Advancing Stereotyping Research: How and Why to Use Linear Mixed-Effects Models in Gender Stereotyping Research

Abigail M. Folberg

University of Kentucky

Markus Brauer

University of Wisconsin-Madison

Carey S. Ryan

University of Nebraska-Omaha

Jennifer S. Hunt

University of Kentucky

In Press: ***Testing, Psychometrics, and Methodology in Applied Psychology***

Correspondence should be addressed to Abigail M. Folberg, Departments of Gender and Women's Studies and Psychology, University of Kentucky, 106B Kastle Hall, Lexington, KY 40506; email: abby.folberg@uky.edu; phone: 248.797.9965. This research was funded by an internal grant from the University of Nebraska-Omaha and a Grant-in-Aid from the Society for the Psychological Study of Social Issues.

Abstract

We present two studies examining the valence of gender stereotypes using linear mixed-effects models (LMEMs) to demonstrate how they can advance stereotyping research. Although LMEMs are common in some domains of psychology (e.g., developmental and cognitive psychology), they are much less commonly used in research on stereotyping. And yet, concerns about the generalizability of results, inflation of type I errors, and ease of handling missing data are equally relevant to work on stereotyping. In this paper, we first summarize what LMEMs are and how they might be applied to work on stereotyping. We then show how LMEMs can be used to analyze data from studies with researcher-generated attributes (Study 1) and participant-generated attributes (Study 2). We also show that LMEMs are particularly appropriate for designs that employ planned missingness (Study 2). Finally, we discuss how LMEMs may allow researchers to resolve conflicting findings in gender stereotype research and how designs with planned missingness allow researchers more flexibility in answering their research questions.

Keywords: Linear Mixed-Effects Models; Random effects models; Crossed Random Effects Models; Stereotypes; Gender Stereotypes

Word Count: 150

Advancing Stereotyping Research: How and Why to Use Linear Mixed-Effects Models in Gender Stereotyping Research

Stereotypes, that is, beliefs about the characteristics of social groups, can impact judgments of group members and lead to discriminatory behavior. As a result, a great deal of research has examined the content of gender and racial stereotypes and their associations with role expectations, discrimination, and prejudice. Researchers who examine stereotypes use a variety of techniques to assess them. Participants may rate stimuli such as faces (Judd, Garcia-Marques, & Yzerbyt, 2019) or attributes (Abele, Hauke, Peters, Louvet, Szymkow, & Duan, 2016; Diekman & Eagly, 2000; Eagly, Nater, Wood, Miller, Kaufmann, & Sczesny, 2019; Glick et al., 2004; Haines, Deaux, & Lotharo, 2016; Hentschel, Heilman & Peus, 2019) on dimensions such as valence (i.e., positivity vs. negativity) and stereotypicality (i.e., stereotypic vs. counterstereotypic attributes). In other research, participants generate their own attributes to describe target groups and then provide ratings of those attributes with respect to valence (Glick, Diebold, Bailey-Werner, & Zhu 1997; Glick et al., 2004).

Regardless of the method used to assess stereotypes, researchers have generally averaged across attributes or other stimuli and analyzed the data using ordinary least squares (OLS) regression. However, OLS analyses limit the types of conclusions that researchers can draw and may even result in misleading conclusions. In OLS analyses, stimuli are treated as fixed, preventing the generalization of results to other stimuli from the same population of stimuli (Brauer & Curtin, 2018; Judd, Westfall, & Kenny, 2012, 2017). In the case of attributes, it means that conclusions are limited to the particular attributes that the researchers selected or that participants generated and cannot be generalized to other stereotype-relevant attributes.

In addition, OLS analyses confound variation due to differences across individual attributes or other stimuli, inflating the values of F and leading to a type I error rate well above 5% (Judd et al., 2012, 2017; Locker, Hoffman, & Bovaird, 2007). Further, traditional OLS analyses, such as repeated measures analysis of variance (ANOVA), are ill-equipped to handle missing data. They are thus not suitable for planned missingness designs. In such designs, participants are randomly assigned to respond to a random

subsample of stimuli, as a means of examining a large number of stimuli without biasing parameter estimates while also minimizing participant attrition (Brauer & Curtin, 2018; Prokopek, 2011). All three of these issues can be resolved using linear mixed-effects models (LMEMs) in which both participants and stimuli (in our case, attributes) are treated as random variables.

In this paper, we argue that LMEM analysis can expand research on gender and other types of stereotyping by enabling a greater range of questions and research designs, as well as facilitating greater generalizability of conclusions. We first explain what LMEMs are, how they might apply to research on stereotyping, and elaborate on the reasons that researchers in this area might want to use LMEMs rather than conventional OLS analyses. We then illustrate how LMEMs can be used to analyze data from two studies on gender stereotypes, one using a student sample and predetermined set of attributes (Study 1) and one using a crowd-sourced sample and participant-generated attributes (Study 2). Finally, we offer suggestions for how LMEMs analyses may help generate and answer new questions in stereotyping research.

Accounting for Non-Independence in Research Designs

Linear mixed-effects models (LMEMs), also known as random effects models, multilevel models, and hierarchical linear models, are often used to address non-independence in data. (For an accessible tutorial to LMEMs see Brauer & Curtin, 2018.) Traditionally, researchers have viewed non-independence as arising in two types of designs: nested, in which each participant provides multiple observations within conditions or groups, and crossed, in which each participant provides multiple observations across conditions or groups (Judd, McClelland & Ryan, 2017). However, as we later demonstrate, many research designs do not fit neatly into either category. Thus, conceptualizing research designs along only these two dimensions may limit the hypotheses that can be tested and the conclusions that may be drawn (Judd et al., 2017).

In nested designs, non-independence occurs when the error terms of observations within higher-order units are correlated, for example, when the responses of individuals (e.g., students) within conditions (e.g., classrooms) are more similar (or dissimilar in the case of negative non-independence) to

each other than to the error terms of observations in other groups (Judd et al., 2017). In crossed designs, non-independence arises when there are multiple observations per participant (or group), resulting in correlated error terms for each participant. For example, researchers examining stereotyping might ask participants to rate a series of male and female targets (e.g., Judd et al., 2019). The data are non-independent because responses for a given participant are more similar (or dissimilar) to each other than to responses generated by other participants.

The application of LMEMs to nested designs is common and is often referred to as hierarchical linear modeling (HLM) or multilevel modeling (MLM) (Raudenbush & Byrk, 2002; Snijders & Bosker, 1999). For example, developmental psychologists may examine changes in children's peer relations over the school year. Thus, time might be nested within students, who are nested within classrooms. Fewer researchers have applied LMEMs to crossed designs and other complex designs, although experts have recommended their use (Baayen, Davidson, & Bates, 2008; Judd et al., 2012, 2017; Locker et al., 2007). For example, studies on psycholinguistics have used LMEMs rather than repeated measures ANOVA to analyze data, treating participants and stimuli as random variables (Baayen et al., 2008; Locker et al., 2007).

LMEMs are equally applicable to data on stereotyping. Indeed, just as words in a psycholinguistics study might be considered a sample of words from a population (i.e., a language) to which researchers wish to generalize (Clark, 1973), so, too, might attributes be considered a sample of attributes from a population of attributes to which researchers wish to generalize. Alternatively, researchers might ask participants to rate target individuals on a particular dimension, and each target individual belongs to one of two groups (e.g., women and men, or Whites and Blacks). Here the researchers' goal is likely to generalize their findings to the entire population of people from which the target individuals were drawn.

Applications of LMEMs to Stereotyping Research

We reference work on gender stereotypes and stereotype change to illustrate how LMEMs might address conflicting research findings. In most of these studies, participants were asked to rate target men

and women on a predetermined list of attributes (Diekman & Eagly, 2000; Eagly et al., 2019; Glick et al., 2004; Haines, et al., 2016; Hentschel et al., 2019). However, others have used participant-generated attributes (Glick et al., 1997, 2004), and the techniques that we review in this article are equally applicable to both types of studies.

Work on gender stereotypes suggests that individuals make inferences about the types of attributes women and men have based on the gender representation in occupational and domestic roles (Diekman & Eagly, 2000; Eagly, Wood & Diekman., 2000; Koenig & Eagly., 2014). Men tend to be better represented in roles requiring displays of physical or social power (e.g., CEO, leader); thus, they are often perceived as agentic (e.g., assertive, independent, competent). In contrast, women are often perceived as communal (e.g., warm, nurturing, moral) because they tend to be better represented in domestic and occupational roles that require caretaking (e.g., stay-at-home mom, nurse, educator). Research suggesting that role information informs stereotypes has been applied to other groups as well (Koenig & Eagly, 2014), including sexual minorities (e.g., Pellegrini, De Cristofaro, Giacomantonio, & Salvati, 2020).

