

# Are Androgynous People More Creative Than Gender Conforming People?

Tingshu Liu and Rodica Ioana Damian  
Department of Psychology, University of Houston

Psychological androgyny refers to possessing both masculine and feminine characteristics. Sandra Bem (1974) proposed that androgynous people are more creative, because they are less limited by gender boundaries. This so-called androgyny-creativity effect contributes to the gender equality movement by ameliorating stereotypes about people who stepped out of gender boundaries. However, the evidentiary value of the available research testing this hypothesis has been limited by suboptimal (by current standards) methodology, such as small samples, antiquated statistical analysis, and inconsistent measurement. The current study attempted to replicate the androgyny-creativity effect in a large sample ( $N = 672$ ), with both self-report and behavioral measures of creativity, and following both original and optimized statistical analyses. We found that androgynous group reported themselves to be more creative than the gender conforming group, but they did not score higher than the latter on behavioral creativity. This suggests that the androgyny-creativity effect (a) could be just a popular lay theory, (b) might only hold for certain types of creativity, and (c) might be a true effect but no longer exist due to societal changes in gender roles.

**Keywords:** gender stereotypes, creativity, androgyny, masculinity, femininity

**Supplemental materials:** <https://doi.org/10.1037/aca0000536.supp>

French writer De Beauvoir (1949) believes that women are oppressed by men and viewed as *the second sex* in society, and this leads to long-term, irreversible gender inequality. Unfortunately, this issue persists today. For example, women are still doing more household chores than men (Cerrato & Cifre, 2018) and gender gaps in productivity and impact in academia have been increasing (Huang et al., 2020). Gender stereotypes do not only reflect the existing gender gaps but also play an important role in reinforcing them. One of the major aspirations of third-wave feminism is to abolish gender stereotypes (Heywood & Drake, 1997), which encourages individuals to behave in ways they prefer; instead of rigidly steering them to behave like their physiological gender. However, gender stereotypes are still pervasive in today's society and they largely impact how people think and behave (Ellemers, 2018). A large meta-analysis study showed that, from the mid-20th century to very recently, the communion stereotype about women has increased and the agency stereotype about men has stayed the same (Eagly et al., 2020).


Counter to the gender stereotypes, Sandra Bem (1974) proposed the concept of androgyny (i.e., those who have both masculine and feminine characteristics) and suggested that androgynous individuals might be more creative—that will be referred to as the androgyny-


creativity effect for convenience. Her work on androgyny was important in the psychological study of gender because she reframed gender-specific traits as malleable, socially constructed beliefs, instead of fixed, innate traits (Dean & Tate, 2017). Therefore, Bem's work served and still serves as an impetus toward gender equality. Additionally, the androgyny-creativity effect fights the stereotype threat or the penalty for crossing the gender line, and encourages more flexible gender roles. Despite the theoretical and practical importance of the androgyny-creativity effect, previous empirical tests of this effect have been riddled with methodological issues, including small samples, antiquated statistical practices, and inconsistent measurement. Thus, given previous empirical evidence, it was unclear if the androgyny-creativity effect was a real-world effect on behavior or just a lay theory. Recent research challenged the lay theory that masculine people are more creative (Proudfoot et al., 2015), so stronger empirical evidence for the androgyny-creativity effect will further clarify this line of research. Therefore, the current study aimed to replicate this effect using a more robust methodology and to respond to the call for conducting more replication studies in the field.

## Gender Stereotypes

Ellemers (2018) referred to gender stereotypes as general expectations about members of a particular gender. Specifically, people generally expect men to behave in masculine ways and women to behave in feminine ways. For example, men are supposedly brave, independent, and insensitive, while women are believed to be tender, dependent, and emotional. Men are expected to protect their family members and provide financial support,

---

Tingshu Liu  <https://orcid.org/0000-0003-4343-9876>

Rodica Ioana Damian  <https://orcid.org/0000-0002-0046-0627>

Correspondence concerning this article should be addressed to Tingshu Liu, Department of Psychology, University of Houston, 3695 Cullen Boulevard, Houston, TX 77004, United States. Email: [tliu31@cougarnet.uh.edu](mailto:tliu31@cougarnet.uh.edu)

while women are expected to raise children and take care of the household. For the current study, gender stereotypes are defined as individuals' beliefs and biases regarding two traditional genders (male and female), including their typical personal characteristics, attitude, behavior, and cognition. Among various stereotypes, gender stereotypes are some of the most influential, fundamental beliefs (Ellemers, 2018), which derive from a complex set of biological, evolutionary, social, and cultural drives. Biologically, the two genders have differences in physical appearance, physiological structure, hormones, and strength (Martin & Parker, 1995); evolutionarily, men adapted to compete with other men over mates and women adapted to choose their mates (Geary, 1999); socially, traditional gender roles of working husbands and housewives are deeply rooted in our minds (Koenig & Eagly, 2014; Martin & Parker, 1995; Sato, 2022); culturally, in most parts of the modern world, men inherit family names and usually properties of the family or group (Geary, 1999; Zanette & Manrubia, 2001). Therefore, gender stereotypes exist in various aspects of our lives and leave a deep influence on our society.

Individuals who violate gender stereotypes are likely to be socially, emotionally, and/or financially punished, and this is true for both genders (Glick et al., 2007; Williams & Tiedens, 2016). For example, a young boy who cries might be told "man up, do not cry like a girl!"; or, a woman might get the advice to spend more time with her family than developing her own career. Gender stereotypes are mentioned everywhere to an extent that it becomes automatic, as a result, our thoughts and behaviors are steered by gender stereotypes even under no social pressure. The traditional bipolar model of Masculinity–Femininity asserts that masculinity and femininity are very different (e.g., see debates in Constantinople, 1973), so it is improper for one to exhibit characteristics that "belong to" the other gender. In other words, gender stereotypes impose expectations for every male to behave like a perfect masculine man and for every female to behave like a perfect feminine woman. In reality, however, individuals possess both masculine and feminine characteristics, in various combinations, on a continuum, as opposed to a fixed level (Leszczynski, 2009).

Gender stereotypes limit the development of flexible gender roles and creativity. When adults purposefully alter children's gender expressions to match their physiological gender, children might lose the chance to explore a wide range of possibilities and in turn fail to realize their true gender self or true ability potential (Ehrensaft, 2012). For example, a girlish boy might not be allowed to play with dolls, dress like a girl, or show his vulnerability, so he would lose the chance to engage in events to explore his feminine gender roles. Instead of proudly identifying himself as a feminine man and reconciling with himself when he grows up, he could be left in continuous struggles to get himself closer to the expected masculine roles. For another example, a girl who is talented in mathematics might face discouraging feedback and repetitive suggestions to quit her math career and choose something else because they *suit a woman better*; as a result, instead of growing up to be a Fields Medal winner, she might end up in an unrelated career. Moreover, research (Gołowska et al., 2013) also showed that when people are instructed to merely think about counterstereotypes, such as a strong woman, they scored higher on flexibility (i.e., measured by a task that asks participants to generate new names for pasta) and had better creative performance (i.e., new ideas for a themed night were developed in the form of both

writing and poster and evaluated by judges). Other researchers (Jönsson & Carlsson, 2000) argued that when individuals are less limited by gender stereotypes, they have more access to problem-solving strategies. This might be one explanation for the finding that bisexual people were more creative than heterosexual and gay or lesbian people (Ben-Zeev et al., 2012; Konik & Crawford, 2004). In other words, crossing the boundaries of traditional gender roles entails greater flexibility.

## Androgyny and Creativity

Along with various practical efforts to break gender stereotypes, Bem (1974) brought up the concept of psychological androgyny to describe individuals who are at the same time masculine and feminine. She (Bem & Lewis, 1975) proposed that androgynous people are more flexible because they are less limited by thinking or behaving in gender-specific ways. Martin and colleagues (2017) argued that, although identifying with one's own gender is important for mental health, being able to identify with the other gender provides additional social benefits and may boost flexibility. In line with this idea, Torrance (1995) argued that creativity requires both independence and sensitivity—a combination of masculine and feminine characteristics. Bem's proposed effect was later extended to the positive relationship between androgyny and one's overall creativity (e.g., Harrington & Anderson, 1981), which will be named the androgyny-creativity effect in the current paper for the convenience of description. Further, androgyny might also be related to fluency and originality. Androgynous people might be more likely to come up with original ideas or solutions by combining thinking styles or strategies that are typical of the other gender (Martin et al., 2017). Although Bem did not specifically mention fluency in her theoretical reasoning, the positive relationship between androgyny and ideational fluency was also found in girls (Hargreaves et al., 1981). Therefore, androgyny could possibly be related to different components of creativity, namely flexibility, originality, and fluency. Historically, researchers have been interested in whether one gender was more creative than the other, but the research has demonstrated mixed results (Baer & Kaufman, 2008). Moreover, there is a bias in the view that more masculine people are more creative (e.g., Harrington & Anderson, 1981), because Proudfoot and colleagues (2015) found that participants judged a product as more creative only because they were told that the product was from men (vs. women). If the androgyny-creativity effect were a replicable, robust effect, it might serve as one explanation for the mixed results from the research on gender and creativity. Specifically, if greater creativity were associated with high androgyny, then any analyses looking at the link between gender and creativity would depend on the ratio of androgynous people in each group, leading to unstable effects—especially when the sample size is not large enough.

