Creativity Ratings of Fashion Outfits Presented on Instagram: Does Gender Matter?

Philipp Barth

Georg Stadtmann

European University Viadrina Frankfurt (Oder)

Department of Business Administration and Economics

Discussion Paper No. 412

August 2019

ISSN 1860 0921

Creativity Ratings of Fashion Outfits Presented on Instagram: Does Gender Matter?

Philipp Barth[a]* and Georg Stadtmann[a]

[a]*Faculty of Business Administration and Economics, European University Viadrina, Frankfurt (Oder), Germany*

Correspondence concerning this article should be addressed to Philipp Barth, Faculty of Business Administration and Economics, European University Viadrina, Grosse Scharrnstrasse 59, 15230 Frankfurt (Oder), Germany. E-mail: barth@europa-uni.de

Creativity Ratings of Fashion Outfits Presented on Instagram: Does Gender Matter?

Creativity assessment can be influenced by social group membership, including gender. This study analyzed implicit biases in the Consensual Assessment Technique (CAT) by examining gender differences in creativity assessment. We asked male (n = 26) and female (n = 39) judges to rate the creativity of fashion outfits presented on Instagram and then examined gender differences in mean creativity ratings and rater consistency (inter-rater reliability). We found no systematic support for gender differences in the level of creativity ratings, but observed that rating consistency was significantly higher for female than for male judges. In an additional qualitative analysis of the implicit criteria that raters applied when assessing creativity, we found that female and male raters attached different relative importance to various assessment criteria. Our study suggests that rater panel composition can indeed affect aspects of creativity assessment, although we do not obtain strong support for a gender-related bias in the CAT methodology.

Keywords: Consensual Assessment Technique; creativity assessment; gender differences; fashion; implicit theories of creativity; Instagram

Introduction

The assessment of creativity can have important implications and consequences. From a business perspective, understanding how and why customers form opinions on creativity is crucial to develop innovative products. From a research perspective, different methods of measuring creativity are grounded on some sort of creativity judgement, such as the assessment of creative products or creative persons (see, e.g., George & Zhou, 2007; Kaufman, Plucker, & Baer, 2008).

One of the most frequently used methods to assess (product) creativity is the *Consensual Assessment Technique* (CAT). It was developed and validated by Amabile (1982a, 1996) and is often referred to as the "gold standard" of creativity assessment (Carson, 2006, as cited in Kaufman, Baer, & Cole, 2009, p. 223). According to the CAT, the best available assessment of a creative product in a given domain can be provided by the collective subjective judgement of a group of experts in that field (Kaufman, Plucker, & Baer, 2008, p. 54). Two of the technique's most prominent features are its use of actual creative products and its claim to mimic "real world assessments" (Kaufman, Plucker, & Baer, 2008, p. 55). The standard procedure of the CAT works as follows (Kaufman, Plucker, & Baer, 2008, p. 56-57). In an initial step, a group of subjects is instructed to create a product according to a prespecified task. Then, in a following step, a panel of judges appropriately familiar with the product domain in question ("experts") is asked to rate the creativity of these products. In this procedure, raters are required to make their judgements independently of one another and to rate each product relative to all other products.

The CAT has been applied as a methodology to measure creativity in a wide range of settings, including different domains, creative tasks, and types of creators. In all of these settings, the CAT proved to be a reliable and valid technique to measure creativity. Reported inter-rater reliabilities mostly ranged between .70 and .90 or higher (see, e.g., Amabile,

1982b, 1985, 1996; Baer, 1997, 1998; Baer, Kaufman, & Gentile, 2004; Hennessey, 1994; Runco, 1989; Tang et al., 2018). Numerous studies have further validated the CAT and examined some of its most prominent features, including the role of raters' expertise (see, e.g., Kaufman & Baer, 2012; Kaufman et al., 2009; Kaufman, Baer, Cole, & Sexton, 2008; Plucker, Kaufman, Temple, & Qian, 2009) or implicit biases against ethnicities or gender (e.g., Kaufman, Baer, Agars, & Loomis, 2010; Kaufman, Niu, Sexton, & Cole, 2010).

In this paper, we draw on previous CAT research and assess biases in the CAT by investigating gender (sex[1]) differences in the creativity assessment of female fashion outfits posted on the social media platform *Instagram*. To do so, we examine the influence of single-sex rater panels on creativity ratings and their consistency. Thereby, we analyze boundary conditions of rater panel composition and further explore the circumstances under which the CAT is a reliable method to measure product creativity. We do this by carrying out a quantitative analysis of creativity ratings and rating consistency (inter-rater reliability) and by complementing it with a qualitative analysis of the implicit criteria that judges apply when providing their creativity assessments.

The question of gender differences in creativity ratings is worthwhile exploring. According to Martin and Wilson (2017, p. 418), "[…] it is known that judgement of creative worth can be influenced […] through membership of social groups, such as gender […]." As most experts are male (Runco, Cramond, & Pagnani, 2010, p. 348), the gender composition of rater panels could introduce a bias into CAT methodology. Any bias—whether conscious or unconscious—can have a severe impact on creativity scores (Kaufman, Niu, et al., 2010, p. 497).

Review of Previous Studies

The amount of creativity research examining gender differences is comparatively limited.

Although differences between males and females have been explored more extensively in recent years, they have never been a major focus in creativity research (Baer & Kaufman, 2008, p. 76). Two reasons potentially account for this fact. On the one hand, it is a topic of great controversy and sensitivity (Abraham, 2016, p. 615). On the other hand, research on sex differences in creativity has so far yielded inconsistent results and a lack of testable theories (Baer & Kaufman, 2008, p. 76). Both aspects possibly discouraged researchers from further investigating the issue.

Gender Differences in Creative Potential and Performance

The vast majority of the sex differences research in the field of creativity focused on either creative potential (ability) or creative performance (achievement; Runco et al., 2010, p. 344). Studies investigating gender differences in *creative potential* yielded inconclusive results. According to Runco et al. (2010, p. 354), about half of the studies found no sex differences in creative ability (e.g., Gralewski & Karwowski, 2013; He & Wong, 2011; Saeki, Fan, & Van Dusen; 2001). In contrast, about a third of the studies reported female advantages in creative potential (e.g., Cheung & Lau, 2010; Mullineaux & Dilalla, 2009), whereas a much smaller amount of research provided (at least partial) evidence for higher creative potential among males (e.g., Lin, Hsu, Chen, & Wang, 2012). Thus, if a case could be made at all, it would suggest that females possess higher creative ability (Baer & Kaufman, 2008, p. 78). More recent research has not yet clarified the question of gender differences in creative potential, reporting evidence for male advantages, female advantages or no gender differences at all, sometimes depending on the specific aspects of creativity or the concrete age groups investigated (e.g., He, 2018; Warren, Mason-Apps, Hoskins, Azmi, & Boyce, 2018; Zhang, Ren, & Deng, 2018).

Studies examining *creative performance* also showed mixed evidence. On the one hand, there was a considerable amount of research finding no gender differences (e.g., Amabile, 1982a; Kaufman, Baer, & Gentile, 2004; Mierdel & Bogner, 2019) and a much lower number of studies providing (partial) evidence for female advantages in creative performance (e.g., Kaufman, Niu, et al., 2010). On the other hand, males have often dominated in real-world creative achievement (e.g., Piirto, 1991; Simonton, 1991). Accordingly, Runco et al. (2010, pp. 353–354) contend that gender differences in creative achievement have historically existed in most fields.

Gender Differences in Creativity Assessment

Another line of research—although mostly unexplored—specifically concerns the topic of this paper, namely gender differences in *creativity assessment*. It traces back to other areas of psychological research, which provided evidence for cognitive differences between men and women in the early phases of information processing, such as perception and attention (Halpern, 2012). Harris (1989, p. 16), for example, found that males and females differed in their preferences for shape and color schemes, thereby empirically supporting the insight that "women and men 'see' colors somewhat differently" (Halpern, 2012, p. 107).