The valence of gender stereotypes also varies across women and men. Researchers have consistently found that women are viewed more positively than are men (Eagly & Mlidanic, 1994; Glick et al., 2004; Sullivan, Moss-Racusin, Lopez, & Williams, 2018), an effect dubbed the "women are wonderful" effect. However, positive perceptions of women generally only occur for perceptions of warmth, or communion and often come at the expense of perceptions of competence (e.g., Fiske, Cuddy, Glick & Xu, 2002; Glick & Fiske, 1996, but see Eagly et al., 2019). Thus, although women may be perceived positively, particularly when they conform to gender stereotypes, they are also often regarded as fragile and weak (Glick & Fiske, 1996).

These gender stereotypes maintain and perpetuate gender inequality. For example, women are less likely than men to be selected for leadership positions (Eagly & Karau, 2002) or careers that are perceived as masculine (Heilman, Wallen, Fuchs, & Tamkins, 2004) because they are perceived as less competent than men. Similarly, the belief that communal qualities (e.g., working with others, being

people-oriented) are incompatible with careers in science, technology, engineering and mathematics (STEM), may deter women from pursuing STEM careers (Diekman, Steinberg, Brown, Belanger, & Clark, 2017). Further, women are often punished for violating gender role expectations (Glick & Fiske, 1996). For example, women, but not men, displaying emotions associated with power (such as anger; Brescoll & Uhlmann, 2008) or behavior that is perceived as self-promotional or assertive (Rudman & Glick, 2001) are viewed as less hirable or lower status.

As stereotypes may inform prejudice and discrimination, some researchers have been interested in whether (and how) they change. Given the above-mentioned impact of gender representation in occupational and domestic roles, it might be expected that stereotypes of women and men become more similar as societies become more egalitarian and women and men occupy similar roles. However, work on stereotype change has yielded mixed results. Meta-analytic (Eagly et al., 2019) and experimental (Diekman & Eagly, 2000) evidence suggests women may be perceived as gaining in agency commensurate with their increased participation in the paid labor force. In contrast, other studies suggest that gender stereotypes are relatively stable over time (Haines et al., 2016), and that they lag significantly behind cultural change (Diekman, Eagly, & Johnston, 2010).

Research on stereotype change has generally concentrated on the extent to which women have gained in agency as they increasingly enter traditionally masculine roles (Diekman & Eagly, 2000; Eagly et al., 2019). However, it is unclear whether there even are gender differences in agency. Some recent studies still find evidence of gender differences in agency (Eagly et al., 2019), whereas others do not (Hentschel et al., 2019). Closely related work on gender-role congruent goals suggests that gender differences do not reliably emerge in goals associated with agency (Diekman et al., 2017; Folberg, Kercher, & Ryan, 2019). Some of these discrepancies may be due to the ways in which stereotypes and goals are conceptualized and modeled (Folberg et al., 2019), as agency and communion may comprise multiple dimensions (Abele et al., 2016; Eagly et al., 2019; Hentschel et al., 2019) or bifactor structures (Folberg et al., 2019).

Most importantly, we argue that treating stereotypic attributes as a fixed variable may also account for some of these apparently contradictory findings. For example, Hentschel et al. (2019) and Eagly et al. (2019) used different sets of items to assess agency and came to different conclusions regarding gender differences in perceptions of agency. Similarly, different items have been used in studies assessing stereotype change (e.g., Haines et al., 2016; Diekman & Eagly, 2000; Eagly et al., 2019). Thus, sampling error due to the specific attribute items used in individual studies may account for differences across studies. Indeed, with the exception of Eagly et al. (2019), who used meta-analytic techniques to assess stereotype change, all other studies used OLS analyses treating attributes as fixed, which prevents conclusions from being generalized to the population of attributes from which they were chosen (Brauer & Curtin, 2018; Clark, 1973; Judd et al., 2012; Judd et al., 2017;).

Benefits of LMEMs, Effect Size, and Power, and Missing Data

LMEMs allow researchers to treat attributes and participants as random variables. As a result, researchers can generalize from their sample of attributes to the entire population of attributes, and from their sample of participants to the population of individuals, as long as the samples were drawn randomly from the populations to which the researchers wish to generalize. LMEMs also provide less biased estimates of *F* and handle missing data much better than do OLS analyses (Brauer & Curtin, 2018). Thus, the use of LMEMs may increase the generalizability of results and accuracy of conclusions for any field of research in which researchers sample stimuli (e.g., words, pictures, attributes) in addition to participants. However, despite some of the benefits of LMEMs, their complexity, relative to OLS analyses, may also present some challenges.

In the sections below, we discuss measures of effect size and power. We close with a discussion of planned missingness designs, which LMEMs can better accommodate than can OLS analyses. Mathematical derivations of estimates of effect size and power are beyond the scope of this article. However, Rights and Sterba (2018, 2019) provide useful discussions of effect size, and Judd et al. (2017) and Westfall, Kenny, and Judd (2014) provide useful and accessible discussions of both effect size and power.

Effect size. Measures of effect size, such as Cohen's *d* or *partial* η^2, are easy to estimate in OLS analyses, and can be computed in most, if not all, statistical software packages. In contrast, researchers disagree on what constitutes an appropriate measure of effect size for LMEM models. Some researchers have calculated measures of effect size for designs with categorical predictors (Nakagawa, Johnson, & Schielzeth, 2017; Judd et al., 2017). For example, Judd et al. (2017) provide an analogue of Cohen's *d* for designs with crossed random variables, although it is limited to designs with one dichotomous predictor.

Others have expressed interest in developing measures of effect size that can be more widely used and easily communicated across a number of different designs, as designs with crossed random variables may also include continuous measures (e.g., Kurebayashi, Hoffmann, Ryan, & Mayamura, 2012). Thus, measures, such as r^2 may be useful in helping researchers communicate effect sizes in LMEMs. Seyla, Dierker, Hedeker, and Mermelstein (2012) developed an analog of Cohen's f^2 for use in crossed designs, although they acknowledge that Cohen's f^2 may not be widely used in all fields. In addition, Rights and Sterba (2018, 2019) provide a very flexible framework for calculating different measures of r^2 for two-level nested designs. They also developed a macro for R (R2MLM) that allows researchers to calculate these measures (Rights & Sterba, 2018). However, as of yet, the framework has not been extended to designs with crossed random variables.

In the present paper, we use "pseudo r^2," an effect-size indicator proposed by Page-Gould, Sharples, and Song (2019), who conducted simulation studies suggesting that converting a value of *t* (or its squared equivalent, *F*) to a value of *r* or r^2 yields a meaningful estimate of effect size. Thus, the pseudo r^2 is simply the *F*-value divided by the sum of the *F*-value and the error (denominator) degrees of freedom. We advocate for this method for a few reasons. First, it is an intuitive measure of effect size that can be used across a number of different designs. Second, it is easy to estimate and easy to understand for researchers accustomed to OLS analyses. Third, some measures of effect sizes used in LMEMs represent the percentage of the total outcome variance accounted for by a predictor of interest (Rights & Sterba, 2018, 2019). However, just like with Cohen's *d* or *partial* η^2 in OLS analyses, we are

generally more interested in the proportional reduction in error accounted for by a specific predictor (or set of predictors), which can be derived from values of *t* or *F* (Judd, et al., 2017).

Statistical power. Researchers may also be interested in how to conduct power analyses in crossed and other complex designs. For a specific hypothesis or model comparison of interest, power is an estimate of the probability of finding an effect of a specific size, assuming it exists, given a specific alpha level for a given sample size (Cohen, 1988). In OLS analyses, power analysis is fairly straightforward and can be conducted easily in most statistical software programs. For a specific hypothesis, if researchers know the estimate of power, effect size, and alpha value, they can determine the sample size necessary to find an effect of interest, assuming it exists. Similarly, if researchers know the sample size, effect size, and alpha value, they can estimate power. Since effect size is relevant for power analysis, some of the points raised in the previous section on effect sizes are also relevant for power analysis. Different effect size indicators will yield different values of statistical power.

Power in the context of designs with crossed random factors has not been as widely examined as it has in nested designs (Westfall et al., 2014). Some programs are available for estimating power in designs with crossed random variables. For example, Westfall et al. created a Shiny app for researchers who want to estimate power for designs with two crossed random variables and one fixed predictor with two conditions. Others have recommended the use of simulation studies, which may be conducted in programs, such as R or Mplus (Brysbaert & Stephens, 2018). However, such studies may require significant programming skills. In any case, the reason that power is somewhat more difficult to estimate in designs with crossed random variables is that in these designs, power is affected by the number of participants, the number of stimuli, the amount of variation between participants, and the amount of variation between stimuli (Judd et al., 2017; Westfall et al., 2014).