Following Bem's line of thinking, several groups of researchers tested the androgyny-creativity effect with the employment of Bem's Sex Role Inventory (Bem, 1974). First, Jönsson and Carlsson (2000) measured creativity by the Creative Functioning Test (Smith & Carlsson, 1990), where students' ( $N = 163$ ) inhibition of subjective interpretations of ambiguous pictures were recorded in their lab. In this measure, after participants recognize the motif of an ambiguous picture, the same picture is shown to the participants in a series of decreasing exposure times, while they do not know it is the same

picture all the time. As the ambiguity of the stimulus increases, less creative participants tend to inhibit their subjective interpretation because of the previous objective recognition, so they scored lower on creativity. The researchers found that androgynous people scored higher on this test than gender-conforming people. Second, Norlander et al. (2000) replicated this result ( $N = 200$ ) using an elaboration measure (Modeus et al., 1987). This measure provides participants with nine squares of incomplete pictures and asks them to finish the pictures in 15 min, where the amount of detail of each picture is evaluated by judges. Though, it was not replicated with their 20-item creativity attitude measure, which asked participants' thoughts on what traits help creativity (e.g., risk-taking). Third, Keller et al. (2007) used the Creativity Styles Questionnaire-Revised (Kumar & Holman, 1997), a self-report measure that contains everyday creativity questions in six domains. They also replicated the same result on students ( $N = 358$ ) in three out of six domains (i.e., Self-Perceived Creative Capacity, Use of Techniques, and Use of Other People).

Although the above findings provided preliminary support for the androgyny-creativity effect, the studies had some methodological limitations. First, the sample sizes used were relatively small. This is important because small sample sizes, or underpowered studies, can produce unstable effects and widely ranging  $p$ -values, which can render results unreliable and irreproducible due to Type I errors (Cumming, 2014). Second, the use of the median-split (to differentiate androgynous from gender-conforming individuals) is not ideal because it treats people as if they belonged to neat categories rather than on a continuum; therefore, losing meaningful variability in the data. Third, the above studies did not use consistent or widely accepted measurements of creativity. Bem and Lewis (1975) did not limit the type of creativity related to androgyny to any one specific type; thus, the researchers who tested her theory empirically used various measurements. Specifically, as mentioned earlier, Jönsson and Carlsson (2000) used Creative Functioning Test (Smith & Carlsson, 1990), which was designed for studying people's creative visual perception; Norlander and colleagues (2000) used a behavioral creativity measure (Modeus et al., 1987) which arguably measures elaboration better than creativity, and elaboration was found to have the lowest factor loading (Auzmendi et al., 1996) among the four creativity indicators (i.e., fluency, flexibility, originality, and elaboration in Guilford, 1956); whereas Keller and colleagues (2007) used a self-report measure of creativity.

Bem's Sex Role Inventory (BSRI) is the most widely used scale to measure gender roles (Hoffman & Borders, 2001) and it is still studied (e.g., Donnelly & Twenge, 2017; Keener & Mehta, 2017; Lips, 2017; Martin et al., 2017) and used frequently in recent literature (e.g., Eggenberger et al., 2021; Mobasser et al., 2022; Lopez-Fernandez et al., 2019; Saint-Michel, 2018). The consistent use of this measure also enables the investigation of long-term value changes in gender roles (e.g., Donnelly & Twenge, 2017), making it difficult to find a substitute. Although some researchers (Hoffman & Borders, 2001) argued that the BSRI items do not represent the masculinity or femininity of the current society anymore, some (Choi et al., 2008) found most masculinity or femininity items to still be rated as highly desirable for a typical man or woman, while other researchers (Donnelly & Twenge, 2017) found that BSRI's

longitudinal results (i.e., men's masculinity and femininity and women's femininity) remained relatively stable since 1974.

The present study was designed to test the androgyny-creativity effect while addressing some of the previous limitations. Specifically, we (a) used a larger and more diverse participant sample, (b) included both self-report and behavioral measures of creativity (the latter was the Unusual Uses Test, a widely used measure of creativity; Guilford, 1956) to exclude the possibility that the effect is only a lay theory (i.e., in the mind of the perceivers self-reporting on both their creativity and androgyny), and (c) conducted a planned analysis that used masculinity and femininity as continuous variables in addition to the media-split analyses that were meant to serve as replications of prior statistical procedures used.

Based on prior evidence, we preregistered the successful replication standard (preregistration can be found here <https://osf.io/g2x8k>). Specifically, to count as a successful replication, the creativity of the androgynous group had to be higher than the gender-conforming group, with an effect size  $d \geq .27$ . Therefore, our two hypotheses were:

*Hypothesis 1:* Androgynous individuals (i.e., people who scored high on both masculinity and femininity scales) would self-report to be more creative than gender-conforming individuals (i.e., people who scored high on one gender-typical scale but score low on the other).

*Hypothesis 2:* Androgynous individuals would score higher on the behavioral creativity measure than gender-conforming individuals.

We did not preregister the replication standard for the analyses using continuous predictors because none of the prior research included them, but we expected these analyses to provide consistent results with our median-split analyses.

## Method

The current study was preregistered via Open Science Framework (OSF, 2020) and can be found here: [https://osf.io/g2x8k?view\\_only=1](https://osf.io/g2x8k?view_only=1)

## Participants

To decide on the appropriate sample size, we conducted an a-priori power analysis, which we preregistered. Prior research (Jönsson & Carlsson, 2000; Norlander et al., 2000) estimated our effect of interest to range between  $r = .27$  and  $.71$ . Based on this

<sup>1</sup> Three minor deviations from the preregistration: First, we had to exclude 33 participants who provided meaningless answers to the behavioral creativity question and one participant who left all creativity questions blank. Therefore, we planned to recruit 750 participants in the preregistration but were only able to get 746. Second, in addition to the planned analyses comparing the androgynous group with the gender conforming group (that included both high masculinity/low femininity and high femininity/low masculinity groups), we conducted an exploratory analysis investigating whether the androgynous group only scored higher on self-report creativity than either high masculinity/low femininity or high femininity/low masculinity group. Third, when we conducted the planned one-side  $t$  test for behavioral creativity with the median split, we got an unexpected very large effect in the opposite direction, so we added a  $t$  test in the opposite direction after adjusting the  $\alpha$  for additional nonblinded testing.

prior work, we concluded that the effect size of interest for a replication attempt should not be lower than  $r = .27$ . Based on the result from G\*Power, when comparing means between two independent groups we needed a total sample of 707 to achieve .9 power to detect an effect of .27 at the  $\alpha$  level of .05.<sup>2</sup>

We recruited 746 participants from Amazon Mechanical Turk (MTurk) and the final sample consisted of 672 people (see Data Cleaning section below for the data cleaning procedure). Of these participants, 342 identified themselves as female (50.9%). The average age was 40.65 years (ranged from 18 to 78, standard deviation ( $SD$ ) = 13.13). The majority of the participants reported their race/ethnicity as White/European American (70.7%), with the rest of the participants being Black/African American (10.9%), Asian/Asian American (9.8%), Latino/Hispanic (4.3%), Multiracial (3.6%), Native American/American Indian (.4%), and Other (.3%). Comparing with the 2020 US Census Bureau data (see Table 1 for details; U.S. Census Bureau, 2022), the compositions of age and gender in our sample were representative; most compositions of race/ethnicity were representative, except for that our sample contained more White/European and less Latino/Hispanic participants compared with Census data. Although MTurk samples do not perfectly match the descriptive characteristics of the population, some of their demographics are indeed representative, such as gender and education level (Berinsky et al., 2012). Therefore, the MTurk is a cost-efficient option by providing an acceptable representation of key demographic aspects at a low cost.

## Procedure and Measures

After consenting to participate (as per protocol HRP-503, approved by the relevant Internal Review Board), the participants answered some demographic questions, and completed a self-report creativity scale, a behavioral creativity task, and a gender self-stereotype scale in the stated order. Two attention check questions were included in the self-report creativity scale and gender self-stereotype scale, respectively.

