H. (2004). The heroism of women and men. *American Psychologist, 59*(3), 163–178. https://doi.org/10.1037/0003-066X.59.3.163

Bettencourt, B. A., Dill, K. E., Greathouse, S. A., Charlton, K., & Mulholland, A. (1997). Evaluations of ingroup and outgroup members: The role of category-based expectancy violation. *Journal of Experimental Social Psychology, 33*(3), 244–275. https://doi.org/10.1006/jesp.1996.1323

Biernat, M., & Manis, M. (1994). Shifting standards and stereotype-based judgments. *Journal of Personality and Social Psychology, 66*(1), 5–20. https://doi.org/10.1037/0022-3514.66.1.5

Bosson, J. K., & Vandello, J. A. (2011). Precarious manhood and its links to action and aggression. *Current Directions in Psychological Science, 20*, 82–86. https://doi.org/10.1006/jesp.1996.1323

Booth-Kewley, S., Edwards, J. E., & Rosenfeld, P. (1992). Impression management, social desirability, and computer administration of attitude questionnaires: Does the computer make a difference? *Journal of Applied Psychology, 77*, 562–566. https://doi.org/10.1037/0021-9010.77.4.562

Brescoll, V. L., & Uhlmann, E. L. (2005). Attitudes toward traditional and nontraditional parents. *Psychology of Women Quarterly, 29*(4), 436–445. https://doi.org/10.1111/j.1471-6402.2005.00244.x

Burleson, B. R., & Kunkel, A. W. (2006). Revisiting the different cultures thesis: An assessment of sex differences and similarities in supportive communication. In K. Dindia & D. J. Canary (Eds.), *Sex differences and similarities in communication* (pp. 137–159). Lawrence Erlbaum Associates Publishers.

Cancian, F. M., & Oliker, S. J. (2000). *Caring and gender*. Rowman & Littlefield.

Croft, A., Atkinson, C., Sandstrom, G. M., Orbell, S., & Aknin, L. B. (2021). Loosening the GRIP (Gender Roles Inhibiting Prosociality) to promote gender equality. *Personality and Social Psychology Review, 25*(1), 66–92. https://doi.org/10.1177/1088868320964615

Croft, A., Schmader, T., & Block, K. (2015). An underexamined inequality: Cultural and psychological barriers to men's engagement with communal roles. *Personality and Social Psychology Review, 19*, 343–370. https://doi.org/10.1177/1088868314564789

Cuddy, A. J., Fiske, S. T., & Glick, P. (2007). The BIAS map: Behaviors from intergroup affect and stereotypes. *Journal of Personality and Social Psychology, 92*(4), 631–648. https://doi.org/10.1037/0022-3514.92.4.631

Eagly, A. H. (2009). The his and hers of prosocial behavior: An examination of the social psychology of gender. *American Psychologist, 64*(8), 644–658. https://doi.org/10.1037/0003-066X.64.8.644

Eagly, A. H., & Crowley, M. (1986). Gender and helping behavior: A meta-analytic review of the social psychological literature. *Psychological Bulletin, 100*(3), 283–308. https://doi.org/10.1037/0033-2909.100.3.283

Eagly, A. H., Johansen-Schmitt, M. C., & van Engen, M. L. (2003). Transformational, transactional, and laissez-faire leadership styles: A meta-analysis comparing women and men. *Psychological Bulletin, 129*(4), 569–591. https://doi.org/10.1037/0033-2909.129.4.569

Eagly, A. H., Nater, C., Miller, D. I., Kaufmann, M., & Sczesny, S. (2020). Gender stereotypes have changed: A cross-temporal meta-analysis of US public opinion polls from 1946 to 2018. *American Psychologist, 75*(3), 301. https://doi.org/10.1037/amp0000494

Farrell, S. K., & Finkelstein, L. M. (2007). Organizational citizenship behavior and gender: Expectations and attributions for performance. *North American Journal of Psychology, 9*(1), 81–95. https://psycnet.apa.org/record/2007-05078-006

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*(2), 175–191. https://doi.org/10.3758/bf03193146

Fisher, J. D., Nadler, A., & Whitcher-Alagna, S. (1982). Recipient reactions to aid. *Psychological Bulletin, 91*(1), 27. https://doi.org/10.1037/0033-2909.91.1.27