Westfall et al. (2014) demonstrated that in designs with two crossed random variables increasing the number of participants will not necessarily increase power. That is, with a fixed number of stimuli, statistical power will asymptote at a particular level even when thousands or even millions of participants are included in the study. Westfall and colleagues refer to this level as the "maximum attainable power"

and demonstrated that under realistic conditions, maximum attainable power is likely to be quite small when fewer than 16 participants and fewer than 16 stimuli are included in the study. Indeed, researchers who are interested in detecting small effects, may need to present a large number of participants with large numbers of stimuli to have adequate statistical power. However, such studies have drawbacks. Presenting participants with large numbers of stimuli may not be feasible or desirable. Long questionnaires may exhaust participants and cause attrition (Sahlqvist, Song, Bull, Adams, Presont, & Ogilvie, 2011), which may bias estimates (Prokopek, 2011).

Planned missingness. One potential remedy is to use planned missingness designs in which participants judge subsamples of stimuli (Brauer & Curtin, 2018; Westfall et al., 2014; Judd et al., 2017). In contrast to missing data caused by participant attrition, in planned missingness designs, participants are randomly assigned to respond to different subsamples of stimuli, such that, across participants, data are "missing completely at random" (MCAR). For example, in a study examining 100 attributes, each participant may rate a random sample of 25 attributes. This procedure can reduce the burden on participants, and although it may lead to a slight decrease in statistical power, data that are missing completely at random will not bias parameter estimates (Prokopek, 2011).

OLS analyses are not well-equipped to handle missing data, and are therefore, not well-suited for the analysis of planned missingness designs. For example, in repeated measures ANOVA, participants with missing data are dropped from the analyses. In contrast, LMEMs can accommodate planned missingness by using restricted maximum likelihood estimation (ReML), which accounts for the fact that due to pure chance some stimuli may be responded to by more participants than other stimuli (Brauer & Curtin, 2018). Researchers who only use OLS analyses thus have a double handicap. They cannot generalize their findings beyond the stimuli that were included in the study and they are constrained with regard to the number of stimuli that they present to participants, because OLS analyses are less well equipped to handle missing data than are LMEMs.

The Present Studies

Research on stereotyping frequently uses designs in which the variable "participants" is crossed with the variable "attributes"; ideally, such designs should be analyzed with LMEMs in which both of these variables should be treated as random rather than fixed. One barrier to using LMEMs in stereotyping research is that researchers tend to have less experience with them than with more conventional OLS analyses. Whereas LMEMs have been discussed in the context of nested designs for several decades, the application of LMEMs to designs with crossed random variables is relatively new (Baayen et al., 2008). Even researchers who have some familiarity with LMEMs may not know how to apply them to research on stereotyping. Indeed, despite calls for the use of these techniques in stereotyping research (Judd et al., 2012, 2017), Judd and colleagues (2019) are to our knowledge, the only researchers who have treated both participants and stimuli as random variables in their analysis of stereotyping data. Our goal in this paper is to show in two studies, one of which uses planned missingness (Study 2), how LMEMs can be applied to stereotyping research.

In Study 1, we use LMEMs to assess the valence of participants' gender stereotypes along eight predetermined attributes. With only eight attributes, our analyses are underpowered, at least for a small or medium-sized effect. Indeed, given the sample size of both participants and attributes in the first study, estimates of power for small and medium effects are .20 and .50, respectively (Westfall et al., 2014). However, our goal is pedagogical; we focus on the application of LMEMs to designs that are commonly used. In Study 2, we show how researchers can employ planned missingness to obtain data for a broad range of attributes while avoiding participant exhaustion and attrition.

STUDY 1

In Study 1, we examined the tendency to view women more positively than men (Eagly & Mlidanic, 1994, Glick et al., 2004; Sullivan et al., 2018). As noted above, this "women are wonderful effect" reflects the positive valence of the communal attributes associated with women (e.g., friendly, caring) relative to many attributes associated with men and agency (e.g., competitive, aggressive). We expected to find evidence of the "women are wonderful" effect, that is, that participants would ascribe more positive (vs. negative) attributes to women than to men. We further expected this effect to be

stronger among female versus male participants (Glick et al., 2004; Rudman & Goodwin, 2004), as studies have suggested that women tend to exhibit more ingroup bias, that is, show a preference for members of their own groups, than do men.

Method

Participants

Participants (N = 239) at a midsize public college in the Northeastern U.S. were recruited for a study on social attitudes in exchange for course credit. Five students failed to indicate their gender or identified as transgender or non-binary and were dropped from the analyses. Of the remaining 234 students, most (75.6%) were women. Participants ranged in age from 18 to 48 years old (M = 20.62, SD = 3.80). Most participants identified as White (42.6%), followed by Black (32.6%), Multiracial (9.1%), Latinx (9.1%), Other (4.4%), and Asian (2.2%). Three participants declined to indicate their ethnicity.

Procedure

Participants completed a percentage estimation task in which they estimated the percentage of women and men, on a 0 to 100% scale, who possessed eight stereotype-relevant attributes commonly used in gender stereotype research. Four attributes were positive in valence (warm, nurturing, ambitious, independent), whereas four attributes were negative in valence (emotional, nagging, aggressive, arrogant). The order of target gender (women vs. men) was counterbalanced across participants. Thus, the study had a 2 (Target Gender: Women vs. Men) X 2 (Attribute Valence: Positive vs. Negative) X 2 (Participant Gender: Women vs. Men) design with the first two factors varying within participants. All predictors (i.e., Target Gender, Attribute Valence, and Participant Gender) were coded 1 and -1.

Notably, the attributes also varied as to whether they were stereotypic of women (i.e., warm, nurturing, nagging, emotional) or men (i.e., ambitious, independent, arrogant, aggressive). None of the effects that we report depended on attribute stereotypicality, and analyses accounting for attribute stereotypicality yielded only one additional significant effect (i.e., attributes stereotypic of women were more commonly ascribed to women, and attributes stereotypic of men were more commonly ascribed to men). Since our purpose in describing the analyses is pedagogical, we focus on the simpler analysis.

Results

We begin with a brief description of how we set up the LMEM analyses (see Brauer and Curtin, 2018, for a more extensive tutorial). An annotated copy of our R code is provided in the supplemental materials.

Issues to Consider Before Estimating LMEMs: Degrees of Freedom, Estimators, and Optimizers

Researchers who use LMEMs have additional issues to consider when building their models, as compared to researchers who use more conventional OLS analyses. Some programs, such as R and SAS, allow researchers to choose the method for calculating denominator degrees of freedom. Different methods are available, and scholars disagree about which method is best (Baayen et al., 2008). Some (Locker et al., 2007) suggest the Satterthewaite method, whereas most recent publications (Brauer & Curtin, 2018; Judd et al., 2012) favor the Kenward-Roger's method, a minor extension of the Satterthwaite method. Both methods yield similar results and usually result in degrees of freedom that contain decimals. In addition, R and other programs allow researchers to specify an optimization procedure (or "optimizer"). LMEM models estimate parameters using restricted maximum likelihood (ReML), which is an iterative process that converges around a solution with the largest log likelihood (see Brauer & Curtin, 2018, for a more detailed explanation of ReML). However, models that are complex or have few data points per cell in the design will sometimes fail to converge around a solution, which optimizers can help address (see Brauer & Curtin, 2018 for a detailed description of ways to address model convergence issues.)

We analyzed the data as a LMEM, using the lme4 package (Bates, Maechler, Bolker, & Walker, 2015) in R with ReML estimation. We also used the Satterthwaite method for calculating denominator degrees of freedom, and the bobyqa optimizer.

In the Supplemental Materials we provide instructions for preparing the data for LMEM analyses, including how to convert the data from wide form to long form, which is required to run LMEMs in R. As some readers may be unaccustomed to thinking about data in long form, we provide sample lines of the dataset in Table 1. Participants provided the same eight judgments of positive and negative attributes for

both male and female targets. Thus, for each participant, there were 16 responses and each participant had 16 rows of data in the data file in long form. Note, however, that there are only eight unique attribute identification numbers, as participants only rated eight unique attributes.