## Demographics

Participants reported the year they were born, their sex (“What sex was originally listed on your birth certificate?”), gender (Do you think of yourself as: male, female, transgender man/trans man/female-to-male, transgender woman/trans woman/male-to-female, genderqueer/gender nonconforming neither exclusively male nor

**Table 1**  
*Descriptive of Our Sample Versus 2020 U.S. Census Bureau (N = 672)*

Descriptives	Our sample	Census
Median age	38	38.2
Female	50.9%	50.8%
White/European	70.7%	60.1%
Black/African American	10.9%	13.4%
Asian/Asian American	9.8%	5.9%
Latino/Hispanic	4.3%	18.5%
Multiracial	3.6%	2.8%
Native American/American Indian	0.4%	1.5%
Other	0.3%	N/A

Note. N/A = not available.

female, other; a recommended question by Centers for Disease Control and Prevention, 2021), and race and ethnicity. Age was calculated as 2020 minus the year they were born, and gender was coded as male = 1 and female = 2. We used gender in the analyses instead of sex (the two variables included the same data with one exception, a trans man, whom we included in the data analysis as a man). Race and ethnicity were coded as White/Caucasian = 1, Latino/Hispanic = 2, Native American/American Indian = 3, Black/African American = 4, Asian/Asian American = 5, Native Hawaiian/Pacific Islander = 6, Multiracial = 7, and Other = 8.

## Self-Report Creativity

Kaufman (2012) developed the Kaufman Domains of Creativity Scale (KDOCS) to measure everyday creativity in five different domains: Self/Everyday, Scholarly, Performance (encompassing writing and music), Mechanical/Scientific, and Artistic domain. The scale consists of 50 items with 10 items under each domain. Participants were asked to rate their degree of creativity compared with others on a 5-point Likert scale, ranging from 1 = *much less creative* to 5 = *much more creative*. The instruction was “Compared to people of approximately your age and life experience, how creative would you rate yourself for each of the following acts? For acts that you have not specifically done, estimate your creative potential based on your performance on similar tasks.” A sample item is “Finding something fun to do when I have no money.” The internal consistency for this scale in this study was .94. The mean score of the 50 items was used to calculate self-report creativity. Higher score stands for higher self-report creativity. The first attention check question was included in this scale and it asked the participants to “answer much less creative for this item.”

## Behavioral Creativity

Guilford (1956) developed the unusual uses task to measure individuals’ creativity, and the brick task was used in this study. The participants were asked to list as many creative uses as they can think of for a brick in 2 min, and “refrain from listing typical uses or uses that are virtually impossible.” The data were randomly split into two equal parts, and four independent raters were assigned into two groups that took half of the data each. They computed and rated three dimensions of creativity: ideational fluency, flexibility, and originality. Ideational fluency was computed as the total number of valid uses a participant generated; flexibility was computed as the number of different categories their generated uses belong to; originality was rated with the Consensual Assessment Technique (CAT; Amabile, 1982), in which two independent raters evaluated how original the uses were on a 10-point scale, ranging from 1 = *not at all original* to 10 = *highly original*, and their scores were averaged to form the originality score for each use. We summed up the originality scores for all the uses a participant came up with and this was the originality score. The

<sup>2</sup> According to the line of research we attempted to replicate (Jönsson & Carlsson, 2000; Keller et al., 2007; Norlander et al., 2000), the ratio of androgynous group, gender conforming group, and in-differentiated group is about 1:2:1 in the population. Because in-differentiated individuals will not be included in the  $t$  test analysis, we first used the allocation rate of 2:1 in G\*Power to get the sum of androgynous group and gender conforming group ( $N = 530$ ), and then calculated the total sample ( $N = 707$ ) based on the ratio.

interrater reliability for the overall behavioral creativity measure was satisfying and consistent with the reliability observed in prior work,  $\alpha = .94$ , especially given the subjective ratings of originality typically vary widely. Interrater reliability and intraclass correlation coefficient (ICC) were satisfying for each of the three dimensions of behavioral creativity: for fluency,  $\alpha = .98$ , ICC = .95; for flexibility,  $\alpha = .94$ , ICC = .88; for originality,  $\alpha = .88$ , ICC = .79.<sup>3</sup> The scores of the three dimensions were z-scored and then averaged, which formed the final behavioral creativity score. Higher score stands for higher behavioral creativity.

### *Gender Self-Stereotypes (Masculinity and Femininity)*

Bem (1974) developed the BSRI to measure the extent to which one is masculine and also the extent to which one is feminine. The scale consists of 60 items, with 20 measuring masculinity, 20 measuring femininity, and the other 20 measuring social desirability. Only the 40 items for masculinity and femininity were used in this study and order of these items was randomized. Participants responded to how well each of the following characteristics describes them on a 7-point Likert scale, ranging from 1 = *never or almost never true* to 7 = *always or almost always true*. A sample item for masculinity is “dominant,” a sample item for femininity is “tender.” The internal consistency for masculinity in this study was .90 and for femininity was .87. The mean score of the first 20 items was used to calculate masculinity and the mean score of the second 20 items was used to calculate femininity. Higher score indicates higher self-report masculinity or femininity. The second attention check question was included in this scale, and it asked participants to “select three for this item.”

### **Data Analysis**

The analysis in this study was conducted using R 3.6.1 (R Core Team, 2020).

### *Data Cleaning*

A total of 746 participants completed our questionnaire (we collected slightly more participants than the power analysis required because we expected some participants to provide low-quality data and be excluded before any analyses). Indeed, we had to exclude 33 participants before any analyses due to their failure to engage with the behavioral creativity task in any meaningful way (e.g., they copy-pasted nonsensical content from the Internet, wrote “good” or “nice,” or gibberish). Then, following the preregistration, we excluded eight individuals who did not provide any information regarding their sex or gender (that was required for the gender self-stereotypes measure we used, following prior research), 30 individuals who did not pass at least one of the two attention checks, and two outliers who scored outside of 3 *SDs* for self-reported creativity. We also excluded one participant who did not answer any creativity questions. The remaining 672 participants were included in the final sample.

The final sample was categorized by median split. This was done only for the sake of replicating the analyses used by previous research as mentioned above, but we also conducted separate analyses with masculinity and femininity as continuous variables. The median for masculinity was 4.6, and the median for femininity

was 4.6 (on scales from 1–7). Among the participants, 335 were categorized as low masculinity, and 337 were categorized as high masculinity; 336 were categorized as low femininity, and 336 were categorized as high femininity. There were 192 individuals in the androgyny group (high on both variables), and 289 individuals (144 female-typed) in gender-conforming group (high on one and low on the other). The R packages used in data cleaning were the *VIM* (v4.8.0; Kowarik & Templ, 2016), the *mosaic* (v1.7.0; Pruim et al., 2017), the *mice* (v3.6.0; van Buuren & Groothuis-Oudshoorn, 2011), the *jmv* (v0.9.6.1; Selker et al., 2021), and the *psych* (v1.8.12; Revelle, 2022).

### *The Analysis*

The assumptions of normality, homoscedasticity, and multicollinearity were not violated, according to the skew, kurtosis, histograms of the outcome variables, variance inflation factors (VIF), as well as the result of the nonconstant variance score test. We first presented descriptive results and compared self-reported masculinity and femininity across the two genders analyzed (male and female).

To address Hypothesis 1 (i.e., androgynous individuals should score higher on self-report creativity than gender-conforming individuals) and following preregistered analyses (see [osf.io/hgc67](https://osf.io/hgc67)), we used the median split to categorize these two groups of participants (see categorization process in the above Data Cleaning section) and did an independent, one-tailed *t* test between them. A one-tailed *t* test was conducted because, in line with prior work that we were trying to replicate, we hypothesized that the androgynous group would score higher than the gender-conforming group. As specified in the preregistration, we planned to interpret nonstatistically significant results as a failure to replicate the original effect. Following a statistically significant *t* test, we planned to conduct an equivalence test (Lakens et al., 2018) with the R package *TOSTER* (v0.3.4; Lakens, 2017) to test if the effect size observed in our data was consistent with the previously reported effect sizes. We calculated our observed effect size using Cohen’s *d* and the R package *rstatix* (v0.6.0; Kassambara, 2020). The planned equivalence test included two one-sided tests to test against the effect being larger than .27 or smaller than 0. If both one-sided tests were statistically significant, the observed effect would have been considered smaller than the *smallest effect size of interest* (SESOI) determined in our preregistration based on prior work,  $d = .27$ , and we could report a replication failure.