Glick, P., & Fiske, S. T. (1997). Hostile and benevolent sexism: Measuring ambivalent sexist attitudes toward women. *Psychology of Women Quarterly, 21*(1), 119–135. https://doi.org/10.1111/j.1471-6402.1997.tb00104.x

Gordon, R. A. (1987). Social desirability bias: A demonstration and technique for its reduction. *Teaching of Psychology, 14*(1), 40–42. https://doi.org/10.1207/s15328023top1401_11

Haines, E. L., & Stroessner, S. J. (2019). The role prioritization model: How communcal men and agentic women can (sometimes) have it all. *Social and Personality Psychology Compass, 13*(12), 1751–9004. https://doi.org/10.1111/spc3.12504

Heilman, M. E., & Chen, J. J. (2005). Same behavior, different consequences: Reactions to men's and women's altruistic citizenship behavior. *Journal of Applied Psychology, 90*(3), 431–441. https://doi.org/10.1037/0021-9010.90.3.431

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*(4), 664–667. https://doi.org/10.1016/j.jesp.2010.01.008

Klein, N., & Epley, N. (2014). The topography of generosity: Asymmetric evaluations of prosocial actions. *Journal of Experimental*

Psychology: General, 143(6), 2366–2379. https://doi.org/10.1037/xge0000025

Klein, N., Grossman, I., Uskul, A. K., Kraus, A., & Epley, N. (2015). It pays to be nice, but not really nice: Asymmetric evaluations of prosociality across seven cultures. *Judgment and Decision Making*, *10*, 355–364. https://www.proquest.com/docview/1701178510

Lamy, L., Fischer-Lokou, J., & Guéguen, N. (2009). Induced reminiscence of love and chivalrous helping. *Current Psychology: A Journal for Diverse Perspectives on Diverse Psychological Issues, 28*(3), 202–209. https://doi.org/10.1007/s12144-009-9059-9

Latané, B., & Darley, J. M. (1970). *The unresponsive bystander: Why doesn't he help?* Appleton-Century-Crofts.

Koenig, A. M., & Eagly, A. H. (2014). Evidence for the social role theory of stereotype content: Observations of groups' roles shape stereotypes. *Journal of Personality and Social Psychology, 107*(3), 371–392. https://doi.org/10.1037/a0037215

Moss-Racusin, C. A., & Johnson, E. R. (2016). Backlash against male elementary educators. *Journal of Applied Social Psychology, 46*(7), 379–393. https://doi.org/10.1111/jasp.12366

Moss-Racusin, C. A., Phelan, J. E., & Rudman, L. A. (2010). When men break the gender rules: Status incongruity and backlash against modest men. *Psychology of Men & Masculinity, 11*(2), 140–151. https://doi.org/10.1037/a0018093

Munsch, C. L., Weaver, J. R., Bosson, J. K., & O'Connor, L. T. (2018). Everybody but me: Pluralistic ignorance and the masculinity contest. *Journal of Social Issues, 74*(3), 551–578. https://doi.org/10.1111/josi.12282

Nadler, A., & Fisher, J. D. (1986). The role of threat to self-esteem and perceived control in recipient reaction to help: Theory development and empirical validation. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 19, pp. 81–122). Academic Press. https://doi.org/10.1016/S0065-2601(08)60213-0

Nadler, A., & Halabi, S. (2015). Helping relations and inequality between individuals and groups. In M. Mikulincer, P. R. Shaver, J. F. Dovidio, & J. A. Simpson (Eds.) *APA handbook of personality and social psychology, (*Vol. 2, pp. 371–393). American Psychological Association. https://doi.org/10.1037/14342-014

Okimoto, T. G., & Heilman, M. E. (2012). The "bad parent" assumption: How gender stereotypes affect reactions to working mothers. *Journal of Social Issues, 68*(4), 704–724. https://doi.org/10.1111/j.1540-4560.2012.01772.x

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*(4), 406–413. https://doi.org/10.1111/j.1471-6402.2008.00454.x

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*(4), 269–281. https://doi.org/10.1111/1471-6402.t01-1-00066

Prentice, D. A., & Miller, D. T. (1993). Pluralistic ignorance and alcohol use on campus: Some consequences of misperceiving the social norm. *Journal of Personality and Social Psychology, 64*(2), 243–256. https://doi.org/10.1037/0022-3514.64.2.243