Setting Up the Model: Fixed Effects, Random Effects, and Variance Components

We regressed participants' target judgments on contrast-coded predictors for Target Gender, Attribute Valence, Participant Gender, and their interactions (Table 3).[1] The so-called "fixed effects" (i.e., the average effect of our predictors across participant and attributes) are reported in the top half of the table. Note that the fixed effects allow researchers to determine whether their hypotheses have been supported. The parameter estimates for the fixed effects can be interpreted just as one would interpret parameter estimates in standard regression. For example, the effect of Target Gender represents the mean difference between judgments of men and women, i.e., whether men are seen as having these eight attributes less than women, or vice-versa. Likewise, the *p*-value associated with Attribute Valence tests the mean difference between positive and negative attributes. We calculated effect sizes from the values of *F* provided in the output (Page-Gould et al., 2019).

The advantage of LMEMs over OLS analyses is that LMEMs allow researchers to estimate random effects. These random effects simply refer to the (multiple) reasons that the model predictions may be erroneous. Extending Barr, Levy, Scheepers, and Tily's (2013) framework, Brauer and Curtin (2018) proposed adding a random intercept for each random variable that causes non-independence. Here, both participants and attributes cause non-independence, and we thus included a by-participant random intercept and a by-attribute random intercept. The by-participant random intercept indicates the amount of variability due to mean differences in participants' judgments. That is, averaging across attributes, some participants likely provide higher judgments than do others. Likewise, the by-attribute random intercept

[1] When we estimated this model, the lme4 package issued a warning indicating a singularity issue. None of the variance components were equal to zero, nor did any of the covariances approach 1 or -1. In other words, there were no obvious reasons for the singularity warning (Bates et al., 2015). We therefore estimated additional models, removing random slopes, consistent with recommendations by Barr et al. (2013) and Brauer and Curtin (2018), trying different methods for calculating degrees of freedom, and different optimizers (see Supplemental Materials for Model Results). All of the effects we report below were also produced in these models, suggesting that the potential singularity issue did not affect our results. We therefore present the results of the original model described above.

indicates the amount of variability due to mean differences in attributes. That is, averaging across participants, some attributes are likely seen as more common than other attributes. All random effects, that is, the random intercepts (mentioned above) and the random slopes (mentioned below), are displayed in the bottom half of Table 2. Note that random effects are sometimes referred to as "variance components."

Brauer and Curtin (2018) further proposed to add a random slope for each predictor that varies within a particular random variable. Random slopes are an indication of the variability in an effect (e.g., the effect of Target Gender) from participant to participant (or from attribute to attribute). For example, the variance estimate for the by-participant random slope of Target Gender ($\sigma^2 = 6.04$) indicates the amount of variability in the Target Gender effect from one participant to another. More simply, it indicates that some participants judged target men and women as more different from each other than did other participants. We can see that this variance estimate is relatively small compared to the other estimates in the bottom of Table 2. Thus, differences in judgments of male and female targets did not vary much across participants. There is also a variance component for the by-attribute random slope of Target Gender ($\sigma^2 = 117.71$), which indicates the extent to which the effect for target gender varies across attributes. That is, participants judged target men and women as being very different from each other on some attributes and very similar to each other on other attributes. The variance component for the by-attribute random slope of Target Gender was comparatively quite large (see the bottom of Table 2). Thus, the target gender effect varied considerably from attribute to attribute.

In Study 1, Target Gender, Attribute Valence, and their interaction all vary within participants (see Table 1 for sample rows of the dataset). We therefore included a by-participant random slope for target gender, a by-participant random slope for attribute valence, and a by-participant random slope for the target gender by attribute valence interaction. In contrast, Target Gender, Participant Gender, and their interaction varied within attribute. We, therefore, included by-attribute random slopes for the effects of Target Gender, Participant Gender, and their interaction.

Random effects are important for a few reasons. First, they allow researchers to verify whether something is wrong in the dataset. Random effects with unusually high variances may indicate a problem (e.g., a data entry error or the presence of outliers). Second, only if the researchers specify the correct random effects will the estimated model produce the correct standard errors for the parameter estimates and thus the correct inferential tests (Barr et al., 2013). Third, sometimes a large random effect may suggest an omitted moderator variable. For example, a large by-participant random slope for Target Gender indicates that some participants judge men and women as radically different from each other, whereas other participants do not. Perhaps the former are high in benevolent sexism, but the latter are not (Glick & Fiske, 1996). In any case, treating the effect of Target Gender as fixed would implicitly argue that the effect of Target Gender (i.e., mean differences in judgments of target women versus target men) is the same across participants, which one cannot know to be true.

Interpreting the Results

Table 2 and Figure 1 present condition means. Condition means and their standard errors were estimated using the emmeans package in R. Note, that the emmeans package accounts for the LMEM model that was estimated in its estimation of means and standard errors. Thus, estimates of the standard errors account for both crossed random variables. If we were to estimate means and standard deviations without accounting for variation between participants and variation between attributes, the estimates of condition means would be accurate. However, the estimates of the standard deviations would be inaccurate because the variability in the ratings stems from two sources: participants and attributes. (Judd et al., 2017).

Tests of the predictors revealed no main effects of Target Gender, Attribute Valence, or Participant Gender. The hypothesized Target Gender X Attribute valence interaction was not significant, nor did the effect of Target Gender depend on Participant Gender. Indeed, only two effects emerged. A marginal Participant Gender X Attribute Valence interaction indicated that female participants were somewhat more likely to ascribe positive versus negative attributes to targets than male participants. However, this interaction was qualified by a significant three-way Participant Gender X Attribute Valence

X Target Valence interaction (Figure 1). Patterns of mean differences indicated that, consistent with previous research, although both men and women exhibited the tendency to ascribe more positive (vs. negative) characteristics to female (vs. male) targets, the effect was more pronounced for women (vs. men). More simply, participants tended to exhibit the "women are wonderful" effect, but this effect was more pronounced among female participants than among male participants.

Discussion

In Study 1, we showed how LMEMs might be applied to commonly used repeated measures designs used in stereotyping studies by replicating the "women are wonderful" effect. We found that women more than men ascribed positive (vs. negative) attributes to target women (vs. men), consistent with Rudman and Goodwin (2004), who similarly found that women have a stronger tendency to exhibit ingroup bias than do men. Further, although our design was underpowered to find small and medium effects (Westfall et al., 2014), we demonstrated how LMEMs, which provide unbiased standard errors of the parameter estimates and more generalizable results than OLS analyses, could be applied to commonly used designs in stereotyping research.

In Study 2, we show how these LMEMs can be used for data from a more complicated design, analyzing the valence judgments of target men and women in stereotypically feminine or masculine roles using participant-generated attributes. Studies that have examined participant-generated attributes (e.g., Glick et al., 1997, 2004) have noted that such procedures are higher in ecological validity because they reflect participants' actual stereotypes. As a result, participants are likely to generate some shared attributes and some unique attributes. Thus, attributes are neither fully nested nor fully crossed with participants. Finally, as we wished to obtain ratings of many target groups, we employed a planned missingness design to avoid exhausting participants or encouraging attrition. Thus, in Study 2, we show how LMEMs can be used for complex designs with data that are missing completely at random.

STUDY 2

We examined participants' evaluations of women and men in roles traditionally occupied by women (i.e., being a stay-at-home parent, being feminine) and roles traditionally occupied by men (i.e.,

having a career, being masculine). We again expected to find evidence of the "women are wonderful" effect. However, we predicted that positive evaluations of women (relative to men) would only emerge for target women described as stay-at-home moms, but not career women. In other words, we expected target women to be judged as "wonderful" so long as were described as conforming to gender role expectations (Glick & Fiske, 1996). We also predicted that participants would exhibit a preference for men in masculine (vs. feminine) roles, consistent with the research on backlash (Rudman et al., 2012; Sullivan et al., 2018) indicating that both male and female targets, who exhibit counterstereotypic behavior, tend to be evaluated negatively. Finally, we explored whether these effects depended on participant gender. Although some work has indicated that ingroup bias is stronger among women (e.g., Rudman & Goodwin, 2004), studies in which gender roles were manipulated have produced inconclusive results. Some researchers have found no evidence of ingroup bias (Heilman et al., 2004), and others report results suggesting that women more negatively evaluate women in stereotypically masculine roles (Garcia-Retamero & López-Zafra, 2006).

Method

Participants

Participants (N = 249) were recruited from Prolific Academic for a study of beliefs about different social groups and social issues and were paid $1.50USD. Eight participants failed to indicate their gender or identified as transgender or non-binary and were dropped from the analyses. Thus, the final sample consisted of 241 participants. Participants ranged in age from 19 to 79 years old (M = 34.20, SD = 11.57). Half identified as men (50.0%), and half identified as women. Most participants identified as heterosexual (83.0%), followed by bisexual (12.0%), gay or lesbian (3.7%), and other or unsure (1.2%). Most participants identified as White (71.8%), followed by Black (9.1%), multiethnic (7.1%), Latinx (5.8%), Asian (4.2%), and other (1.2%). Fewer than 1% of participants identified as Native American or East Indian. Most participants had completed at least some college or had a college degree (60.2%), followed by individuals with post-graduate education (20.3%), associates degrees (10.0%), high school diploma (8.3%), and individuals who did not complete high school (1.2%).