In addition to the above analyses, we also preregistered analyses for one exploratory research question, that is, whether the effect replicated when using continuous variables and interaction effects instead of the median split. We conducted a regression analysis, where self-report

<sup>3</sup> Due to the available research assistant working hours, we randomly divided the data into two portions and assigned them to raters in Groups A and B, respectively, with each group having two raters. To check whether the two groups rated in a similar level of leniency, we compared their results. Fluency for group A ( $M = 6.00$ ) was higher than group B ( $M = 5.47$ ),  $t(657) = 2.63$ ,  $p = .009$ ; also, originality for group A ( $M = 25.26$ ) was higher than group B ( $M = 21.49$ ),  $t(623.34) = 3.99$ ,  $p < .001$ . Flexibility for the two groups was not significantly different,  $t(654) = .70$ ,  $p = .482$ . Therefore, raters in Group A were more lenient than raters in Group B for two out of three dimensions of creativity. The main results for each of the two groups respectively supported the same conclusions as the overall sample, and the details were provided in the online supplemental materials.

creativity was predicted by the continuous variables masculinity, femininity, their interaction, and gender. The R package used for the regression models was the *lavaan* (v0.6-5; Rosseel, 2012), and we used Full Information Maximum Likelihood to account for missing data. The R packages *interactions* (v1.1.0; Long, 2019), *effects* (v4.1-2; Fox & Weisberg, 2019), and *ggplot2* (v3.2.1; Wickham, 2016) were used to conduct the simple slopes analysis and make simple slope figures. Moreover, as requested by an anonymous reviewer, we conducted the same analyses for each of the subdimensions of self-reported creativity.<sup>4</sup> We did not have a priori predictions for these analyses, so we used a Bonferroni correction for six analyses ( $p < .0083$ ), but the results should still be interpreted with caution.

To address Hypothesis 2, that androgynous individuals scored higher on behavioral creativity than gender-conforming individuals, and our exploratory question, whether the effect was consistent across continuous versus median split data, we conducted the same kinds of analyses described above with behavioral creativity (instead of self-reported creativity) as the outcome of interest. Moreover, as requested by an anonymous reviewer, the same analyses were also conducted with each subdimension of behavioral creativity.<sup>5</sup> Similarly, we did not have a priori predictions for these analyses, so we used a Bonferroni correction for four analyses ( $p < .0125$ ), but the results should still be interpreted with caution.<sup>6</sup>

## Results

The descriptive statistics and intercorrelations for the study variables are shown in Table 2. As expected, male participants ( $n_1 = 330$ ,  $M = 4.77$ ,  $SD = .88$ ) reported their masculinity to be significantly higher than female participants ( $n_2 = 342$ ,  $M = 4.38$ ,  $SD = .90$ ),  $t(670) = 5.61$ ,  $p < .001$ ,  $d = .43$ ; female participants reported their femininity ( $n_2 = 342$ ,  $M = 4.77$ ,  $SD = .81$ ) to be significantly higher than male participants ( $n_1 = 330$ ,  $M = 4.41$ ,  $SD = .82$ ),  $t(670) = 5.71$ ,  $p < .001$ ,  $d = .44$ . Notably, the distributions of masculinity and femininity across the two gender identities had a big overlap, as demonstrated in Figure 1 and Figure 2. In other words, masculinity was not a distinct characteristic for men and femininity was not a distinct characteristic for women, which is consistent with prior work and theory (Bem, 1977; Bem & Lewis, 1975; Leszczynski, 2009).

Masculinity (vs. femininity) was more strongly correlated with self-reported creativity ( $r = .53$ ,  $p < .001$  vs.  $r = .27$ ,  $p < .001$ ), though gender identity itself was not significantly correlated with self-reported creativity ( $r = -.06$ ,  $p = .141$ ). Women ( $M = 3.17$ ) self-reported lower creativity than men ( $M = 3.24$ ), but they ( $M = .13$ , standardized) scored higher than men ( $M = -.14$ , standardized) on behavioral creativity ( $r = .15$ ,  $p < .001$ , for variable gender, male = 1 and female = 2), which we discuss in the next section. Notably, the correlation between self-reported and behavioral creativity was not statistically significant ( $r = .07$ ,  $p = .065$ ), which suggests that the two measures assessed distinct constructs.

### Analysis for Self-Report Creativity

#### With Median Split

We did an independent, one-tailed  $t$  test on creativity between the androgyny group and the gender-conforming group, but the Levene's test suggested that the assumption of the equal variance

was violated,  $F(1, 479) = 8.47$ ,  $p = .004$ . Therefore, we used Welch's  $t$  test instead, and the androgyny group ( $n_1 = 192$ ,  $M = .64$ ,  $SD = .99$ ) reported a significantly higher creativity than the gender conforming group ( $n_2 = 289$ ,  $M = -.09$ ,  $SD = .84$ ),  $t(363.19) = 8.33$ ,  $p < .001$ ,  $d = .79$ .

Given the  $t$  test was statistically significant, we continued to conduct two one-sided  $t$  tests to see whether the effect was larger than the *smallest effect size of interest* (SESOI) determined in our preregistration, .27 (see Lakens et al., 2018, for an introduction to the equivalence testing). The result of the equivalence test was nonsignificant,  $t(362.42) = 5.48$ ,  $p = 1.000$ , given equivalence bounds of .000 and .248 (on a raw scale) and an  $\alpha$  of .05. Therefore, combining the results of Welch's  $t$  test and the equivalence test, the statistical equivalence was not concluded. Instead, the effect that androgynous individuals scored higher on self-report creativity than gender-conforming individuals was statistically larger than  $d = .27$ . Therefore, Hypothesis 1 was supported, indicating that androgynous individuals self-reported significantly higher creativity than gender-conforming individuals (though the effect size was significantly larger than what previous literature had indicated).

To further explore the pattern, we also tested the androgynous group against male-typed and female-typed groups separately. Two independent, one-tailed  $t$  test on creativity were conducted, and results showed that androgyny group ( $n_1 = 192$ ,  $M = .64$ ,  $SD = .99$ ) reported a significantly higher creativity than the male-typed group ( $n_3 = 145$ ,  $M = .12$ ,  $SD = .89$ ),  $t(335) = 4.93$ ,  $p < .001$ ,  $d = .54$ , as well as the female-typed group ( $n_4 = 144$ ,  $M = -.30$ ,  $SD = .73$ ),  $t(333.96) = 9.94$ ,  $p < .001$ ,  $d = 1.07$ . Therefore, we excluded the possibility that the significant result was only due to one subgroup (male- or female-typed).

#### Without Median Split

Median split is not an ideal statistical procedure, because it drops part of the valid data, so we also conducted regression analysis with masculinity and femininity as continuous factors. Full Information Maximum Likelihood (FIML) was used to deal with missing data. As presented in Table 3, Model 1 that predicts creativity with masculinity and femininity was tested, and both predictors were significant. For masculinity,  $\beta = .50$ ,  $p < .001$ ; for femininity,  $\beta = .17$ ,  $p < .001$ . Model 2 was then tested, where the interaction between masculinity and femininity was added into the model, and the interaction term was significant,  $\beta = .06$ ,  $p = .041$ , masculinity ( $\beta = .50$ ,  $p < .001$ ), and femininity ( $\beta = .16$ ,  $p < .001$ ) were also significant. In Model 3 we added gender, but it did not explain significantly more variance in creativity, so Model 2 was

<sup>4</sup>The subdimensions of self-report creativity include self/everyday creativity, scholar creativity, performance creativity, mechanical/scientific creativity, and artistic creativity. We conducted the  $t$  test that compares subdimension creativity in androgynous and gender-conforming groups, as well as same series of regression models that predict each creativity subdimension. Equivalence testing was not included, because we did not have a preregistered *smallest effect size of interest* (SESOI) for each of the effects.

<sup>5</sup>The subdimensions of behavioral creativity include fluency, flexibility, and originality.

<sup>6</sup>In actual analyses, the  $\alpha$  level was further corrected to .00625, because the results were not significant, and we conducted an additional opposite direction  $t$  test for each subdimension.

**Table 2**  
*Descriptive Statistics and Correlations for the Main Variables (N = 672)*

Variables	<i>M</i>	<i>SD</i>	1	2	3	4	5
1. Self-report creativity	3.20	.60	(.95)				
2. Behavioral creativity	11.22	5.40	.07	(.94)			
3. Masculinity	4.57	.91	.53***	.003	(.90)		
4. Femininity	4.59	.83	.27***	-.04	.20***	(.87)	
5. Gender	.51	.50	-.06	.15***	-.21***	.22***	—

*Note.* Reliability for each scale was included in the parentheses; the means and standard deviations are unstandardized; gender was coded in dummy variables, 0 = male, 1 = female.

\*\*\*  $p < .001$ .

the best fit to the data. Therefore, both masculinity and femininity were positively associated with self-report creativity, where the latter had a weaker effect than the former.