To the best of our knowledge, only one study has so far examined gender differences in creativity assessment, namely Kaufman, Niu, et al. (2010). Although the mere analysis of sex differences in creativity assessment was not their main purpose, Kaufman, Niu, et al. (2010) examined this aspect in the course of studying interaction effects between gender and ethnicity among raters and ratees. They had a sample of college students of different gender and ethnicity write a poem and a story and asked nonexpert student judges to rate the creativity of these products. In their analyses, they could not detect differences in creativity ratings between male and female judges. However, using Cronbach's coefficient alpha as a

measure of inter-rater reliability, they found female creativity ratings to be significantly more consistent than male ratings for both poems (.92 vs. .73) and stories (.94 vs. .70).

Present Studies

Rationale

The main goal of this paper was to further validate the CAT by exploring gender-related biases in the methodology, thereby contributing to an area of creativity research that has hardly been examined. Using female fashion outfits posted on Instagram as ratable products followed a specific logic: Sex differences in creativity assessment could be particularly pronounced in contexts that are—either by their nature or by social implications—more familiar or more controversial to one sex than to the other. Female fashion seems to have the potential to elicit different degrees of controversy and familiarity among women and men, thereby serving as a likely source of bias. Finding evidence for or against biases in such a susceptible context could help particularly well to understand the extent of such biases.

Research Objectives

Based on the main goal of this study, we pursued three specific objectives.

(1) Creativity ratings: Analysis of gender differences in creativity ratings of female fashion outfits;
(2) Inter-rater reliability: Examination of gender differences in inter-rater reliability as a proxy for rating consistency;
(3) Implicit theories of creativity: Investigation and comparison of the implicit theories of creativity underlying the creativity judgements made by female and male raters.

The examination of the first two research objectives was supposed to include analyses of a main sample of female fashion outfits as well as analyses of various subsamples based on the type of outfit presented and the type of user presenting the outfit. With this additional subsample analysis, we intended to verify the robustness of our main analysis and obtain evidence on the influence of certain sample characteristics. The examination of the third research question aimed to understand how judgements were made and what criteria raters employed when making their assessments, thereby qualifying the insights gained in the previous two research objectives.

Method

Creative Products

As rateable artifacts, we used fashion outfits worn by Instagram users whose social media accounts focused on fashion or lifestyle topics. We downloaded pictures of these outfits from Instagram and later presented them to the judges. By employing products that were neither created for experimental purposes (nonexperimental) nor under experimentally controlled conditions (nonparallel), we followed a procedure applied by Baer et al. (2004, p. 114). Deviating from their approach, we used creative products that were additionally neither made nor combined following any instructions (noninstructed). Instead, users wore and presented the fashion outfits voluntarily.

We selected the fashion outfits in accordance with the following procedure:

(1) We defined two types of German female Instagram users, namely well known "social media influencers" with at least 100,000 followers as well as less known users with less than 10,000 followers. For each type, we identified 15 users,

serving as the type of user subsamples in our data analysis. By limiting the selection to female German users, we controlled for potential rating effects induced by the gender or the nationality of the outfit wearers.

(2) We defined three categories of fashion outfits for which suitable pictures were collected. These categories served as the type of outfit subsamples in our data analysis and as a mechanism to ensure a balanced, unbiased compilation of rateable products. For the *standard fashion outfit* category (hereafter *standard*), only "everyday" outfits, which can be seen regularly and lack extraordinary features, were chosen. For the *non-standard fashion outfit* category (*non-standard*), more unusual, extravagant outfits were selected. Finally, for the *revealing fashion outfit* category (*revealing*), outfits with either high levels of sex appeal or high visibility of skin were included.

(3) We compiled a preselection of up to five fashion outfits for each Instagram user and each type of fashion outfit and—together with a female co-selector with a distinct interest in fashion—chose the most suitable fashion outfit for each user and each type of outfit from this preselection.

This process resulted in a main sample of 90 fashion outfits (2 types of Instagram users x 15 Instagram users per type x 3 types of outfits per Instagram user). For the rating process, all pictures were scaled to the same standardized size.

Survey Instrument

The creativity rating process (see "Rating Procedure") was based on the CAT and administered on computers. For that reason, we designed a computer-based online survey that was formulated in German language. In this survey, we asked participants to rate the creativity as well as the likeability of each fashion outfit. The likeability category was further

broken down into the likeability of the fashion outfit itself and the likeability of the person wearing that outfit. This procedure was supposed to facilitate analyses of influences resulting from the likeability of a certain Instagram user and to account for the fact that outfits could only be presented in connection with the users wearing them. The entire rating process was randomized. Additionally, we used open questions to ask the raters for the reasons that led them to either rate a fashion outfit as highly creative or highly uncreative. In this context, raters were requested to provide at least three reasons as to why a fashion outfit was (or was not) creative, following an approach of Loewenstein and Mueller (2016, p. 324).

One aspect of this setup did not follow the original CAT methodology, but can be regarded as adequate in the context of this study. In fact, we asked raters for their criteria of fashion creativity, but did so *after* all ratings had been completed. Thereby, we adhered to the requirement that judges are not supposed to explain or defend their assessments (Kaufman, Plucker, Baer, 2008, p. 57).

Raters

We contacted students of a German public university who had signed up to a mailing list for upcoming experiments and asked them for their participation as raters in this study. By using students as nonexpert judges, we followed the approach of other studies investigating biases in the CAT (e.g., Kaufman, Baer, et al., 2010; Kaufman, Niu, et al., 2010). To control for influences of cultural norms as well as possible, all participants were required to have completed their high school diplomas (A-Levels) in Germany. Sixty-five students aged 19 to 32 years (M = 23.4, SD = 2.75) served as judges, of which 26 were males aged 20 to 28 years (M = 23.6, SD = 2.37) and 39 females aged 19 to 32 years (M = 23.2, SD = 2.96). The judges were predominantly undergraduate students (65%) majoring in business (55%), cultural science (25%) or law (20%). All raters were paid for their participation in this study.

Although the CAT usually requires expert raters—the most objective, accurate and valid source of creativity assessments (e.g., Kaufman, Plucker, & Baer, 2008, p. 59)—, the use of nonexpert judges is not only in line with the specific purpose of this study, but also a likely source of valid creativity judgements of fashion outfits.

On the one hand, this study is primarily interested in the investigation of assessment biases induced by the gender of subjects in general, and not by domain experts in particular. Accordingly, accurate creativity ratings are not essential for this study. Following this line of argumentation—originally provided by Kaufman, Baer, et al. (2010, p. 202) —, this study can and actually should use nonexpert raters.

On the other hand, the actual meaning of the requirement of expert judges is still debatable. Although expertise is usually associated with close familiarity with the domain in which a product was created (Amabile, 1996, p. 73), previous empirical research could neither provide conclusive evidence on the required level of expertise nor on when and why the use of expert raters is indeed essential (Cseh & Jeffries, 2019, p. 161). Actually, it was pointed out long ago that the necessary level of expertise might depend on the product domain in question, with some domains being sufficiently simple to allow for raters with minimal familiarity in the respective domain (Amabile, 1982a, p. 1009). Fashion might in fact be such a domain: Freeman, Son, and McRoberts (2015) found that nonexperts and experts did not differ in their creativity evaluations of fashion design illustrations, reporting a high and significant correlation between the judgements of the two rater groups ($r = .83$, $p < .001$). Thereby, Freeman et al. (2015) present evidence that both expert and nonexpert judges can provide reliable and valid assessments in a fashion context. Accordingly, even if not necessary for the purpose of this study, the creativity judgements provided by our nonexpert rater panel might be accurate and valid after all.