Procedure

Participants were asked to provide evaluations of target men and women in different social roles drawn from Glick and colleagues (2015). Four of the target groups were described as occupying feminine roles (i.e., stay-at-home mom, feminine woman, stay-at-home dad, feminine man), whereas four of the target groups were described as occupying masculine roles (i.e., career man, masculine men, career women, masculine women).[2]

For each target, participants were instructed to list the first five attributes that came to mind when they thought about that target group. Then, participants rated the valence of each attribute on a 1 (*very negative*) to 7 (*very positive*) scale. The order in which targets were presented was counterbalanced across participants.

Because generating five attributes and assigning five valence scores to each of ten target groups (i.e., 100 items total) seemed likely to exhaust participants, we used a planned missingness design, randomly presenting to participants three female target groups and three male target groups (six target groups in total). Note that although some participants generated fewer than five attributes, participants never generated the same attribute more than once for a given target. However, participants could possibly generate the same attribute for two different target groups. For example, a participant might judge both stay-at-home dads and stay-at-home moms as "caring."

Participants generated a total of 5699 attributes for all targets, 1861 (32.7%) of which were unique attributes.[3] We were interested in whether evaluations of women and men depended on whether they were described in roles complementing feminine or masculine stereotypes. Thus, the study had a 2 (Target Gender: Women vs. Men) X 2 (Target Role: Feminine vs. Masculine) X 2 (Participant Gender:

[2] Participants also provided ratings of feminist women and feminist men, which were not included in the present analyses. However, we are interested here in whether the valence of target group judgments depended on whether targets were in traditionally feminine or masculine roles. As feminists are not considered traditionally feminine or masculine (Glick et al., 2015), we did not include them in these analyses.

[3] Some attributes were combined during data cleaning. For example, words that were misspelled, such as "inteligent" were treated as the same attribute as words that were spelled correctly. Further, words that were written in plural (e.g., leaders) were combined with words in the singular (leader). Analyses with and without the cleaned data yielded similar conclusions.

Women vs. Men) design with repeated measures on the first two factors. All predictors were coded 1 and -1.

Results

We again estimated a LMEM model using the lme4 package (Bates et al., 2015) in R with restricted maximum likelihood estimation and the bobyqa optimizer. The data were again in long form, consistent with Study 1. We regressed ratings of valence on Target Gender, Target Role, Participant Gender, and their interactions.

As was the case in Study 1, we considered participants and attributes to be random variables. Both variables caused non-independence, because each participant provided multiple ratings and some of the attributes were judged by more than one participant. Following Brauer and Curtin (2018), we therefore estimated by-participant and by-attribute random intercepts. As in Study 1, these random intercepts indicate the amount of variability in judgments of target valence due to participants and attributes, respectively.

The effects of Target Gender, Target Role, and their interaction varied within participants. Thus, we estimated by-participant random slopes of these effects. However, unlike Study 1, we did not estimate any by-attribute random slopes. In Study 1, attributes systematically varied with respect to valence, and participants rated the extent to which each attribute described a target group. We were interested in whether differences between positive and negative attributes could be accounted for by other variables in the model (e.g., target gender, participant gender). Thus, it was important to statistically control for between-attribute differences. Here, we were not interested in the attributes themselves, as participants provided the valence ratings directly; that is, participants rated how positively or negatively they perceived the group with respect to each attribute they generated.

Table 4 displays condition means calculated with the emmeans package. Table 5 lists the fixed effects (top) and random effects (bottom). Tests of the variance components indicated that the majority of the variability in the data was due to differences between attributes; target judgments did not vary much across participants. Estimates of the fixed effects revealed no main effect of Participant Gender. However,

a significant main effect of Target Gender emerged indicating that, as expected, participants viewed women (M = 4.70, SE = 0.06) more positively than men (M = 4.55, SE = 0.06). A main effect of Target Role also emerged indicating that targets in feminine roles (M = 4.68, SE = 0.06) were regarded more positively than were targets in masculine roles (M = 4.57, SE = 0.06). Only one interaction between Target Gender and Target Role emerged (Figure 2). However, inconsistent with expectations, participants perceived women in feminine and masculine roles approximately equally positively, whereas men in masculine roles were judged, surprisingly, less positively than men in feminine roles. None of these effects depended on Participant Gender.

Discussion

In Study 2, we analyzed perceptions of target men and women in feminine and masculine roles using LMEMs. We again showed how LMEMs may be easily used to analyze data with complex designs and that LMEMs can easily account for data that are missing completely at random, for example, when using planned missingness designs and when using participant-generated attributes. Were we to analyze these data using OLS analyses, participants would be dropped from the analyses because of missing data. The use of LMEMs, therefore, allowed us to use data from all participants and to assess participants responses to a larger number of targets without exhausting participants (Brauer & Curtin, 2018; Prokopek, 2011).

Differences between attributes accounted for the majority of variability in the data. Further, consistent with expectations, we found evidence of the "women are wonderful" effect. That is, participants judged target women more positively than target men, regardless of role. We expected a Target Gender X Target Role interaction indicating that the women-are wonderful effect would be stronger for women in feminine (vs. masculine) roles. However, we were surprised to find the opposite. Men and women in feminine roles were judged equally positively, whereas women in masculine roles were judged more positively than men in masculine roles. In retrospect, these results are consistent with Glick et al. (2004), who suggested that the women are wonderful effect is also driven by perceptions that men are viewed as "bad but bold," as men are respected and also resented for their power. Perhaps the

negative stereotypes associated with men were attenuated for men in feminine (vs. masculine) roles. Nevertheless, somewhat positive evaluations of men in feminine roles is inconsistent with work on backlash suggesting that men who are perceived as feminine are often viewed negatively (Rudman et al., 2012; Sullivan et al., 2018).

General Discussion

We have demonstrated how LMEMs can be applied to studies on stereotyping using researcher-generated attributes (Study 1), participant-generated attributes (Study 2), and planned missingness designs (Study 2). As previously noted, few researchers examining stereotyping have utilized LMEMs—despite the clear benefits of doing so.

One reason to use LMEMs is to obtain more accurate estimates of *F*. In a re-analysis of previously published data, Judd et al. (2012) demonstrated that previously published effects were overestimated using OLS analyses. Using data from the Implicit Associations Test (Greenwald & Banaji, 1995), Wolsiefer, Westfall, and Judd (2016) similarly demonstrated that OLS analyses overestimated traditional test statistics by approximately 60%. Thus, LMEMs generally produce more accurate test statistics than do OLS analyses.

Further, inconsistencies across studies may be better addressed by LMEMs, as LMEMs allow researchers to generalize from their sample of stimuli (in our case, attributes) to the entire population of stimuli. Indeed, there is much concern about the reproducibility of psychological science. Some of this concern has been addressed through study preregistration and calls for more open science. Undoubtedly, these tactics are positive steps forward. However, less attention has been paid to issues of measurement (Flake, Pek, & Hehman, 2017; Hussey & Hughes, 2018) and sampling of stimuli (Westfall, Judd, & Kenny, 2015), which may also produce inconsistencies across studies, including conflicting results pertaining to stereotype change (e.g., Diekman & Eagly, 2000; Eagly et al., 2019; Haines et al., 2016).

Some researchers have called for editors and authors to include constraints on generality (COG) statements in their papers, in which authors explicitly specify the target population of a study in terms of participants and stimuli (Simons, Shoda, & Lindsay, 2017). Although most researchers are accustomed to

thinking about how their sample of participants may (or may not) generalize to a specific population, they may be less used to thinking about stimuli (e.g., attributes) as being a sample from a population of stimuli. However, as Westfall et al. (2015) demonstrate, there is reason to be concerned about stimuli-specific results.

LMEMs address this concern by specifically modeling the variability between the stimuli used in the study and allowing inferences about the entire population of stimuli from which the stimuli used in the study were drawn. LMEMs, may, therefore, allow researchers to more directly examine the extent to which items used to assess agency and communion in gender stereotyping and person perception research, more generally, generalize across studies. Further, researchers may use LMEMs to determine whether conflicting results pertaining to gender differences in agency (Eagly et al., 2019; Folberg et al., 2019; Hentschel et al. 2019), or changes in stereotypes over time (Eagly et al., 2019; Haines et al., 2016) are due to the specific stimuli used in a given study.