The interaction effect was visualized in Figure 3, where femininity was treated as a moderator. As shown, when femininity was scored 1 *SD* below the mean, the slope of masculinity was significant,  $\beta = .45, p < .001$ ; when femininity was scored the mean, the slope of masculinity was significant,  $\beta = .50, p < .001$ ; when femininity was scored 1 *SD* above the mean, the slope of masculinity was significant,  $\beta = .56, p < .001$ . Moreover, the positive relationship between masculinity and self-report creativity was stronger when femininity was higher. This is congruous with the result from median split analysis on self-report creativity, confirming that people higher (vs. lower) in androgyny self-reported higher creativity (notably, these analyses included all participants, not just high androgyny and gender-conforming individuals, but also, the so-called undifferentiated individuals, that is, the people who scored low on both masculinity and femininity).

### Added Analyses for Subdimensions

As mentioned in the analyses section, the same analyses for subdimensions of self-report creativity (i.e., self/everyday creativity, scholar creativity, performance creativity, mechanical/scientific creativity, and artistic creativity) were requested by an anonymous reviewer, so we did not have them preregistered and the results should be interpreted with caution.<sup>7</sup> We found similar results for *t* tests in all subdimensions (all significant in the same direction as the main result), but not for regression models. Specifically, the interaction between masculinity and femininity was significant for only mechanical/scientific creativity. For gender, men reported themselves to be higher on mechanical/scientific creativity and women reported themselves to be higher on artistic creativity, and there were no gender differences in other subdimensions.

## Analysis for Behavioral Creativity

### With Median Split

We did an independent, one-tailed *t* test on behavioral creativity between androgynous and gender-conforming group, testing whether the former had higher scores than the latter, and the result was not statistically significant,  $t(358.42) = -2.21, p = .986, d = -.21$ . Because, based on the descriptive statistics, the effect appeared to be in the opposite direction than predicted by prior literature, we conducted an additional, one-tailed *t* test in the opposite direction to see whether the gender-conforming group had

higher scores than the androgynous group. Because the additional test was not included in the preregistration and we already knew the nonsignificant results in the opposite direction, the significance threshold was reduced to .025 to account for  $\alpha$  inflation. The result was statistically significant,  $t(358.42) = -2.21, p = .014, d = -.21$ , indicating that gender conforming group (vs. androgynous group) indeed scored higher on behavioral creativity. Therefore, Hypothesis 2, that androgynous individuals scored higher on behavioral creativity than gender-conforming individuals, was not supported.

### Without Median Split

Similarly, we also did regression analysis with masculinity and femininity as continuous factors, due to the limitations of the median split. As presented in Table 4, the best model predicting creativity was the final model with the following predictors: masculinity ( $\beta = .05, p = .193$ ), femininity ( $\beta = -.09, p = .025$ ), interaction between the two ( $\beta = -.08, p = .017$ ), and gender ( $\beta = .33, p < .001$ ). A simple slopes analysis showed that the slope of masculinity when femininity was 1 *SD* below the mean was significant,  $\beta = .13, p = .009$ ; the slope of masculinity when femininity was scored the mean was not significant,  $\beta = .05, p = .180$ ; the slope of masculinity when femininity was 1 *SD* above the mean was not significant,  $\beta = -.03, p = .621$ . Moreover, as shown in Figure 4, the positive association between masculinity and behavioral creativity was weaker for participants who were more feminine. Therefore, this shows a consistent pattern with the result of our median split analysis, suggesting that people higher in androgyny did not score significantly higher on behavioral creativity than people lower in androgyny.

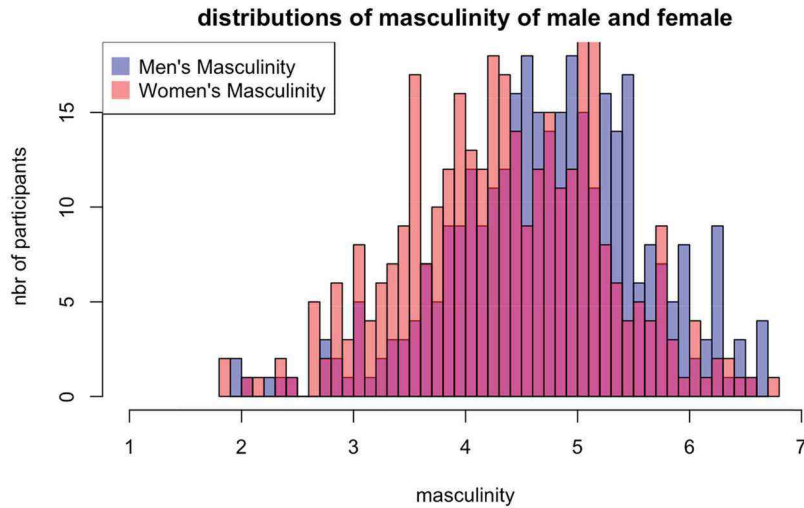
### Added Analyses for Subdimensions

As mentioned in the analyses section, the same analyses for subdimensions of behavioral creativity (i.e., fluency, flexibility, and originality) were requested by the reviewer, so we did not have them preregistered and the results should be interpreted with caution.<sup>8</sup> Not all results of the *t* test were significant, but the effects were in the same direction as the main result. Specifically, the results of flexibility replicated the main results, but that of fluency and originality did not—for these two outcomes, only gender remained as a significant predictor in the regression model (i.e., both favored women).

<sup>7</sup> The full results can be found in Tables 1S and 2S in the online supplemental materials.

<sup>8</sup> The full results can be found in Tables 3S and 4S in online supplemental materials.

**Figure 1**  
*Distribution of Masculinity of Two Genders*



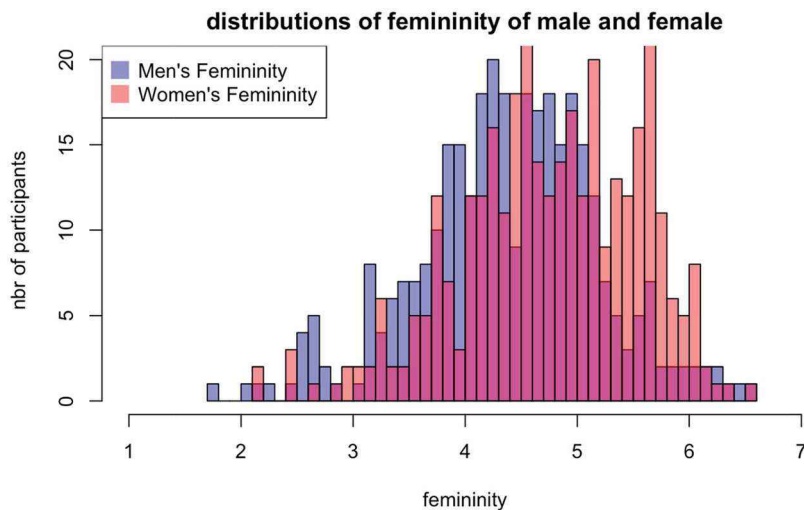
*Note.* See the online article for the color version of this figure.

## Discussion

The current study aimed to replicate the androgyny-creativity effect, namely, people who are high in both masculine and feminine characteristics are more creative (Bem, 1974; Harrington & Anderson, 1981). We investigated whether the effect could be replicated under different conditions: (a) self-reported versus behavioral creativity and (b) categorical (using median split like in the original studies that found the effect) versus continuous treatment of the masculinity and femininity variables. We originally aimed to recruit at least 707 people to achieve .9 power to detect an effect of .27, we only managed to get a sample of 672 after eliminating 30 people who failed attention checks. Overall, we found that regardless of

whether a median split was used in the data analysis, the androgyny-creativity effect replicated (and was even larger than anticipated) for *self-reported* creativity, but *not* for *behavioral* creativity. The results of self-report creativity replicated Keller et al. (2007) but not Norlander et al. (2000), and the results of behavioral creativity did not replicate Norlander et al. (2000) and Jönsson and Carlsson (2000). In other words, androgynous individuals reported themselves to be more creative than gender-conforming individuals, but the former did not score higher than the latter on a creativity performance test. Therefore, our results provided partial support for the hypothesis that androgynous individuals are more creative than gender-conforming individuals.

**Figure 2**  
*Distribution of Femininity of Two Genders*



*Note.* See the online article for the color version of this figure.



**Table 3**  
Regression Models Predicting Self-Report Creativity ( $N = 672$ )

Models	Self-report creativity	
	$\beta$	$R^2$
Model 1		.31
Masculinity	.50*	
Femininity	.17*	
Model 2		.32*
Masculinity	.50*	
Femininity	.16*	
M $\times$ F	.06*	
Model 3		.32
Masculinity	.51*	
Femininity	.16*	
M $\times$ F	.06*	
Gender	.04	

Note. All variables were  $z$ -scored. For gender, 0 = male, 1 = female. M  $\times$  F = interaction between masculinity and femininity.  
\*  $p < .05$ .