Rating Procedure

The online survey and the rating procedure were conducted in the computer laboratory of a German public university. Each rater worked on his or her own computer shielded by partition walls (cubicles). Thereby, each judge was able to make his or her judgments independently. In order to be able to provide ratings relative to all other fashion outfits, all raters received physical booklets with pictures of all fashion outfits, allowing them to compare outfits at all times. In line with the CAT, we asked raters to assign ratings based on their own personal and subjective definitions of creativity. We gave them as much time as needed to complete the survey and rate all outfits. All ratings had to be made on a 7-point Likert scale, ranging from 1 (*very uncreative*) to 7 (*very creative*). All raters completed the entire survey.

Data Analysis and Results

We first computed correlations between creativity ratings and the other two dimensions of subjective assessment, namely the likeability of fashion outfits and the likeability of the persons wearing these outfits. This analysis provided evidence that the different subjective assessments we asked for were independent of one another. We found low positive mean correlations between the creativity and the likeability of fashion outfits (.08) as well as between the creativity of fashion outfits and the likeability of the persons wearing the outfits (.16). Thus, raters were apparently able to distinguish between these concepts and responded to the dimensions they were asked for. In addition, the latter correlation indicates that there were no serious confounding effects induced by the likeability of the "creators" (wearers) of the outfits, whose identity was revealed to all raters through the Instagram pictures.

In a next step, we analyzed our data concerning the three research objectives outlined above. All analyses were based on single-sex rater groups composed entirely of either female or male judges.

Creativity Ratings

We investigated gender differences in creativity ratings for the main sample of all 90 fashion outfits and for all subsamples. To compare the mean ratings of male and female rater panels in each sample, we computed independent means t tests. All tests were carried out at a significance level of $\alpha = 10\%$.

Table 1 shows mean creativity ratings for male and female rater groups and each of the investigated samples.

TABLE 1 TO BE INSERTED HERE

With the exception of one subsample (*non-standard*), male raters provided higher mean ratings than female raters in the main sample as well as in all subsamples. However, these gender differences in mean creativity ratings were only statistically significant in two subsamples and our hypothesis tests did not yield any systematic patterns. Male raters provided higher scores for the *revealing* subsample ($t(63) = -1.96$, $p = .054$, $d = -0.50$), but female judges assigned higher mean ratings for the non-standard subsample ($t(63) = 1.79$, $p = .079$, $d = 0.45$). As the *non-standard* subsample did not appear to be normally distributed, the latter result was validated using the Mann−Whitney U test and remained the same ($U = 645$, $p = .065$, $r = .23$). For all other examined samples, mean ratings between female and male judges did not differ significantly.

Inter-Rater Reliability

Cronbach's Coefficient Alpha

Inter-rater reliability as a measure of rating consistency was calculated for male and female rater panels using Cronbach's coefficient alpha. As coefficient alpha is a point estimate, we computed 95% confidence intervals, which provide additional information on the precision and location of each point estimate as well as on statistical significance (Belia, Fidler, Williams, & Cumming, 2005, p. 389). In contrast to Kaufman, Niu, et al. (2010) who applied the Duhachek and Iacobucci methodology (Duhachek & Iacobucci, 2004; Iacobucci & Duhachek, 2003) to calculate confidence intervals, we used a more recent approach by Bonett and Wright (2015). We tested for significant differences between inter-rater reliabilities of male and female raters using the testing procedure and the 95%-confidence-interval-based decision rules outlined by Bonett and Wright (p. 7).

Table 2 gives an overview of Cronbach's coefficient alphas, their confidence intervals, and the results of the significance tests.

TABLE 2 TO BE INSERTED HERE

Generally, and irrespective of the examined sample, we found high inter-rater reliabilities for female and—in most cases—male rater panels. For female raters, Cronbach's coefficient alphas amounted to levels of .90 or higher in all samples investigated, thereby clearly exceeding the common minimum standards of reliability of .70 (Nunnally & Bernstein, 1994, pp. 264–265) or .80 (Lance, Butts, & Michels, 2006, p. 206). Male raters consistently showed lower inter-rater reliabilities in each sample investigated, with coefficient alphas exceeding levels of .90 in the main sample as well as in the type of user subsamples, but only ranging between .65 and .83 in the type of outfit subsamples. Testing

for differences between these inter-rater reliabilities using the Bonett and Wright approach (2015) yielded significant gender differences in all but one subsample, namely *revealing*.

Spearman–Brown Adjusted Alpha

Inter-rater reliability increases with the number of judges (Kaufman et al., 2009, p. 224). To account for the different sizes of female (n = 39) and male (n = 26) rater panels, we additionally calculated adjusted alphas using the Spearman–Brown prophecy formula (see, e.g., Nunnally & Bernstein, 1994, p. 232). The Spearman–Brown adjusted alpha allows for a comparison of rater panels under the assumption that groups have equal numbers of judges. We standardized both single-sex rater panels to sizes of 10 raters, thereby creating a panel size approximating usual group sizes reported in the literature (see, e.g., Kaufman, Plucker, & Baer, 2008, p. 58). Again, we used the Bonett and Wright (2015) approach to test for significant differences.

Table 3 displays Spearman–Brown adjusted alphas, their confidence intervals, as well as the results of the Bonett and Wright (2015) significance tests.

TABLE 3 TO BE INSERTED HERE

For the female rater panel, Spearman–Brown adjusted alphas reached at least acceptable levels, exceeding the minimum threshold of .70 in all examined samples and attaining levels of up to .90. Again, adjusted alphas for male raters were constantly lower, sometimes clearly failing to reach the proposed minimum reliability of .70, in particular in the type of outfit subsamples. Testing for gender differences between the respective adjusted alphas yielded the same results as in the previous analysis. Except in the *revealing* subsample, female judges rated the creativity of fashion outfits significantly more consistently than male raters did.

Implicit Theories of Creativity

Approach

Finally, we analyzed the creativity assessment criteria used by male and female raters to examine their implicit theories about what aspects of the fashion outfits made an outfit creative or uncreative. To do so, we used content analysis (e.g., Krippendorff, 2013) and partly followed the procedure applied by Loewenstein and Mueller (2016). Specifically, we replicated the criteria coding method used during the first phase of their cultural consensus analysis (p. 324). Our analysis included three main steps:

(1) Each explanation was unitized to identify distinct, conceptually different statements on why a certain fashion outfit had been regarded as creative or uncreative. This was necessary since raters had sometimes provided more than one explanation in a single response. Thereby, 240 (212) initial explanations on why outfits were regarded as particularly creative (uncreative) resulted in 253 (233) distinct statements and—after eliminating statements that could not be categorized—in 230 (225) usable distinct statements. Thus, 455 usable distinct statements were used for our qualitative analyses.

(2) Each of the 455 statements was evaluated and assigned to a category of (preferably) distinct criteria for particularly creative or uncreative fashion outfits. Since categories can still contain nuances of similarity with other categories, we additionally followed Long's (2014, pp. 186–187) approach of grouping different categories under frames of similar meaning. The process of categorization and framing was highly iterative and repeatedly included steps of generating, merging, or deleting categories.

(3) We examined the frequencies of mentioning the different frames for both male and female raters. Then, we derived proportions from these frequencies and subsequently tested for differences at a significance level of $\alpha = 10\%$ using chi-square tests of independence. This last step was supposed to identify gender differences and similarities in implicit theories of creative fashion and to reconcile these with the results of our analyses of gender differences in creativity ratings. For the final step of our analysis, we used all distinct statements provided by the raters, even if they referred to the same category.

Identified Frames and Categories

Following the outlined procedure, twenty-four criteria categories for highly creative and uncreative fashion outfits were identified (see Appendix A for descriptions of the criteria categories). Based on similarities in their meanings, these categories were subsumed under seven frames (see Appendix B for descriptions of the frames). Following the approach used by Long (2014, p. 188), we comprehensively illustrated all frames and categories as well as the relationships between one another in cases of overlaps, similarities or ambiguities (see Appendix C for an illustration of the relationships between the frames).

Female and Male Raters' Use of Criteria

Table 4 summarizes the results of our content analysis.