In addition, LMEMs offer researchers more flexibility than do conventional OLS analyses. Fixed effects and effects from OLS analyses can be thought of as the average effect across a sample of stimuli. For example, the estimates of the fixed effects in Table 3 and Table 5 are the average effects of our predictors of interest. However, as our analyses demonstrated, there is variability in the size of those effects from participant to participant and, also, from attribute to attribute. For example, in Study 1, the effect of Target Gender did not vary much for participants, but it varied quite a bit for attribute. This variability in effects is what Yeager et al. (2019) refer to as "effect heterogeneity."

OLS analyses do not permit researchers to model "effect heterogeneity" due to attributes or other stimuli, as they treat attributes as fixed. However, LMEMs allow researchers to treat *both* participants and attributes as random variables, thus, enabling researchers to assess "effect heterogeneity" both with regard to participants and with regard to attributes. Knowing the extent to which effects vary across attributes is important for a few reasons. First, it gives researchers an indication of how "particular" attributes might be; a great amount of attribute-to-attribute variability in effects suggest that the effects of interest may be large for some attributes but non-existent for others. Second, it allows researchers to ask direct questions

about the attributes themselves, such as whether attribute-level moderators qualify their effects of interest. For example, research on gender stereotypes might examine whether the salience of different attributes affects perceptions of target stereotypicality.

In addition, LMEMs more easily handle missing data than do OLS analyses (Brauer & Curtin, 2018; Prokopek, 2011), which may be one way to present participants with the large number of stimuli required to have adequate power (Westfall et al., 2014). Further, so long as data are missing completely at random, as would be the case if participants are randomly assigned to subsets of stimuli, missing data will not bias parameter estimates (Prokopek, 2011). Thus, planned missingness provides a powerful tool for researchers examining a broad range of targets or stimuli. Using LMEMs enables researchers to take advantage of such designs and, more generally, have greater choices available to them in the designs that they choose.

Women are Wonderful and Backlash Towards Masculine Men

In both studies, we found some evidence of the "women are wonderful" effect. That is, participants viewed women more positively than they viewed men However, in Study 1 this effect only emerged for women (vs. men). In Study 2, we found that the "women are wonderful" effect depended on role information. Unexpectedly, however, participants judged men in *feminine* roles more positively than men in masculine roles.

The latter finding is inconsistent with research suggesting that men perceived as feminine may be judged to have the least positive attributes (Rudman et al., 2012; Sullivan et al., 2018). However, as noted above, it is consistent with research suggesting that men are perceived as "bad but bold," that is, that men are simultaneously valued and resented for their power (Glick et al., 2004). Thus, perhaps men in gender-role congruent roles are viewed especially negatively as men may also enact their masculinity through dominance and aggression (Bosson & Vandello, 2011; Vandello & Bosson, 2013). Indeed, since the rise of the #metoo movements, started by Tarana Burke in 2006 and reinvigorated in 2017, individuals have engaged in a broader cultural discussion about negative aspects of masculinity. For example, some work indicates that perceptions that men should be strong, stoic, and domineering may result in bullying and

abuse at work that negatively affects both men and women (Berdahl, Cooper, Glick, Livingston, & Williams, 2018; Glick, Berdahl, & Alonso, 2018). Thus, recognition of the negative elements of masculinity may explain participants' particularly negative responses to men in masculine roles.

Limitations and Additional Considerations

The goal of this paper was to show how LMEMs can be applied to two studies examining gender stereotypes by examining the well-established "women are wonderful" effect. Thus, we discuss the limitations of our analyses as a way of illustrating broader considerations in the use of LMEMs.

Deciding whether variables are random or fixed or whether to model by-participant and by-attribute effects is not always straightforward (Brauer & Curtin, 2018). One might argue that the attributes chosen for Study 1 were selected because they have some particular "ideal" quality (e.g., they are the most agentic or the most communal attributes that exist) and thus would be more appropriately treated as a fixed variable. But one could also argue that these are a sample of attributes commonly used to assess perceptions of agency and communion in gender stereotyping research and that the goal is to be able to generalize conclusions to the broader population of attributes used to assess agency and communion; in this case, attributes would be more appropriately treated as a random variable. Similarly, one might argue that by-attribute random slopes should be modeled when participants rate some of the same sets of attributes (Study 2), whereas others might agree with our strategy to only include by-participant random slopes but not by-attribute random slopes in Study 2. Experts can reasonably disagree about these decisions. However, as Brauer and Curtin, among others, have recommended, it is generally best to let the hypotheses guide the modeling strategy.

As we acknowledged, the analyses in Study 1 were underpowered (Westfall et al., 2014) so that, for example, failure to find significant a Target Gender X Attribute Valence interaction should not be interpreted as disconfirming evidence of the "women are wonderful" effect (Eagly & Mladinic, 1994). Study 1 was intended as a pedagogical tool for showing how LMEMs can be applied to repeated measures designs, and, in fact, despite being underpowered, it replicated the previous finding in which

women associate women with positive attributes more strongly than men do (Glick et al., 2004; Rudman & Goodwin, 2004).

Further, LMEMs may improve the generalizability of results for researchers conducting stereotyping research. However, increased generalizability should not be interpreted to mean that LMEM results generalize from a given sample to any population, for example, from a sample of U.S. participants to participants in a non-Western, non-industrialized country. The need to explicitly define one's populations of interest (both in terms of stimuli and participants) also applies to researchers who use LMEMs (Simons et al., 2017).

Conclusion

We have demonstrated how LMEMs can be used to advance research on stereotyping in two separate studies by producing more accurate test statistics, improving the generalizability of results, and allowing for more flexibility in design considerations. Further, LMEMs can be used to address inconsistencies in gender stereotyping, which might emerge as a result of sampling error due to stimuli. We also discussed how a close examination of effect heterogeneity may generate new hypotheses that can be tested in subsequent research. Finally, we showed how LMEMs can easily handle data that are missing at random, allowing researchers greater flexibility to present participants with a larger number of stimuli, which may be required to have adequate power for LMEM analyses. We hope that these examples stimulate increased use of LMEMs in stereotyping research, helping to advance the reproducibility of the findings and the conclusions that can be drawn from them.

References

Abele, A. E., Hauke, N., Peters, K., Louvet, E., Szymkow, A., & Duan, Y. (2016). Facets of the fundamental content dimensions: Agency with competence and assertiveness—Communion with warmth and morality. *Frontiers in Psychology, 7,* Article ID 1810. https://doi.org/10.3389/fpsyg.2016.01810

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59,* 390–412. http://dx.doi.org/10.1016/j.jml .2007.12.005

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68,* 255–278. http://dx.doi.org/10.1016/j.jml.2012 .11.001

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67,* 1–48. https://doi.org/10.18637/jss.v067.i01

Berdahl, J. L., Cooper, M., Glick, P., Livingston, R. W., & Williams, J. C. (2018). Work as a masculinity contest. *Journal of Social Issues, 74,* 422-448. https://doi.org/10.1111/josi.12289

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/10.1177/0963721411402669

Brauer, M., & Curtin, J. J. (2018). Linear mixed-effects models and the analysis of non-independent data: A unified framework to analyze categorical and continuous independent variables that vary within subjects and/or within items. *Psychological Methods, 23,* 389-411. https://doi.org/10.1037/met0000159

Brescoll, V. L., & Uhlmann, E. L. (2008). Can an angry woman get ahead? Status conferral, gender, and expression of emotion in the workplace. *Psychological Science, 19*, 268–275. https://doi.org/10.1111/j.1467-9280.2008.02079.x

Clark, H. (1973). The language as fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12,* 335–359. http://dx.doi.org/10.1016/S0022- 5371(73)80014-3

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Diekman, A. B., & Eagly, A. H. (2000). Stereotypes as dynamic constructs: Women and men of the past, present, and future. *Personality and Social Psychology Bulletin, 26*, 1171-1188. https://doi.org10.1177/0146167200262001

Diekman, A. B., Eagly, A. H., & Johnston, A. M. (2010). Social structure. In J. F. Dovidio, M. Hewstone, P. G. Glick, & V. M. Esses (Eds.), *The Sage handbook of prejudice, stereotyping, and discrimination* (pp. 209–224). Thousand Oaks, CA: Sage. https://doi.org/10. 4135/9781446200919.n13

Diekman, A. B., Steinberg, M., Brown, E. R., Belanger, A. L., & Clark, E. K. (2017). A goal congruity model of role entry, engagement, and exit: Understanding communal goal processes in STEM gender gaps. *Personality and Social Psychology Review, 21*(2), 142–175. https://doi.org/10.1177/1088868316642141

Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review, 109*, 573–598. https://doi.org/10.1037/0033-295X.109.3.573