One possible explanation for these findings is that the androgyny-creativity effect on self-reported (but not behavioral) creativity is driven by a lay theory. Specifically, if people believed that more androgynous people were more creative and they perceived themselves to be more androgynous following the self-reported masculinity-femininity test, then they might have also subsequently seen themselves as more creative. One possible reason why the androgyny-creativity lay theory might exist is that it might make being androgynous more desirable (given that creativity is seen as a positive characteristic), so it might relieve some of the conflict experienced between one's innate, androgynous gender identity and societal pressures to conform to more stereotypical gender roles (Sullivan et al., 2018; Yousaf et al., 2015). Lay theories aside, given our null findings for behavioral creativity, future research should be cautious toward assuming a positive

**Table 4**  
Regression Models Predicting Behavioral Creativity ( $N = 672$ )

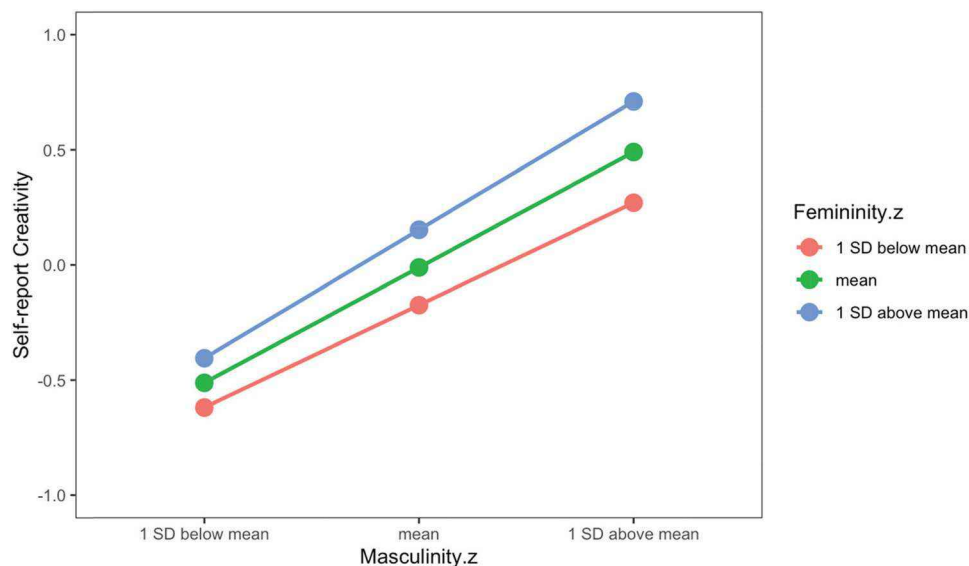
Models	Behavioral creativity	
	$\beta$	$R^2$
Model 1		.00
Masculinity	.02	
Femininity	-.04	
Model 2		.01*
Masculinity	.01	
Femininity	-.04	
M $\times$ F	-.09*	
Model 3		.04*
Masculinity	.05	
Femininity	-.09*	
M $\times$ F	-.08*	
Gender	.33*	

Note. All variables were  $z$ -scored. For gender, 0 = male, 1 = female. M  $\times$  F = interaction between masculinity and femininity.  
\*  $p < .05$ .

relationship between androgyny and creativity in general. This is not to say, however, that androgyny is not desirable for other reasons beyond creativity, including psychological health and well-being (Martin et al., 2017). Further, if this is the case, it would be interesting to explore how and why this lay theory may have developed, although this topic might lie in the intersection of multiple fields, including psychology, gender study, and sociology. The existence of such a lay theory in itself might suggest that flexible gender roles are encouraged by the current social values.

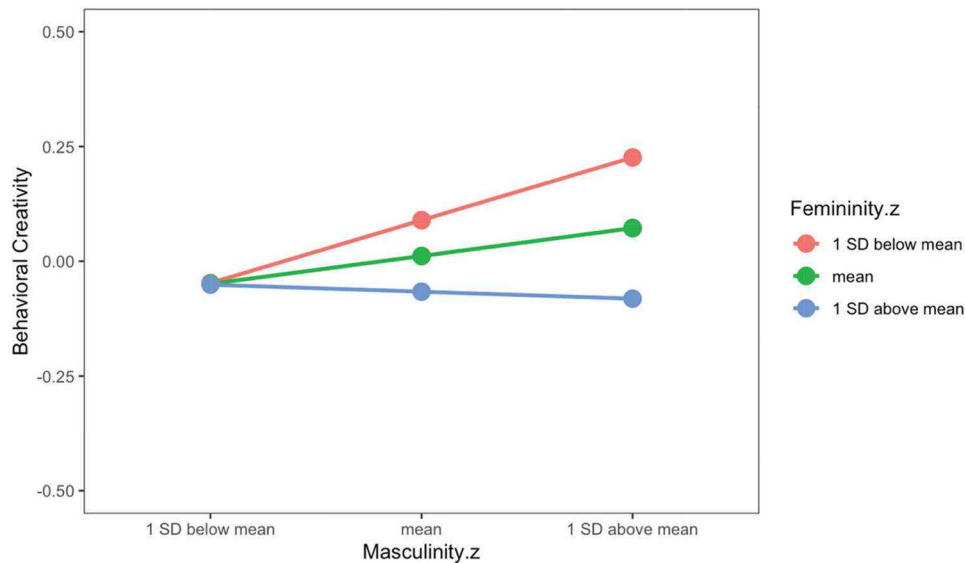
Another possible explanation is that androgyny is only related to certain types of creativity—self-report creativity but not verbal performance creativity. Theoretically, the assessments we included measure two aspects of creativity, so they should be correlated with each other, but the data suggested the opposite ( $r = .07$ ). If the two measures are unrelated, the proposed androgyny-creativity

**Figure 3**  
Simple Slopes on Self-Report Creativity



Note. .z indicates  $z$ -score. See the online article for the color version of this figure.

**Figure 4**  
Simple Slopes on Behavioral Creativity



Note. .z indicates z-score. See the online article for the color version of this figure.

effect deserves a further, refined definition. It is also possible that the lack of correlation was due to issues involved with the measurements themselves. Although the Kaufman-Domains of Creativity Scale (Kaufman, 2012) and Unusual Uses Test (Guilford, 1956) are both widely used assessments, self-report creativity measures suffer from bias and inattentiveness (McKibben & Silvia, 2017; Ng & Feldman, 2012) and divergent thinking tests received criticism for their lack of validity (Said-Metwaly et al., 2017) and ignoring of creativity's role in solving problems (Plucker & Makel, 2010). Anyways, future research should test androgyny's possible association with various types of creativity and sort out which type(s) of creativity is related to androgyny.

A third possible explanation is the influence of social changes. As social values and culture develop, the current understanding of gender roles might be different from that of about 50 years ago when Bem's Sex Role Inventory was invented (Bem, 1974), so the androgyny-creativity effect that existed then might disappear now—but people still believe in it. If this is the case, then the big effect we found on self-report creativity might be the lingering values of the past decades. However, research found that the gender stereotypes did not change a lot from the 1940s to now (Eagly et al., 2020; Haines et al., 2016), with the only exception that women's communion has increased—so the stereotype was even strengthened. Though, to be rigid, future studies should still include women's communion as a moderator to rule out this possibility.

Notably, one additional finding was that although women did not self-report as more creative than men, they scored higher on behavioral creativity. Although most previous studies found no difference between men and women in creativity test scores, a limited number of studies found an advantage for women, which is in line with the present findings (for a review, see Baer & Kaufman, 2008). One possible explanation for this finding is the specific type of creativity measure we used, which involved intense verbalization of the abstract ideas;

indeed, women generally outperform men in verbal ability tasks (Hirnstein et al., 2012).

Moreover, from a methodological point of view, the current study provided the following insight. As stated above, the androgyny-creativity effect was replicated for *self-reported* creativity, but *not* for *behavioral* creativity. Thus, although self-report measures of creativity can provide valuable information (Silvia et al., 2012), creativity researchers should ideally use a multimeasure approach, especially for replication studies. From a practical point of view, the current study provided the following implications. First, people who are both masculine and feminine believe they are more creative. Therefore, there are potential benefits to encouraging the free exploration of gender roles because it might boost self-confidence in creativity in society. Second, although the data only showed a tentative relationship between creativity and androgyny, to build a more creative system, it is still not suggested to restrict people's behaviors and strategy options to one stereotypic gender role that is consistent with their biological sex. Instead, it could be better to allow their free exploration of gender roles, including but not limited to dressing in both masculine and feminine ways, engaging in both activities for men and women, and choosing the career they are truly passionate about.

### Limitation

One limitation of the present study is that we used different measurements of creativity compared with prior studies. This was necessary because prior research did not make measures publicly available, the papers did not contain enough information to replicate the measures, and they did not use widely used or validated measures of behavioral or self-report creativity. Because we did not have the access to measurements in the original studies, we used K-DOCS for self-reported creativity (Kaufman, 2012) and the brick task for behavioral creativity (Guilford, 1956), both widely used and well-validated

measures, which nevertheless measure different aspects of creativity (Snyder et al., 2019). This could be a reason why the results observed for behavioral creativity did not replicate. However, the results observed for self-reported creativity suggested that the effect was in a consistent direction (though larger in the present study) across different measurements under the same construct.