TABLE 4 TO BE INSERTED HERE

We found gender differences in the relative importance (ranking) of several frames (especially for those mentioned less frequently). For instance, the criteria *appropriate* and *appeal/likeability* seemed to play a much less important role for female judges than for male judges in identifying highly creative and highly uncreative outfits. The difference between

the proportions of statements assigned to *appropriate* by female raters (5%) and male raters (18%) was significant ($\chi^2 = 20.10$, $p < .001$), as was the difference between the proportions assigned to *appeal/likeability* by women (1%) and men (3%; $\chi^2 = 3.20$, $p = .074$). Similarly, the criterion *artful* seemed to be more important for the identification of creative fashion for female judges (7%) than for male judges (2%; $\chi^2 = 5.44$, $p = .020$). The same trend could be observed for the *noticeable* criterion, with female raters assigning relatively more distinct statements to this frame (29%) compared to male raters (21%; $\chi^2 = 3.52$, $p = .061$). However, the latter observation was not robust with regard to the method used. When we reran our analysis using *distinct categories of statements* instead of using *all statements*—thereby eliminating all categories that were mentioned multiple times by a single rater—the difference failed to reach statistical significance.

Finally, there was partial evidence that females attached more relative importance to their top three frames (84%) compared to males (77%). The respective proportions differed between female and male raters at the 10% level ($\chi^2 = 2.74$, $p = .098$). Again, we reran our analysis using only distinct categories of statements instead of using all statements, but the test failed to reach statistical significance.

Discussion

Interpretation and Implications

This study yielded three main findings, each of them relating to one of the three research objectives outlined above.

Creativity Ratings

First, the results of this study suggest that males and females do not rate the creativity of female fashion outfits differently. This result is consistent with previous findings of

Kaufman, Niu, et al. (2010), who did not detect any significant rating differences between men and women either.

Inter-Rater Reliability

Second, our main analysis suggests that female raters assess the creativity of fashion outfits more consistently than male raters do. Irrespective of the measure used (Cronbach's coefficient alpha or Spearman–Brown adjusted alpha), inter-rater reliability was significantly higher for females than for males in the main sample as well as in most subsamples. Again, this result is in line with previous research. One the one hand, it is consistent with general evidence from psychological research indicating that males are more variable than females (overrepresentation in the distributional tails) and therefore show different (normal) distributional curves in several areas (see, e.g., Halpern, 2012; Pinker & Spelke, 2005). On the other hand, it matches previous findings from creativity research. For instance, He (2018) and Karwowski et al. (2016) provided evidence that males show greater variability in creative ability than females, which was partially supported by Lau and Cheung (2015). Kaufman, Niu, et al. (2010) presented support for higher rating consistency among females than among males. The latter result specifically matches the insights from this study.

However, one aspect of our results relating to the subsample analysis is striking. *Revealing* was the only subsample (and the only type of outfit) for which female judges did not provide more consistent ratings than males, thereby deviating from the general pattern of findings. This suggests that rating consistency could in fact be influenced by the product to be rated. *Revealing* included fashion outfits with high skin visibility, thereby forming the only subsample with a potentially distinct (sexual) appeal to each gender. Although currently purely speculative, it could be the case that male raters place relatively equal emphasis on factors that attract them, thereby inducing more consistent ratings than in other subsamples.

Similarly, it could be the case that some females are more distracted by these factors than others are, resulting in lower consistency among them. Both approaches could explain why inter-rater reliability differed between female and male raters in all but the *revealing* subsample. However, since the "promiscuity" of outfits was barely mentioned as a criterion for highly creative or uncreative outfits in our qualitative analysis (four times in total), it is likely that the internal processes underlying such convergence of inter-rater reliabilities proceed unconsciously. In either case, products containing some aspect of controversy or sexual connotation could affect the rating consensus of males and females differently.

Implicit Theories of Creativity

Third, our content analysis suggests that the relative importance assigned to certain creativity assessment criteria differs between male and female judges. We found that females based their judgements more often on artistic, imaginative and playful components of fashion outfits (represented by the frame *artful*), whereas males assigned more weight to the mere appeal of an outfit and its appropriateness in a given context (represented by the frames *appeal/likeability* and *appropriate*).

Yet, this variety in rating criteria only found matching expression in rating consistency (for which we detected gender differences), but not in levels of creativity ratings (for which we did not detect gender differences). The latter part is particularly surprising. Two explanations are possible. First, gender differences in assessment criteria were only detected in frames with minor importance (*artful*, *appeal/likeability*, and *appropriate*). Accordingly, the frames mentioned more frequently (*original*, *noticeable*, *variety*) might have just been the major assessment criteria, thereby marginalizing the influence of other criteria. Second, the actual effects on mean creativity ratings induced by the criteria for

which gender differences were found could have just compensated one another, resulting in the same average score.

The fact that we did find gender differences in rating consistency appears to be in line with the detected sex differences in rating criteria. Compared to male raters, female judges used higher shares of their statements to name the three most important frames, thereby showing higher agreement concerning important rating criteria. This agreement might have just resulted in higher rating consistency.

Limitations and Directions for Future Research

This study has limitations and provides several avenues for future research. First, although the raters assessed real-world products, these were presented using photographs retrieved from Instagram. Accordingly, raters could only take into consideration what they saw on the pictures, which can be influenced by camera angles or picture details. Haptic factors like fabrics could not be examined. However, this limitation affected all raters and all rateable outfits equally and can thus have no effect on the relative creativity ratings the CAT is based on.

Second, some specific limitations of this study may affect the generalizability of our findings. Therefore, future research should examine the robustness of our results concerning various aspects. On the one hand, it is possible (and maybe even likely) that the degree of familiarity with female fashion outfits differed between male and female raters. Females might have higher expertise with fashion specifically designed for them. As a result, discriminating between the relative influences of gender and expertise is not possible based on this study. Although such a bias is conceivable for all raters (even experts) and all products with any kind of "gender tendency," future research could replicate our study with experts serving as judges, thereby eliminating potential influences of different expertise

levels among raters. In the same manner, the generalizability of our findings could be further supported by replicating our study with male fashion outfits as rateable products. On the other hand, this study investigated boundary conditions of gender biases in the CAT as only single-sex rater panels were analyzed. Gender differences might be less distinctive when analyzing the more realistic scenario of mixed-sex rater panels. Future research could therefore investigate mixed group compositions and their effects on biases in the CAT. Furthermore, with 90 rateable fashion outfits in the main sample and 30 outfits in each of the type of outfit subsamples, sample sizes were rather small. Sample size affects statistical power and increases the likelihood of Type II errors. Therefore, our findings should be validated based on larger samples of rateable products.

Finally, in addition to the research directions derived from the limitations of this study, future research could also examine the more puzzling results of this study. On the one hand, the relationship between creativity ratings and underlying rating criteria could be further analyzed to find out whether male and female judges assign different relative importance to certain criteria in other contexts as well, and how such differences or similarities relate to gender differences in the levels of creativity ratings. On the other hand, the role of potentially controversial products for sex differences in creativity judgments—such as revealing fashion outfits—could be further investigated, thereby taking up our finding that revealing fashion outfits were the only type of outfit in which gender differences in inter-rater reliability were not detected.

Contribution and Conclusion

This study contributes to the creativity and the CAT literature by providing additional empirical evidence to the limited body of research on sex differences in creativity assessments. It did so by using a research design that—for the first time—provided an

additional source of insight into gender differences in creativity assessment by connecting creativity ratings to the underlying rating criteria.

Our study suggests that rater panel composition in the CAT can indeed affect creativity assessment, in particular rating consistency. We confirmed previous results presented by Kaufman, Niu, et al. (2010) and provided support for the robustness of their findings. Since we did not find strong evidence for gender biases in the CAT, this study further validates the CAT as a reliable method to assess product creativity.