Eagly, A. H., & Mladinic, A. (1994). Are people prejudiced against women? Some answers from research on attitudes, gender, stereotypes, and judgments of competence. *European Review of Social Psychology, 5,* 1-35. https://doi.org/10.1080/14792779543000002

Eagly, A. H., Nater, C., Wood, W., Miller, D. I., Kaufmann, M., & Sczesny, S. (2019). Gender stereotypes have changed: A cross-temporal meta-analysis from 1946-2018. *American Psychologist.* Advance online publication. https://doi.org/10.1037/amp0000494

Eagly, A. H., Wood, W., & Diekman, A. B. (2000). Social role theory of sex differences and similarities: An appraisal. In T. Ecks & H. M. Trautner (Eds.), *The developmental social psychology of gender*

(pp. 123-174). Marwah, NJ: Erlbaum. https://doi.org/10.4135/9781446249222.n49

Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology, 82*, 878–902. https://doi.org/10.1037/0022-3514.82.6.878

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science, 8*, 370-378. https://doi.org/10.1177/1948550617693063

Folberg, A. M., Kercher, K., & Ryan, C. S. (2019, online). The hidden role of dominance in career interests: A bifactor analysis of agentic and communal goal orientations. *Sex Roles.* https://doi.org/10.1007/s11199-019-01104-1

Garcia-Retamero, R., & López-Zafra, E. (2006). Prejudice against women in male-congenial environments: Perceptions of gender role congruity in leadership. *Sex Roles*, 55, 51-61. https://doi.org/10.1007/s11199-006-9068-1

Glick, P., Berdahl, J. L., & Alonso, N. M. (2018). Development and validation of the Masculinity Contest Culture Scale. *Journal of Social Issues, 74,* 449-476. https://doi.org/10.1111.josi.12280

Glick, P., Diebold, J., Bailey-Werner, B., & Zhu, L. (1997). The two faces of Adam: Ambivalent sexism and polarized attitudes toward women. *Personality and Social Psychology Bulletin, 23*, 1323-1334. https://doi.org/10.1177/01461672972312009

Glick, P., & Fiske, S. T. (1996). The Ambivalent Sexism Inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology, 70*, 491–512. https://doi.org/10.1037/0022-3514.70.3.491

Glick, P., Lameiras, M., Fiske, S.T., Eckes, T., Masser, B., Volpato, C., et al. (2004). Bad but bold: Ambivalent attitudes toward men predict gender inequality in 16 nations. *Journal of Personality and Social Psychology, 86,* 713–728. https://doi.org/10.1037/0022-3514.86.5.713

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102*, 4-27. https://doi.org/10.1037/0033-295X.102.1.4

Haines, E. L., Deaux, K., & Lofaro, N. (2016). The times they are a-changing … or are they not? A comparison of gender stereotypes, 1983–2014. *Psychology of Women Quarterly, 40*, 353-363. https://doi.org/10.1177/0361684316634081

Heilman, M. E., Wallen, A. S., Fuchs, D., & Tamkins, M. M. (2004). Penalties for success: Reactions to women who succeed at male gender-typed tasks. *Journal of Applied Psychology, 89*, 416-427. https://doi.org/10.1037/0021-9010.89.3.416

Hentschel, T., Heilman, M. E., & Peus, C. V. (2019). The multiple dimensions of gender stereotypes: A current look at men's and women's characterizations of others and themselves. *Frontiers in Psychology.* Advance online publication. https://doi.org/19.3389/fpsyg.2019.00011

Hussey, I., & Hughes, S. (2018, November 19). Hidden invalidity among fifteen commonly used measures in social and personality psychology. https://doi.org/10.31234/osf.io/7rbfp

Judd, C. M., Garcia-Marques, T., & Yzerbyt, V. (2019). The complexity of dimensions of social perception: Decomposing bivariate associations with crossed random factors. *Journal of Experimental Social Psychology, 82,* 200-207, https://doi.org/10.1016/j.jesp.2019.01.008

Judd, C. M., McClelland, G. H., & Ryan, C. S. (2017). *Data analysis: A model comparison approach to regression, ANOVA, and beyond* (3rd ed.). New York, NY: Routledge.

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology, 103,* 54–69. http://dx.doi.org/10.1037/a0028347

Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual Review of Psychology, 68,* 601–625. http://dx.doi.org/10.1146/ annurev-psych-122414-033702

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*, 371–392. https://doi.org/10.1037/a0037215

Kurebayashi, K., Hoffman, L., Ryan, C. S., & Murayama, A. (2012). Japanese and American perceptions of group entitativity and autonomy. *Journal of Cross-Cultural Psychology, 43*, 349-364. https://doi.org/10.1177/0022022110388566

Locker, L., Jr., Hoffman, L., & Bovaird, J. A. (2007). On the use of multilevel modeling as an alternative to items analysis in psycholinguistic research. *Behavioral Research Methods, 39*, 723-730. https://doi.org/10.3758/BF03192962

Nakagawa, S., Johnson, P.C.D., & Schielzeth, H. (2017). The coefficient of determination R^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface, 14,* 20170213. https://doi.org/10.1098/rsif.2017.0213

Page-Gould, E., Sharples, A., & Song, S. (2019, October). Effect Sizes for Models of Longitudinal Data. In P. E. Shrout (Chair), Modeling Mediation Processes in Longitudinal Data. Symposium conducted at the annual meeting of the Society for Experimental Social Psychology, Toronto, ON, Canada. Retrieved from: https://osf.io/af4h5/

Pellegrini, V., De Cristofaro, V., Giacomantonio, M., & Salvati, M. (2020, online). Why are gay leaders perceived as ineffective? The role of the type of organization, sexual prejudice, and gender stereotypes. *Personality and Individual Differences, 157.*https://doi.org/10.1016/j.paid.2020.109817

Pokropek, A. (2011). Missing by design: Planned missing-data designs in social science. *Research & Methods, 20,* 81-105.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models* (2nd ed.). Thousand Oaks, CA: Sage.

Rights, J. D., & Sterba, S. K. (2018). A framework of R-squared measures for single-level and multilevel regression mixture models. *Psychological Methods, 23*, 434–457. https://doi.org/10.1037/met0000139

Rights, J. D., & Sterba, S. K. (2019, online). New recommendations for the use of R-squared differences in multilevel model comparisons. *Multivariate Behavioral Research.* https://doi.org/10.1080/00273171.2019.1660605

Rudman, L. A., & Glick, P. (2001). Prescriptive gender stereotypes and backlash toward agentic women. *Journal of Social Issues, 57*, 743–762. https://doi.org/10.1111/0022-4537.00239

Rudman, L. A., & Goodwin, S. A. (2004). Gender differences in automatic in-group bias: Why do women like women more than men like men? *Journal of Personality and Social Psychology, 87*, 494-509. https://doi.org/10.1037/0022-3514.87.4.494

Sahlqvist, S., Song, Y., Bull, F., Adams, E., Preston, J., & Ogilvie, D. (2011). Effect of questionnaire length, personalization and reminder type on response rate to a complex postal survey: Randomized control trial. *BMC Medical Research Methodology, 10,* 87-102. https://doi.org/10.1186/1479-5868-10-87

Seyla, A. S., Rose, J. S., Dierker, L. C., Hedeker, D.H., & Mermelstein, R. J. (2012). A practical guide to calculating Cohen's f^2, a measure of local effect size, from PROC MIXED. *Frontiers in Psychology, 3,* 111. https://doi.org/10.3389.fpsyg.2012.00111

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science, 12*, 1123-1128. https://doi.org/10.1177/1745691617708630

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). London, UK: Sage Publishers.

Sullivan, J., Moss-Racusin, C., Lopez, M., & Williams, K. (2018). Backlash against gender stereotype-violating preschool children, *PLoS ONE, 13,* e0195503. https://doi.org/10.1371/journal.pone.0195503

Wolsiefer, K., Westfall, J., & Judd, C. M. (2016). Modeling stimulus variation in three common implicit attitude tasks. *Behavior Research Methods, 49,* 1210.