Another limitation is that we only used one self-report creativity and one behavioral creativity measure. Given that previous studies with inconsistent measures presented different results, it might have been better for us to have included multiple measures for each type of creativity to provide a wider test of replicability. Further, the behavioral creativity measure we used (i.e., the Unusual Uses Test) assessed divergent thinking, which does not demonstrate the full picture of creativity that involves preparation, incubation, illumination, and verification (Wallas, 1926). Although it measures participants' actual performance, it only assesses verbal creativity. Therefore, future studies on androgyny should aim to include multiple measures of creativity.

## Conclusion

We conducted a conceptual replication of the androgyny-creativity effect with rigorous preregistration and found an interesting pattern, that prior results for self-reported creativity were replicated but that for behavioral creativity were not. There are several possible explanations for our findings: (a) The androgyny-creativity effect may be the consequence of a lay theory, where androgynous people judge themselves to be more creative, though they do not actually score higher on a verbal creativity test. (b) The androgyny-creativity effect might only stay true for certain types of creativity, such as self-report creativity. (c) Although the possibility is small, the change in gender roles in social value might have changed, so the androgyny-creativity effect might no longer exist in the current society. Future research is needed to further study and clarify these possibilities.

## References

- Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology*, 43(5), 997–1013. <https://doi.org/10.1037/0022-3514.43.5.997>
- Auzmendi, E., Villa, A., & Abedi, J. (1996). Reliability and validity of a newly constructed multiple-choice creativity instrument. *Creativity Research Journal*, 9(1), 89–95. [https://doi.org/10.1207/s15326934crj0901\\_8](https://doi.org/10.1207/s15326934crj0901_8)
- Baer, J., & Kaufman, J. C. (2008). Gender differences in creativity. *The Journal of Creative Behavior*, 42(2), 75–105. <https://doi.org/10.1002/j.2162-6057.2008.tb01289.x>
- Bem, S. L. (1974). The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology*, 42(2), 155–162. <https://doi.org/10.1037/h0036215>
- Bem, S. L. (1977). On the utility of alternative procedures for assessing psychological androgyny. *Journal of Consulting and Clinical Psychology*, 45(2), 196–205. <https://doi.org/10.1037/0022-006X.45.2.196>
- Bem, S. L., & Lewis, S. A. (1975). Sex role adaptability: One consequence of psychological androgyny. *Journal of Personality and Social Psychology*, 31(4), 634–643. <https://doi.org/10.1037/h0077098>
- Ben-Zeev, A., Dennehy, T. C., & Kaufman, J. C. (2012). Blurring boundaries: Bisexual versus lesbian and heterosexual women's self-assessed creativity. *Journal of Bisexuality*, 12(3), 347–359. <https://doi.org/10.1080/15299716.2012.702614>
- Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis*, 20(3), 351–368. <https://doi.org/10.1093/pan/mpr057>
- Centers for Disease Control and Prevention. (2021). *Collecting sexual orientation and gender identity information*. <https://www.cdc.gov/hiv/clincicians/transforming-health/health-care-providers/collecting-sexual-orientation.html>
- Cerrato, J., & Cifre, E. (2018). Gender inequality in household chores and work-family conflict. *Frontiers in Psychology*, 9, 1330. <https://doi.org/10.3389/fpsyg.2018.01330>
- Choi, N., Fuqua, D. R., & Newman, J. L. (2008). The Bem Sex-Role Inventory: Continuing theoretical problems. *Educational and Psychological Measurement*, 68, 881–900. <https://doi.org/10.1177/0013164408315267>
- Constantinople, A. (1973). Masculinity-femininity: An exception to a famous dictum? *Psychological Bulletin*, 80(5), 389–407. <https://doi.org/10.1037/h0035334>
- Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, 25(1), 7–29. <https://doi.org/10.1177/0956797613504966>
- De Beauvoir, S. (1949). *The second sex*. Knopf.
- Dean, M. L., & Tate, C. C. (2017). Extending the legacy of Sandra Bem: Psychological androgyny as a touchstone conceptual advance for the study of gender in psychological science. *Sex Roles*, 76(11–12), 643–654. <https://doi.org/10.1007/s11199-016-0713-z>
- Donnelly, K., & Twenge, J. M. (2017). Masculine and feminine traits on the Bem Sex-Role Inventory, 1993–2012: A cross-temporal meta-analysis. *Sex Roles*, 76(9–10), 556–565. <https://doi.org/10.1007/s11199-016-0625-y>
- Eagly, A. H., Nater, C., Miller, D. I., Kaufmann, M., & Sczesny, S. (2020). Gender stereotypes have changed: A cross-temporal meta-analysis of U.S. public opinion polls from 1946 to 2018. *American Psychologist*, 75(3), 301–315. <https://doi.org/10.1037/amp0000494>
- Eggenberger, L., Fordschmid, C., Ludwig, C., Weber, S., Grub, J., Komlenac, N., & Walther, A. (2021). Men's psychotherapy use, male role norms, and male-typical depression symptoms: Examining 716 men and women experiencing psychological distress. *Behavioral Sciences*, 11(6), 83. <https://doi.org/10.3390/bs11060083>
- Ehrensaft, D. (2012). From gender identity disorder to gender identity creativity: True gender self child therapy. *Journal of Homosexuality*, 59(3), 337–356. <https://doi.org/10.1080/00918369.2012.653303>
- Ellemers, N. (2018). Gender stereotypes. *Annual Review of Psychology*, 69(1), 275–298. <https://doi.org/10.1146/annurev-psych-122216-011719>
- Fox, J., & Weisberg, S. (2019). *An {R} companion to applied regression* (3rd ed.). Sage. <https://socialsciences.mcmaster.ca/jfox/Books/Companion/>
- Geary, D. C. (1999). Evolution and developmental sex differences. *Current Directions in Psychological Science*, 8(4), 115–120. <https://doi.org/10.1111/1467-8721.00027>
- Glick, P., Gangl, C., Gibb, S., Klumpner, S., & Weinberg, E. (2007). Defensive reactions to masculinity threat: More negative affect toward effeminate (but not masculine) gay men. *Sex Roles*, 57(1), 55–59. <https://doi.org/10.1007/s11199-007-9195-3>
- Gocłowska, M. A., Crisp, R. J., & Labuschagne, K. (2013). Can counterstereotypes boost flexible thinking? *Group Processes & Intergroup Relations*, 16(2), 217–231. <https://doi.org/10.1177/1368430212445076>
- Guilford, J. P. (1956). The structure of intellect. *Psychological Bulletin*, 53(4), 267–293. <https://doi.org/10.1037/h0040755>
- 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(3), 353–363. <https://doi.org/10.1177/0361684316634081>
- Hargreaves, D., Stoll, L., Farnworth, S., & Morgan, S. (1981). Psychological androgyny and ideational fluency. *British Journal of Social Psychology*, 20(1), 53–55. <https://doi.org/10.1111/j.2044-8309.1981.tb00473.x>
- Harrington, D. M., & Anderson, S. M. (1981). Creativity, masculinity, femininity, and three models of psychological androgyny. *Journal of*