References

Abraham, A. (2016). Gender and creativity: An overview of psychological and neuroscientific literature. *Brain Imaging and Behavior, 10*, 609–618. https://doi.org/10.1007/s11682-015-9410-8

Amabile, T. M. (1982a). 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

Amabile, T. M. (1982b). Children's artistic creativity: Detrimental effects of competition in a field setting. *Personality and Social Psychology Bulletin, 8*(32), 573–578. https://doi.org/10.1177/0146167282083027

Amabile, T. M. (1985). Motivation and creativity: Effects of motivational orientation on creative writers. *Journal of Personality and Social Psychology, 48*(2), 393–399. https://doi.org/10.1037/0022-3514.48.2.393

Amabile, T. M. (1996). *Creativity in context: Update to "the social psychology of creativity."* Boulder, CO: Westview.

Baer, J. (1997). Gender differences in the effects of anticipated evaluation on creativity. *Creativity Research Journal, 10*(1), 25–31. https://doi.org/10.1207/s15326934crj1001_3

Baer, J. (1998). Gender differences in the effects of extrinsic motivation on creativity. *Journal of Creative Behavior, 32*(1), 18–37. https://doi.org/10.1002/j.2162-6057.1998.tb00804.x

Baer, J., & Kaufman, J. C. (2008). Gender differences in creativity. *Journal of Creative Behavior, 42*(2), 75–105. https://doi.org/10.1002/j.2162-6057.2008.tb01289.x

Baer, J., Kaufman, J. C., & Gentile, C. A. (2004). Extension of the Consensual Assessment Technique to nonparallel creative products. *Creativity Research Journal, 16*(1), 113–117. https://doi.org/10.1207/s15326934crj1601_11

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods, 10*(4), 389–396. https://doi.org/10.1037/1082-989X.10.4.389

Bonett, D. G., & Wright, T. A. (2015). Cronbach's alpha reliability: Interval estimation, hypothesis testing, and sample size planning. *Journal of Organizational Behavior, 36*, 3–15. https://doi.org/10.1002/job.1960

Carson, S. (2006). *Creativity and mental illness*. Invitational panel discussion hosted by Yale's Mind Matters Consortium, New Haven, CT., April 19, 2006.

Cheung, P. C., & Lau, S. (2010). Gender differences in the creativity of Hong Kong school children: Comparison by using the new electronic Wallach–Kogan Creativity Tests. *Creativity Research Journal, 22*(2), 194–199. https://doi.org/10.1080/10400419.2010.481522

Cseh, G. M., & Jeffries, K. K. (2019). A scattered CAT: A critical evaluation of the Consensual Assessment Technique for creativity research. *Psychology of Aesthetics, Creativity, and the Arts, 13*(2), 159–166. https://doi.org/10.1037/aca0000220

Duhachek, A., & Iacobucci, D. (2004). Alpha's Standard Error (ASE): An accurate and precise confidence interval estimate. *Journal of Applied Psychology, 89*(5), 792–808. https://doi.org/10.1037/0021-9010.89.5.792

Freeman, C., Son, J., & McRoberts, L. B. (2015). Comparison of novice and expert evaluations of apparel design illustrations using the Consensual Assessment Technique. *International Journal of Fashion Design, Technology and Education, 8*(2), 122–130. https://doi.org/10.1080/17543266.2015.1018960

George, J. M., & Zhou, J. (2007). Dual tuning in a supportive context: Joint contributions of positive mood, negative mood, and supervisory behaviors to employee creativity. *The Academy of Management Journal, 50*(3), 605–622. https://doi.org/10.5465/amj.2007.25525934

Gralewski, J., & Karwowski, M. (2013). Polite girls and creative boys? Students' gender moderates accuracy of teachers' ratings of creativity. *Journal of Creative Behavior, 47*(4), 290–304. https://doi.org/10.1002/jocb.36

Halpern, D. F. (2012). *Sex differences in cognitive abilities* (4th ed.). New York: Psychology Press.

Harris, L. J. (1989). Two sexes in the mind: Perceptual and creative differences between women and men. *Journal of Creative Behavior, 23*(1), 14–25. https://doi.org/10.1002/j.2162-6057.1989.tb00514.x

He, W. (2018). A 4-year longitudinal study of the sex-creativity relationship in childhood, adolescence, and emerging adulthood: Findings of mean and variability analyses. *Frontiers in Psychology, 9*, 2331. https://doi.org/10.3389/fpsyg.2018.02331

He, W., & Wong, W. (2011). Gender differences in creative thinking revisited: Findings from analysis of variability. *Personality and Individual Differences, 51*, 807–811. https://doi.org/10.1016/j.paid.2011.06.027

Hennessey, B. A. (1994). The consensual assessment technique: An examination of the relationship between ratings of product and process creativity. *Creativity Research Journal, 7*(2), 193–208. https://doi.org/10.1080/10400419409534524

Iacobucci, D., & Duhachek, A. (2003). Advancing alpha: Measuring reliability with confidence. *Journal of Consumer Psychology, 13*(4), 478–487. https://doi.org/10.1207/S15327663JCP1304_14

Karwowski, M., Jankowska, D. M., Gajda, A., Marczak, M., Groyecka, A., & Sorokowski, P. (2016). Greater male variability in creativity outside the WEIRD world. *Creativity Research Journal, 28*(4), 467–470. https://doi.org/10.1080/10400419.2016.1229978

Kaufman, J. C., & Baer, J. (2012). Beyond new and appropriate: Who decides what is creative? *Creativity Research Journal, 24*(1), 83–91. https://doi.org/10.1080/10400419.2012.649237

Kaufman, J. C., Baer, J., Agars, M. D., & Loomis, D. (2010). Creativity stereotypes and the Consensual Assessment Technique. *Creativity Research Journal, 22*(2), 200–205. https://doi.org/10.1080/10400419.2010.481529

Kaufman, J. C., Baer, J., & Cole, J. C. (2009). Expertise, domains, and the Consensual Assessment Technique. *Journal of Creative Behavior, 43*(4), 223–233. https://doi.org/10.1002/j.2162-6057.2009.tb01316.x

Kaufman, J. C., Baer, J., Cole, J. C., & Sexton, J. D. (2008). A comparison of expert and nonexpert raters using the Consensual Assessment Technique. *Creativity Research Journal, 20*(2), 171–178. https://doi.org/10.1080/10400410802059929

Kaufman, J. C., Baer, J., & Gentile, C. A. (2004). Differences in gender and ethnicity as measured by ratings of three writing tasks. *The Journal of Creative Behavior, 38*(1), 56–69. https://doi.org/10.1002/j.2162-6057.2004.tb01231.x

Kaufman, J. C., Niu, W., Sexton, J. D., & Cole, J. C. (2010). In the eye of the beholder: Differences across ethnicity and gender in evaluating creative work. *Journal of Applied Social Psychology, 40*(2), 496–511. https://doi.org/10.1111/j.1559-1816.2009.00584.x

Kaufman, J. C., Plucker, J. A., & Baer, J. (2008). *Essentials of creativity assessment*. Hoboken, NJ: John Wiley & Sons.

Krippendorff, K. (2013). *Content analysis: An introduction to its methodology* (3rd ed.). Thousand Oaks, CA: Sage.