Westfall, J., Judd, C. M., & Kenny, D. A. (2015). Replicating studies in which samples of participants respond to samples of stimuli. *Perspectives on Psychological Science, 10,* 390–399. https://doi.org/10.1177/ 1745691614564879

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General, 143*, 2020-2045. https://doi.org/10.1037/xge0000014

Yeager, D. S., Hanselman, P., Walton, G. M., Crosnoe, R., Miller, C., et al., (2019). A focus on heterogeneity reveals where a brief scalable psychological intervention improves adolescents' educational trajectories and where it does not. Retrieved from: https://osf.io/vf2xj/download/?version=4&displayName=Yeager%20et%20al%20National%20Study%203-1-18-2018-03-02T19%3A34%3A43.118Z.pdf

Table 1

Sample Rows of the Dataset in Study 1

Participant ID	Attribute	Attribute ID	Judgment	Target Gender	Attribute Valence	Participant Gender
1	Nurturing	1	100	Female	Positive	Female
1	Warm	2	100	Female	Positive	Female
1	Emotional	3	99	Female	Negative	Female
1	Nagging	4	93	Female	Negative	Female
1	Ambitious	5	73	Female	Positive	Female
1	Independent	6	51	Female	Positive	Female
1	Arrogant	7	19	Female	Negative	Female
1	Aggressive	8	15	Female	Negative	Female
1	Nurturing	1	100	Male	Positive	Female
1	Warm	2	41	Male	Positive	Female
1	Emotional	3	2	Male	Negative	Female
1	Nagging	4	40	Male	Negative	Female
1	Ambitious	5	30	Male	Positive	Female
1	Independent	6	100	Male	Positive	Female
1	Arrogant	7	57	Male	Negative	Female
1	Aggressive	8	70	Male	Negative	Female
2	Nurturing	1	61	Female	Positive	Male
2	Warm	2	21	Female	Positive	Male
2	Emotional	3	69	Female	Negative	Male
2	Nagging	4	41	Female	Negative	Male
2	Ambitious	5	44	Female	Positive	Male
…..	…		…	…	…	…
234	Aggressive	8	70	Male	Negative	Male

Table 2

Target Ratings as a Function of Participant Gender, Target Gender, and Attribute Valence in Study 1

	Female Participants (n = 177)				Male Participants (n = 57)			
	Target Women		Target Men		Target Women		Target Men	
	M	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
Positive Attributes	72.37	5.33	56.01	6.32	68.27	5.79	56.64	6.72
Negative Attributes	56.40	5.34	53.96	6.32	57.61	5.61	53.64	6.73

Table 3

Parameter Estimates and Variance Components for the LMEM in Study 1

Fixed Effects	*b*	*SE*	*F*	*df*	*p*	r^2
Intercept	59.04	2.26	681.85	7.96	<.001	
Target Gender	3.90	3.85	1.02	6.03	.350	.15
Attribute Valence	3.42	2.14	2.54	6.40	.159	.28
Participant Gender	0.64	0.91	0.51	108.90	.479	.00
Target Gender X Attribute Valence	1.92	3.86	0.25	6.04	.637	.04
Target Gender X Participant Gender	0.80	0.54	2.19	7.75	.178	.22
Attribute Valence X Participant Gender	1.09	0.54	4.05	22.32	.056	.15
Target Gender X Attribute Valence X Participant Gender	1.56	0.55	7.95	8.58	.021	.48

Variance Components	σ^2	*SD*
Participant		
Intercept	113.07	10.63
Target Gender	6.04	2.46
Attribute Valence	24.51	4.95
Target Gender X Attribute Valence	8.64	2.94
Attribute		
Intercept	34.68	5.89
Target Gender	117.71	10.85
Participant Gender	0.35	0.59
Target Gender X Participant Gender	1.23	1.11
Residual	263.81	16.24

Note. $N_{observations}$ = 3614, $N_{participants}$ = 232, $N_{attributes}$ = 8.

Table 4

Judgments of Valence as a Function of Participant Gender, Target Gender, and Target Role in Study 2

	Female Participants (n = 121)				Male Participants (n = 120)			
	Target Women		Target Men		Target Women		Target Men	
	M	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
Feminine Roles	4.70	0.09	4.72	0.10	4.70	0.09	4.62	0.10
Masculine Roles	4.71	0.10	4.42	0.09	4.71	0.10	4.45	0.09

Table 5

Parameter Estimates and Variance Components for the LMEM model in Study 2

Fixed Effects	*b*	*SE*	*F*	*df*	*p*	r^2
Intercept	4.63	0.06	6955.39	681.43	<.001	
Target Gender	0.08	0.02	13.94	243.42	<.001	.05
Target Role	0.06	0.03	4.60	385.06	.032	.01
Participant Gender	0.01	0.04	0.03	219.85	.863	.00
Target Gender X Target Role	-0.06	0.03	4.09	211.05	.044	.02
Target Gender X Participant Gender	-0.01	0.02	0.25	227.78	.615	.00
Target Role X Participant Gender	0.02	0.02	0.59	218.70	.445	.00
Target Gender X Target Role X Participant Gender	-0.02	0.03	0.33	202.45	.566	.00

Variance Components	σ^2	*SD*
Participant		
Intercept	0.34	0.58
Target Gender	0.04	0.20
Target Role	0.06	0.24
Target Gender X Target Role	0.15	0.87
Attribute		
Intercept	2.27	1.51
Residual	0.75	0.87

Note. $N_{observations}$ = 5699, $N_{participants}$ = 241, $N_{attributes}$ = 1861..

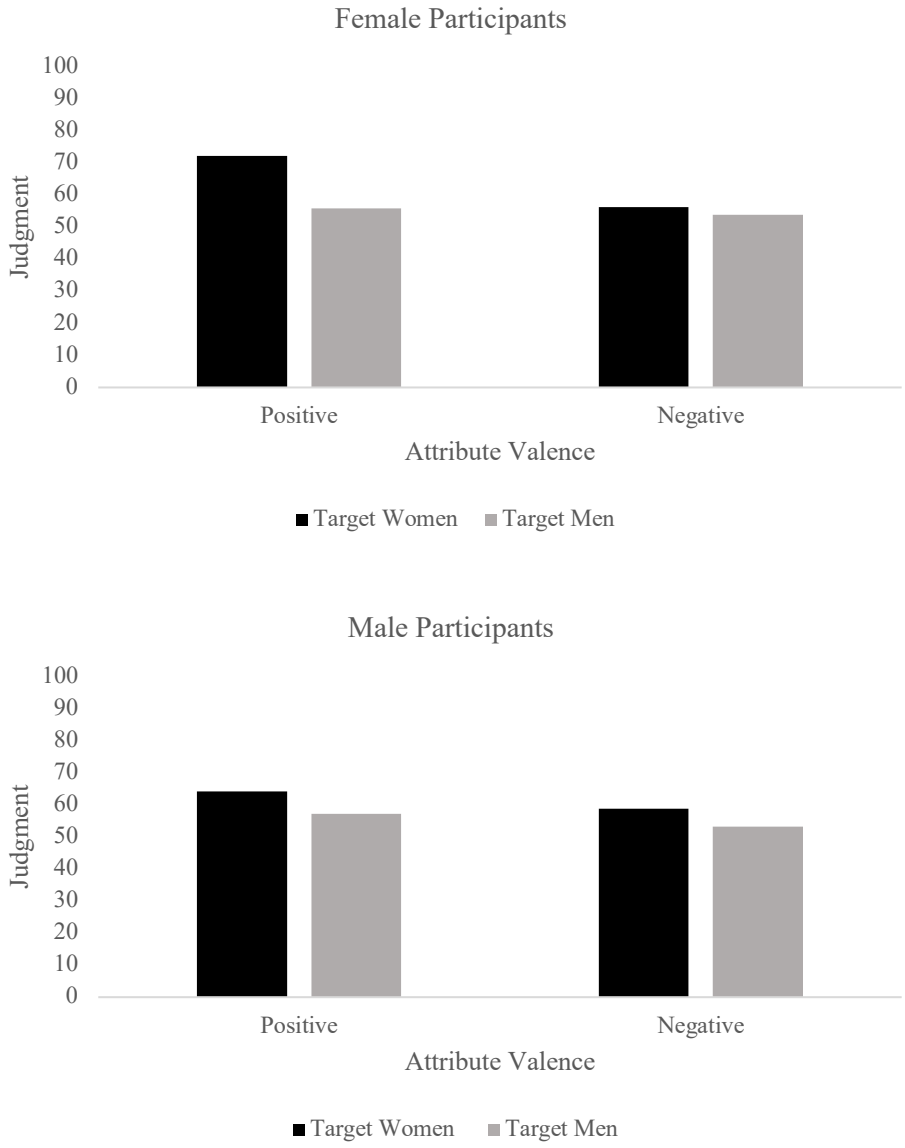

Figure 1. Target ratings as a function of Target Gender, Attribute Valence, and Participant Gender in Study 1. Female participants' judgments are presented in the top graph. Male participants' judgments are presented in the bottom graph. Women (vs. men) ascribed more positive (vs. negative) attributes to target women as compared to target men.

Figure 2. Judgments of valence as a function of Target Role and Target Gender in Study 2. Participants judged target women more positively than target men, although difference only emerged for targets in masculine (vs. feminine) roles.