- Personality and Social Psychology*, 41(4), 744–757. <https://doi.org/10.1037/0022-3514.41.4.744>
- Heywood, L., & Drake, J. (1997). *Third wave agenda: Being feminist, doing feminism*. University of Minnesota Press.
- Hirstein, M., Freund, N., & Hausmann, M. (2012). Gender stereotyping enhances verbal fluency performance in men (and women). *Zeitschrift Für Psychologie Mit Zeitschrift Für Angewandte Psychologie*, 220, 70–77. <https://doi.org/10.1027/2151-2604/a000098>
- Hoffman, R. M., & Borders, L. D. (2001). Twenty-five years after the Bem Sex-Role Inventory: A reassessment and new issues regarding classification variability. *Measurement & Evaluation in Counseling & Development*, 34(1), 39–55. <https://doi.org/10.1080/07481756.2001.12069021>
- Huang, J., Gates, A. J., Sinatra, R., & Barabási, A. L. (2020). Historical comparison of gender inequality in scientific careers across countries and disciplines. *Proceedings of the National Academy of Sciences*, 117, 4609–4616.
- Jönsson, P., & Carlsson, I. (2000). Androgyny and creativity: A study of the relationship between a balanced sex-role and creative functioning. *Scandinavian Journal of Psychology*, 41(4), 269–274. <https://doi.org/10.1111/1467-9450.00198>
- Kassabara, A. (2020). rstatix: Pipe-friendly framework for basic statistical tests (R package version 0.6.0) [Computer software]. <https://CRAN.R-project.org/package=rstatix>
- Kaufman, J. C. (2012). Counting the muses: Development of the Kaufman domains of creativity scale (K-DOCS). *Psychology of Aesthetics, Creativity, and the Arts*, 6(4), 298–308. <https://doi.org/10.1037/a0029751>
- Keener, E., & Mehta, C. (2017). Sandra Bem: Revolutionary and generative feminist psychologist. *Sex Roles*, 76(9–10), 525–528. <https://doi.org/10.1007/s11199-017-0770-y>
- Keller, C. J., Lavish, L. A., & Brown, C. (2007). Creative styles and gender roles in undergraduates students. *Creativity Research Journal*, 19(2–3), 273–280. <https://doi.org/10.1080/10400410701397396>
- 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>
- Konik, J., & Crawford, M. (2004). Exploring normative creativity: Testing the relationship between cognitive flexibility and sexual identity. *Sex Roles*, 51(3/4), 249–253. <https://doi.org/10.1023/B:SERS.0000037885.22789.83>
- Kowarik, A., & Templ, M. (2016). Imputation with the R Package VIM. *Journal of Statistical Software*, 74, 1–16. <https://doi.org/10.18637/jss.v074.i07>
- Kumar, V. K., & Holman, E. R. (1997). *The creativity styles questionnaire-revised* [Unpublished psychological test]. Department of Psychology, West Chester University of Pennsylvania.
- Lakens, D. (2017). Equivalence tests: A practical primer for t-tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8, 355–362. <https://doi.org/10.1177/1948550617697177>
- Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, 1(2), 259–269. <https://doi.org/10.1177/2515245918770963>
- Leszczynski, J. P. (2009). A state conceptualization: Are individuals' masculine and feminine personality traits situationally influenced? *Personality and Individual Differences*, 47(3), 157–162. <https://doi.org/10.1016/j.paid.2009.02.014>
- Lips, H. M. (2017). Sandra Bem: Naming the impact of gendered categories and identities. *Sex Roles*, 76(9–10), 627–632. <https://doi.org/10.1007/s11199-016-0664-4>
- Long, J. A. (2019). interactions: Comprehensive, user-friendly toolkit for probing interactions (R package version 1.1.0) [Computer software]. <https://cran.r-project.org/package=interactions>
- Lopez-Fernandez, O., Williams, A. J., & Kuss, D. J. (2019). Measuring female gaming: Gamer profile, predictors, prevalence, and characteristics from psychological and gender perspectives. *Frontiers in Psychology*, 10, 898. <https://doi.org/10.3389/fpsyg.2019.00898>
- Martin, C. L., Cook, R. E., & Andrews, N. C. Z. (2017). Reviving androgyny: A modern day perspective on flexibility of gender identity and behavior. *Sex Roles*, 76(9–10), 592–603. <https://doi.org/10.1007/s11199-016-0602-5>
- Martin, C. L., & Parker, S. (1995). Folk theories about sex and race differences. *Personality and Social Psychology Bulletin*, 21(1), 45–57. <https://doi.org/10.1177/0146167295211006>
- McKibben, W. B., & Silvia, P. J. (2017). Evaluating the distorting effects of inattentive responding and social desirability on self-report scales in creativity and the arts. *The Journal of Creative Behavior*, 51(1), 57–69. <https://doi.org/10.1002/jocb.86>
- Mobasseri, S., Stein, D. H., & Carney, D. R. (2022). The accurate judgment of social network characteristics in the lab and field using thin slices of the behavioral stream. *Organizational Behavior and Human Decision Processes*, 168, 104103. <https://doi.org/10.1016/j.obhdp.2021.09.002>
- Modeus, N., Ståhlbröst, U., Wester, G., & Ögren, G. (1987). *Att vara människa* [To be a human]. Natur och Kultur.
- Ng, T. W., & Feldman, D. C. (2012). A comparison of self-ratings and non-self-report measures of employee creativity. *Human Relations*, 65(8), 1021–1047. <https://doi.org/10.1177/0018726712446015>
- Norlander, T., Erixon, A., & Archer, T. (2000). Psychological androgyny and creativity: Dynamics of gender-role and personality trait. *Social Behavior and Personality*, 28(5), 423–435. <https://doi.org/10.2224/sbp.2000.28.5.423>
- Open Science Framework. (2020). *Replication of androgyny-creativity effect*. [https://osf.io/g2x8k?view\\_only=](https://osf.io/g2x8k?view_only=)
- Plucker, J. A., & Makel, M. C. (2010). Assessment of creativity. In J. C. Kaufman & R. J. Sternberg (Eds.), *The Cambridge handbook of creativity* (pp. 48–73). Cambridge University Press. <https://doi.org/10.1017/CBO9780511763205.005>
- Proudfoot, D., Kay, A. C., & Koval, C. Z. (2015). A gender bias in the attribution of creativity: Archival and experimental evidence for the perceived association between masculinity and creative thinking. *Psychological Science*, 26(11), 1751–1761. <https://doi.org/10.1177/0956797615598739>
- Pruim, R., Kaplan, D. T., & Horton, N. J. (2017). The mosaic package: Helping students to 'think with data' using R. *The R Journal*, 9(1), 77–102. <https://doi.org/10.32614/RJ-2017-024>
- R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <http://www.R-project.org/>
- Revelle, W. (2022). psych: Procedures for personality and psychological research (Version 2.2.3) [Computer software]. Northwestern University. <https://CRAN.R-project.org/package=psych>
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1–36. <https://doi.org/10.18637/jss.v048.i02>
- Said-Metwaly, S., Van den Noortgate, W., & Kyndt, E. (2017). Approaches to measuring creativity: A systematic literature review. *Creativity. Theories—Research—Applications*, 4(2), 238–275. <https://doi.org/10.1515/ctra-2017-0013>
- Saint-Michel, S. E. (2018). Leader gender stereotypes and transformational leadership: Does leader sex make the difference? *M@n@Gement*, 21(3), 944–966. <https://doi.org/10.3917/mana.213.0944>
- Sato, K. (2022). Who is happier in Japan, a housewife or working wife? *Journal of Happiness Studies*, 23(2), 509–533. <https://doi.org/10.1007/s10902-021-00411-3>
- Selker, R., Love, J., Dropmann, D., & Moreno, V. (2021). jmv: The 'jamovi' analyses (R package version 2.0). <https://CRAN.R-project.org/package=jmv>

- Silvia, P. J., Wigert, B., Reiter-Palmon, R., & Kaufman, J. C. (2012). Assessing creativity with self-report scales: A review and empirical evaluation. *Psychology of Aesthetics, Creativity, and the Arts*, 6(1), 19–34. <https://doi.org/10.1037/a0024071>
- Smith, G. J. W., & Carlsson, I. (1990). *CFT: Test på kreativ funktion* [The CFT: A test of the creative function]. Psykologiförlaget.
- Snyder, H. T., Hammond, J. A., Grohman, M. G., & Katz-Buonincontro, J. (2019). Creativity measurement in undergraduate students from 1984–2013: A systematic review. *Psychology of Aesthetics, Creativity, and the Arts*, 13(2), 133–143. <https://doi.org/10.1037/aca0000228>
- Sullivan, J., Moss-Racusin, C., Lopez, M., & Williams, K. (2018). Backlash against gender stereotype-violating preschool children. *PLoS ONE*, 13(4), e0195503. <https://doi.org/10.1371/journal.pone.0195503>
- Torrance, E. P. (1995). *Why fly: A philosophy of creativity*. Ablex Publishing Corporation.
- U.S. Census Bureau. (2022). *QuickFacts*. <https://www.census.gov/quickfacts/fact/table/U.S./PST045221#qf-headnote-a>
- van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45, 1–67. <https://doi.org/10.18637/jss.v045.i03>
- Wallas, G. (1926). *The art of thought*. Jonathan Cape.
- Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag.
- Williams, M. J., & Tiedens, L. Z. (2016). The subtle suspension of backlash: A meta-analysis of penalties for women's implicit and explicit dominance behavior. *Psychological Bulletin*, 142(2), 165–197. <https://doi.org/10.1037/bul0000039>
- Yousaf, O., Popat, A., & Hunter, M. S. (2015). An investigation of masculinity attitudes, gender, and attitudes toward psychological help-seeking. *Psychology of Men & Masculinity*, 16(2), 234–237. <https://doi.org/10.1037/a0036241>
- Zanette, D. H., & Manrubia, S. C. (2001). Vertical transmission of culture and the distribution of family names. *Physica A*, 295(1–2), 1–8. [https://doi.org/10.1016/S0378-4371\(01\)00046-2](https://doi.org/10.1016/S0378-4371(01)00046-2)

Received March 7, 2022

Revision received October 6, 2022

Accepted October 17, 2022 ■