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria: What did they really say? *Organizational Research Methods, 9*(2), 202–220. https://doi.org/10.1177/1094428105284919

Lau, S., & Cheung, P. C. (2015). A gender-fair look at variability in creativity: Growth in variability over a period versus gender comparison at a time point. *Creativity Research Journal, 27*(1), 87–95. https://doi.org/10.1080/10400419.2015.992685

Lin, W. L., Hsu, K. Y., Chen, H. C., & Wang, J. W. (2012). The relations of gender and personality traits on different creativities: A dual-process theory account. *Psychology of Aesthetics, Creativity, and the Arts, 6*(2), 112–123. https://doi.org/10.1037/a0026241

Loewenstein, J., & Mueller, J. (2016). Implicit theories of creative ideas: How culture guides creativity assessments. *Academy of Management Discoveries, 2*(4), 320–348. https://doi.org/10.5465/amd.2014.0147

Long, H. (2014). More than appropriateness and novelty: Judges' criteria of assessing creative products in science tasks. *Thinking Skills and Creativity, 13*, 183–194. https://doi.org/10.1016/j.tsc.2014.05.002

Martin, L., & Wilson, N. (2017). Defining creativity with discovery. *Creativity Research Journal, 29*(4), 417–425. https://doi.org/10.1080/10400419.2017.1376543

Mierdel, J., & Bogner, F. X. (2019). Is creativity, hands-on modeling and cognitive learning gender-dependent? *Thinking Skills and Creativity, 31*, 91–102. https://doi.org/10.1016/j.tsc.2018.11.001

Mullineaux, P. Y., & Dilalla, L. F. (2009). Preschool pretend play behaviors and early adolescent creativity. *Journal of Creative Behavior, 43*(1), 41–57. https://doi.org/10.1002/j.2162-6057.2009.tb01305.x

Nunnally, J. C., & Bernstein, I. A. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Piirto, J. (1991). Why are there so few? (creative women: Visual artists, mathematicians, musicians). *Roeper Review, 13*(3), 142–147. https://doi.org/10.1080/02783199109553340

Pinker, S., & Spelke, E. (April 22, 2005). *The science of gender and science: Pinker vs. Spelke: A debate sponsored by Harvard's Mind Brain and Behavior Inter-Faculty Initiative.* Retrieved July 24, 2018, from the Edge Foundation Web site: http://edge.org/3rd_culture/debate05/debate05_index.html

Plucker, J. A., Kaufman, J. C., Temple, J. S., & Qian, M. (2009). Do experts and novices evaluate movies the same way? *Psychology & Marketing, 26*(5), 470–478. https://doi.org/10.1002/mar.20283

Runco, M. A. (1989). Parents' and teachers' ratings of the creativity of children. *Journal of Social Behavior and Personality, 4*(1), 73–83.

Runco, M. A., Cramond, B., & Pagnani, A. R. (2010). Gender and creativity. In J. C. Chrisler, & D. R. McCreary (Eds.), *Handbook of gender research in psychology* (pp. 343–357). https://doi.org/10.1007/978-1-4419-1465-1_17

Saeki, N., Fan, X., & Van Dusen, L. (2001). A comparative study of creative thinking of American and Japanese college students. *Journal of Creative Behavior, 35*(1), 24–36. https://doi.org/10.1002/j.2162-6057.2001.tb01219.x

Simonton, D. K. (1991). Emergence and realization of genius: The lives and works of 120 classical composers. *Journal of Personality and Social Psychology, 61*(5), 829–840. https://doi.org/10.1037/0022-3514.61.5.829

Tang, M., Werner, C., Cao, G., Tumasjan, A., Shen, J., Shi, J., & Spörrle, M. (2018). Creative expression and its evaluation on work-related verbal tasks: A comparison of Chinese and German samples. *Journal of Creative Behavior, 52*(1), 91–103. https://doi.org/10.1002/jocb.134

Warren, F., Mason-Apps, E., Hoskins, S., Azmi, Z., & Boyce, J. (2018). The role of implicit theories, age, and gender in the creative performance of children and adults. *Thinking Skills and Creativity, 28*, 98–109. https://doi.org/10.1016/j.tsc.2018.03.010

Zhang, W., Ren, P., & Deng, L. (2018). Gender differences in the creativity–academic achievement relationship: A study from China. *Journal of Creative Behavior*. https://doi.org/10.1002/jocb.387

Footnotes

[1] The terms “gender differences” and “sex differences” will be used interchangeably in this paper. With both terms, we intend to refer to both biological as well as psychosocial aspects of differences between men and women.

Appendix

Appendix A

Description of Criteria Categories

Appeal/likeability refers to the likeability of a fashion outfit with regard to various aspects of an outfit. Highly appealing or likeable fashion outfits were associated with high levels of creativity and included labels such as "pleasant" or "beautiful." Unappealing or unaesthetic fashion outfits were associated with particularly low levels of creativity and incorporated labels such as "unpleasant" or "trashy."

Artistic relates to the amount of art incorporated in a fashion outfit. This category was mentioned as an indicator for particularly creative outfits only. Accordingly, artistic or sophisticated uses of fashion outfits and combinations of single fashion items indicated highly creative fashion outfits.

Authentic refers to the degree of authenticity of a fashion outfit. Authentic outfits were considered highly creative, whereas unauthentic outfits without any aspect of individuality were considered highly uncreative.

Contrast (within outfits) is concerned with the generation of contrasts within a certain fashion outfit. Fashion outfits representing some sort of "inner contrast"—such as between single components of an outfit or between colors—indicated highly creative outfits. Fashion outfits without any indication of such contrasts were considered particularly uncreative. They were associated with statements indicating that only very similar items or colors had been used.

Courageous is concerned with the magnitude of courage incorporated into a fashion outfit. Particularly creative outfits were associated with a lot of courage necessary to use and wear a certain outfit in the way that it was used and worn. Labels assigned included

buzzwords such as “courageous” or “daring.” Uncreative fashion outfits were not associated with courage at all, including labels such as “streamlined.”

Distinct/different (across outfits) relates to the distinction and difference from other outfits with regards to how usual, common or ordinary certain outfits are, but not with regards to the levels of “spectacle” and attention associated with them as in the *flamboyant* category (see *flamboyant* for clear differentiation). Highly distinct and different outfits in comparison to other outfits indicated high levels of creativity. They were labeled as unusual, unique, exceptional, extraordinary or different. Outfits without such signs of difference or distinction indicated low levels of creativity and were denoted as usual, standard, common, ordinary or mainstream. Besides, low levels of creativity were also associated with types of outfits representing some kind of basic combination, such as t-shirt and jeans or a mere bikini.

Diverse (within outfits) is concerned with the variety of certain aspects (e.g., color or combinations) within a fashion outfit. Highly diverse outfits were associated with high levels of creativity and were identified based on comments relating to various colors, variations of single items, or the like. Outfits with low levels of diversity were considered uncreative and included labels such as “monotonous” or “single-colored.” Compared to the *contrast (within outfits)* category, this category referred more to the use of various aspects or features instead of a clear contrast between such aspects or features (see *contrast (within outfits)* for clear differentiation).

Ease of fabrication is concerned with the difficulty of production of a certain fashion outfit. This category was only mentioned as a criterion for particularly uncreative outfits, namely in cases in which outfits seemed easy to produce and did not include any aspect of difficulty.

Expensive is concerned with the price of a fashion outfit. This category was only stated as a criterion for especially uncreative outfits, namely in cases in which outfits looked rather cheap or inexpensive.

Flamboyant refers to the levels of extravagance and salience transferred by a fashion outfit. Flamboyant outfits were associated with particularly high levels of creativity. Such outfits were labeled as outlandish, extravagant, striking, or eye-catching, among others. Uncreative outfits were described as unpretentious, unflashy, or unspectacular. This category is very similar to the *distinct/different (across outfits)* category. While the *distinct/different (across outfits)* category is more concerned with the question of how usual and common certain outfits are, this category relates more to the levels of "spectacle" and attention associated with them (see *distinct/different (across outfits)* for clear differentiation).

Functional is concerned with the functionality and combinability of a certain fashion outfit or individual fashion items included in an outfit. This category was only used as a criterion for highly creative outfits. Accordingly, outfits or single items included in an outfit were regarded as particularly creative if they were easily combinable or highly functional.

Imaginative is concerned with the extent of fantasy and inspiration as well as the number of ideas used to develop a fashion outfit. Highly creative fashion outfits were characterized by labels such as imaginative or fanciful. Particularly uncreative fashion outfits were associated with descriptions like "unimaginative," "joyless," or "uninspired."

Interesting is concerned with the question of whether outfits were interesting or not. This category was mentioned as an indicator for particularly creative outfits only. Thus, high creativity was related to interesting outfits.