Prentice, D. A., & Miller, D. T. (1996). Pluralistic ignorance and the perpetuation of social norms by unwitting actors. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 28, pp. 161–209). Academic Press. https://doi.org/10.1016/S0065-2601(08)60238-5

Roter, D. L., Hall, J. A., & Aoki, Y. (2002). Physician gender effects in medical communication: A meta-analytic review. *JAMA, 288*(6), 756–764. https://doi.org/10.1001/jama.288.6.756

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*(2), 157–176. https://doi.org/10.1037/0022-3514.87.2.157

Rudman, L. A., & Mescher, K. (2013). Penalizing men who request a family leave: Is flexibility stigma a femininity stigma? *Journal of Social Issues, 69*(2), 322–340. https://doi.org/10.1111/josi.12017

Rudman, L. A., Mescher, K., & Moss-Racusin, C. A. (2013). Reactions to gender egalitarian men: Perceived feminization due to stigma-by-association. *Group Processes & Intergroup Relations, 16*(5), 572–599. https://doi.org/10.1177/1368430212461160

Ruiz, A. G. (2019). White knighting: How help reinforces gender differences between men and women. *Sex Roles, 81*(9–10), 529–547. https://doi.org/10.1007/s11199-019-01018-y

Sandstrom, G.M., Croft, A., Gibson, H., Carter, A. J. (2022). People draw on gender stereotypes to judge question-askers, but there is no such thing as a gender-stereotypic question [Manuscript submitted for publication]. Department of Psychology, University of Sussex. Access preprint: https://psyarxiv.com/7eq8j/

Sandstrom, G. M., Schmader, T., Croft, A., & Kwok, N. (2019). A social identity threat perspective on being the target of generosity from a higher status other. *Journal of Experimental Social Psychology, 82*, 98–114. https://doi.org/10.1016/j.jesp.2018.12.004

Shnabel, N., Bar-Anan, Y., Kende, A., Bareket, O., & Lazar, Y. (2016). Help to perpetuate traditional gender roles: Benevolent sexism increases engagement in dependency-oriented cross-gender helping. *Journal of Personality and Social Psychology, 110*(1), 55–75. https://doi.org/10.1037/pspi0000037

Taynor, J., & Deaux, K. (1973). When women are more deserving than men: Equity, attribution, and perceived sex differences. *Journal of Personality and Social Psychology, 28*(3), 360–367. https://doi.org/10.1037/h0035118

Van Grootel, S., Van Laar, C., Meeussen, L., Schmader, T., & Sczesny, S. (2018). Uncovering pluralistic ignorance to change men's communal self-descriptions, attitudes, and behavioral intentions. *Frontiers in Psychology, 9*, 1344. https://doi.org/10.3389/fpsyg.2018.01344

Vandello, J. A., Bosson, J. K., Cohen, D., Burnaford, R. M., & Weaver, J. R. (2008). Precarious manhood. *Journal of Personality and Social Psychology, 95*(6), 1325–1339. https://doi.org/10.1037/a0012453

Willer, R. (2009). Groups reward individual sacrifice: The status solution to the collective action problem. *American Sociological Review, 74*, 23–43. https://doi.org/10.1177/000312240907400102

Willer, R., Feinberg, M., Irwin, K., Schultz, M., & Simpson, B. (2010). The trouble with invisible men: How reputational concerns motivate generosity. In S. Hitlin & S. Vaisey (Eds.), *Handbook of the sociology of morality* (pp. 315–330). Springer. https://doi.org/10.1007/978-1-4419-6896-8_17

Wood, W., & Eagly, A. H. (2012). Biosocial construction of sex differences and similarities in behavior. In M. Olson, James & P. Zanna, Mark (Eds.), *Advances in experimental social psychology* (Vol. 46, pp. 55–123). Burlington: Academic Press. https://doi.org/10.1016/B978-0-12-394281-4.00002-7

Yoder, J., Hogue, M., Newman, R., Metz, L., & LaVigne, T. (2002). Exploring the moderators of gender differences: Contextual differences in door-holding behavior. *Journal of Applied Social Psychology, 32*, 1682–1686. https://doi.org/10.1111/j.1559-1816.2002.tb02769.x

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.