Love for detail is concerned with the question of whether fashion outfits show high degrees of love for detail or whether they do not. Fashion outfits with considerable love for detail—representing truly creative outfits—were associated with the inclusion of accessories

or tiny features considered as kinds of "x-factors." Outfits not including any (additional) features or accessories, on the other hand, were considered highly uncreative.

Memorable is concerned with the question of whether outfits were memorable or not. High levels of creativity were associated with memorable outfits that observers were likely to keep in mind. Low levels of creativity were represented by outfits that observers could hardly remember.

New refers to the levels of newness and novelty of fashion outfits. Statements associated with high creativity included criteria such as new, original, or innovative, and referred to outfits that were never or seldom seen before. Uncreative outfits were characterized by labels such as vintage, "not innovative," or "often seen before."

Playful refers to the degrees of playfulness and experimentation used in a fashion outfit or in the combination of single fashion items. This category was mentioned as an indicator for particularly creative outfits only. Highly creative outfits were labeled as playful or experimental.

Popular is concerned with the question of whether a certain outfit was currently in trend or popular. This category was only used as a criterion for highly creative outfits, indicating that currently popular outfits were at the same time highly creative.

Proper relates to the general suitability and appropriateness of a fashion outfit or features of that outfit in a certain context, for example concerning the harmony between single items included in an outfit or regarding the suitability of an outfit for a certain (such as everyday) use. Highly creative outfits were labeled as suitable, adequate, harmonious or useful. Highly uncreative fashion outfits were regarded as motley, impractical or inappropriate.

Quantity is concerned with the mere number of certain aspects used in a fashion outfit (e.g., number of colors or number of items). High quantities of certain aspects—indicating

highly creative outfits—included fashion outfits composed of multiple aspects, whereas particularly uncreative outfits were associated with few aspects, for example a maximum of just one or two fashion items included in an outfit.

Revealing relates to outfits that showed high amounts of skin. This category was mentioned as an indicator for particularly uncreative outfits only. Particularly low levels of creativity were associated with outfits allowing for a high visibility of skin.

Surprise is concerned with the levels of surprise and excitement incorporated into a fashion outfit. Particularly surprising or exciting fashion outfits were equally considered highly creative. Very uncreative outfits, on the other hand, were described by labels such as boring or predictable.

Thoughtful is concerned with the degree of thought that was invested into choosing a fashion outfit or combining single fashion items. This category was mentioned as an indicator for particularly creative outfits only. Accordingly, if a certain outfit and the combination of the fashion items included in it seemed to be carefully considered and chosen, it indicated a high level of creativity.

Updates tradition refers to incorporating traditional fashion items into a current fashion outfit. This category was only used as a criterion for highly creative outfits. Raters using this criterion suggested that outfits including some kind of vintage or old item were highly creative.

Appendix B

Description of Frames

Appeal/likeability is composed of only one category and is in essence no frame. The category *appeal/likeability* relates to the likeability of a fashion outfit with regard to its various aspects, for example its color, its sewing pattern or the items included in it.

Appropriate includes the categories *proper* and *thoughtful*. This frame points out how a fashion outfit or features of it are appropriate in a given context. It includes criteria such as the suitability of an outfit (with regard to various aspects) as well as the careful consideration of the items included in an outfit.

Artful includes the categories *artistic*, *imaginative*, *love for detail*, and *playful*. This frame integrates categories relating to small, artful details and features of an outfit as well as to the willingness to try out and use inspiration.

Noticeable includes the categories *contrast (within outfits)*, *courageous*, *flamboyant*, *interesting*, *memorable*, *revealing*, and *surprise*. In a broad sense, this frame comprises categories that are concerned with the prominence, salience or excitement of a fashion outfit.

Original includes the categories *distinct/different (across outfits)* and *new*. This frame is composed of categories referring to the unusualness or the novelty of fashion outfits.

Variety includes the categories *diverse (within outfits)* and *quantity*. This frame thus comprises categories concerned with the variety and mere number of certain aspects in a given fashion outfit.

Other is not a frame of similar meanings in itself, but rather a collection of categories that did not share a clearly identifiable meaning with any other category. This collection includes categories that were mentioned only infrequently, namely *authentic*, *ease of fabrication*, *expensive*, *functional*, *popular*, and *updates tradition*.

Appendix C

Frames and Criteria Categories Used to Evaluate Fashion Outfits

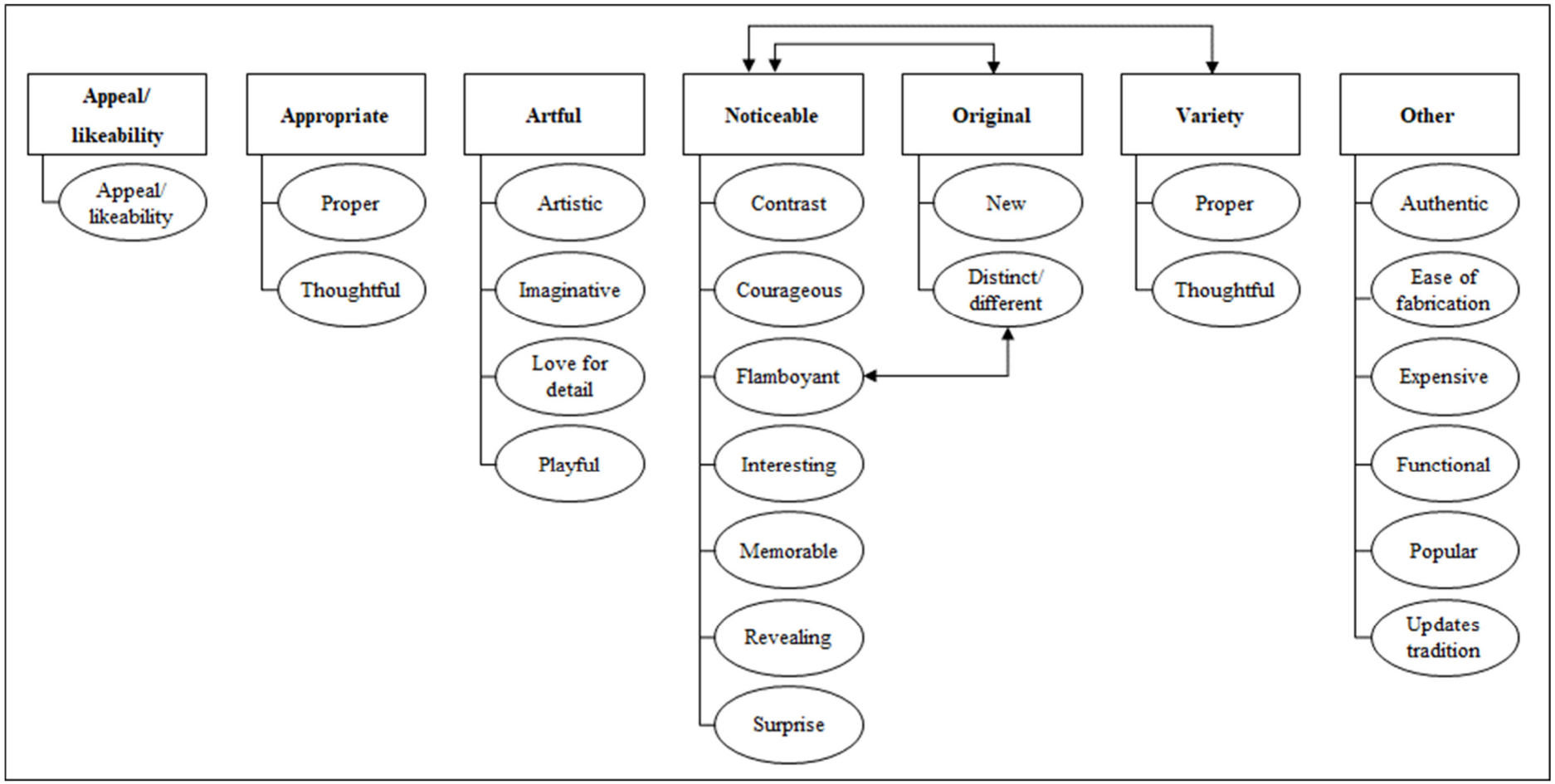

Figure C1. Frames and Criteria Categories. Each frame and each category is illustrated using the same (randomly selected) line pattern. The

frames are rectangular, whereas the categories are illustrated by circles. The relationships between frames and categories are shown by lines. In case of a relation across frames or across categories (i.e., in case there are similarities or ambiguities), a two-sided arrow is used. Such similarities or ambiguities are resolved in Appendices A and B.

Tables

Table 1. Levels of Creativity Ratings: Comparison Between Genders

		Female raters (n = 39)		Male raters (n = 26)		Gender differences		
Sample	j	M	SD	M	SD	$t(63)$	p	Cohen's d
Main sample	90	3.99	0.51	4.09	0.44	-0.80	.429	-0.20
Type of outfit								
- Standard	30	3.08	0.67	3.32	0.67	-1.35	.183	-0.34
- Non-standard	30	5.31	0.69	5.00	0.64	1.79	.079	0.45
- Revealing	30	3.58	0.76	3.95	0.69	-1.96[A]	.054	-0.50
Type of user								
- >100,000 followers	45	4.12	0.50	4.18	0.45	-0.44	.663	-0.11
- < 10,000 followers	45	3.86	0.58	4.00	0.48	-1.02	.311	-0.26

Notes. Differences in mean creativity ratings between female and male raters were tested for significance using independent means t tests (two-tailed). j = number of outfits.

[A] Based on a Shapiro−Wilk test, the data from the female rater sample did not appear to be normally distributed. Therefore, we additionally applied a two-tailed Mann−Whitney U test for nonparametric data. Results are consistent with those of the independent means *t* test ($U = 645$, $p = 0.065$, $r = .23$).

Table 2. Inter-Rater Reliability (Cronbach's Coefficient Alpha): Comparison Between Genders

		Female raters (n = 39)			Male raters (n = 26)			Gender differences		
Sample	j	α_F	*LL* of 95% CI	*UL* of 95% CI	α_M	*LL* of 95% CI	*UL* of 95% CI	*LL* of 95% CI[A]	*UL* of 95% CI[A]	Test Result[B]
Main sample	90	.97	.96	.98	.91	.88	.93	.03	.09	$\alpha_F > \alpha_M$
Type of outfit										
- Standard	30	.93	.89	.96	.83	.73	.90	.02	.21	$\alpha_F > \alpha_M$
- Non-standard	30	.92	.88	.96	.66	.43	.81	.11	.49	$\alpha_F > \alpha_M$
- Revealing	30	.91	.85	.95	.82	.71	.90	-.01	.21	Inconclusive
Type of user										
- >100,000 followers	45	.96	.95	.98	.90	.86	.93	.03	.11	$\alpha_F > \alpha_M$
- < 10,000 followers	45	.97	.96	.98	.92	.89	.95	.02	.09	$\alpha_F > \alpha_M$

Notes. Differences in Cronbach's coefficient alphas between female and male raters were tested for significance using the Bonett and Wright (2015) approach. α_F = Cronbach's coefficient alpha for female raters. α_M = Cronbach's coefficient alpha for male raters.

j = number of outfits.

[A] 95% confidence interval for difference of two Cronbach's coefficient alpha reliabilities based on Bonett and Wright (2015).

[B] Hypothesis test of H_0: $\alpha_F = \alpha_M$ with a significance level of 5% based on decision rules provided by Bonett and Wright (2015): If $LL > 0$, then reject H_0 and accept $\alpha_F > \alpha_M$. If $UL < 0$, then reject H_0 and accept $\alpha_F < \alpha_M$. Otherwise, results are inconclusive.

Table 3. Inter-Rater Reliability (Spearman–Brown Adjusted Alpha): Comparison Between Genders

		Female raters (n = 10)			Male raters (n = 10)			Gender differences		
Sample	j	$SB\alpha_F$	*LL* of 95% CI	*UL* of 95% CI	$SB\alpha_M$	*LL* of 95% CI	*UL* of 95% CI	*LL* of 95% CI[A]	*UL* of 95% CI[A]	Test Result[B]
Main sample	90	.89	.88	.90	.80	.77	.82	.06	.11	$SB\alpha_F > SB\alpha_M$
Type of outfit										
- Standard	30	.77	.74	.81	.65	.57	.74	.03	.21	$SB\alpha_F > SB\alpha_M$
- Non-standard	30	.75	.72	.80	.42	.23	.60	.16	.53	$SB\alpha_F > SB\alpha_M$
- Revealing	30	.71	.67	.77	.64	.55	.73	-.03	.18	Inconclusive
Type of user										
- >100,000 followers	45	.87	.86	.89	.78	.74	.82	.05	.14	$SB\alpha_F > SB\alpha_M$
- < 10,000 followers	45	.90	.89	.91	.82	.79	.85	.05	.11	$SB\alpha_F > SB\alpha_M$

Notes. Cronbach's coefficient alphas were standardized to an equal panel size of 10 based on the Spearman–Brown prophecy formula. Differences in Spearman–Brown adjusted alphas between female and male raters were tested for significance using the Bonett and Wright (2015) approach. j = number of outfits. $SB\alpha_F$ = Spearman–Brown adjusted alpha for female raters. $SB\alpha_M$ = Spearman–Brown adjusted alpha for male raters.

[A] 95% confidence interval for difference of two Spearman–Brown adjusted alpha reliabilities based on Bonett and Wright (2015).

[B] Hypothesis test of H_0: $SB\alpha_F = SB\alpha_M$ with a significance level of 5% based on decision rules provided by Bonett and Wright (2015): If $LL > 0$, then reject H_0 and accept $SB\alpha_F > SB\alpha_M$. If $UL < 0$, then reject H_0 and accept $SB\alpha_F < SB\alpha_M$. Otherwise, results are inconclusive.

Table 4: Statements Concerning Fashion Outfits: Breakdown of Frames

	All raters (N = 65)		Female raters (*n* = 39)		Male raters (*n* = 26)		Gender differences	
Frames of similar meaning	n^A	%	n^A	%	n^A	%	$\chi^2(1)$	*p*
Appeal/likeability	9	2%	3	1%	6	3%	3.20	.074
Appropriate	47	10%	15	5%	32	18%	20.10	<.001
Artful	25	5%	21	7%	4	2%	5.44	.020
Noticeable	117	26%	81	29%	36	21%	3.52	.061
Original	168	37%	102	36%	66	38%	0.18	.671
Variety	80	18%	53	19%	27	16%	0.75	.386
Other	9	2%	7	2%	2	1%	0.97	.324
Total	455[B]	100%	282	100%	173	100%	–	–
Top 3[C]	365	80%	236	84%	134	77%	2.74	.098

Notes. All distinct statements provided by female and male raters on why outfits were regarded as particularly creative or uncreative were collected, allocated to a category and then subsumed under seven frames of similar meaning (see Appendix A for descriptions of the criteria categories and Appendix B for descriptions of the frames). Table 4 illustrates the number of statements and the respective proportions assigned

to these frames. Differences between the proportions assigned to the various frames by female and male raters were tested for significance using chi-square tests of independence. The analysis was based on all statements provided by the raters even if a certain category had been mentioned multiple times by a single rater (weighting of frequencies). An additional analysis based on distinct categories only (elimination of categories mentioned multiple times by a single rater, i.e. no double counting) yielded similar results.

[A] *n* refers to the number of statements assigned to each frame.

[B] All statements that could not be allocated to a category (31) were not included in the analysis.

[C] Top 3 frames include *original*, *noticeable*, and *variety* for female raters and *original*, *noticeable*, and *appropriate* for male raters