

If you're funny and you know it: Personality, gender, and people's ratings of their attempts at humor

By: [Paul J. Silvia](#), Gil Greengross, Katherine N. Cotter, Alexander P. Christensen, and Jeffrey M. Gredlein

Silvia, P.J., Greengross, G., Cotter, K.N., Christensen, A.P., & Gredlein, J.M. If you're funny and you know it: Personality, gender, and people's ratings of their attempts at humor. *Journal of Research in Personality* 92, 104089. <https://doi.org/10.1016/j.jrp.2021.104089>

***© 2021 Elsevier Inc. Reprinted with permission. This version of the document is not the version of record. ***



This work is licensed under a [Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License](#).

Abstract:

In seven studies ($n = 1133$), adults tried to create funny ideas and then rated the funniness of their responses, which were also independently rated by judges. In contrast to the common “funnier than average” effect found for global self-ratings, people were relatively modest and self-critical about their specific ideas. Extraversion ($r = 0.12$ [0.07, 0.18], $k = 7$) and openness to experience ($r = 0.09$ [0.03, 0.15], $k = 7$) predicted rating one's responses as funnier; women rated their responses as less funny ($d = -0.28$ [-0.37, -0.19], $k = 7$). The within-person correlation between self and judge ratings was small but significant ($r = 0.13$ [0.07, 0.19], $k = 7$), so people had some insight into their ideas' funniness.

Keywords: Humor | Comedy | Creativity | Personality | Gender | Discernment

Article:

More than a few college students have complained that their professor isn't as funny as they think they are, but having an inflated notion of one's humor ability isn't merely an academic problem. Long ago, Omwake (1937) proposed that people “are more reluctant to admit a defective sense of humor than a poor ear for music, a lack of physical skill or endurance, or even an inferior intelligence” (p. 692), and research bears this out. When asked for global judgments of their humor skills, people show an overly positive and self-serving bias. Nevertheless, people's judgments of their global features can vary greatly from their judgments of specific ideas, skills, and behaviors, especially when they are judging something that is concrete and easily compared (Alicke & Govorun, 2005). In the present research, then, we examined the overlooked question of how people judge the funniness of their specific attempts at humor. People might view themselves as “funnier than average” overall, but are they more modest and circumspect when judging their specific jokes and quips?

1.1. People see themselves as “funnier than average”

The “better-than-average effect”—a pervasive finding in social judgment research—refers to people’s tendency to see themselves as better than average when asked to rate themselves on a range of traits, abilities, and features (Alicke & Govorun, 2005). Some of the oldest demonstrations of the better-than-average effect, it turns out, come from the domain of humor. In her landmark study, Omwake (1937) asked a sample of 599 high-school and college students to “estimate your position in your class” (e.g., freshman, sophomore) relative to the average for a broad range of traits. Many of the traits reflected creating and delivering humor, such as “ability to feel at ease while telling a joke” and “ability to improve on jokes which you have heard.” Other traits reflected appreciating humor, such as “tendency to enjoy a clever dirty pun” and “tendency to enjoy a good Scotchman joke.” Of all the traits, “possession of a sense of humor” received the second-highest rating (after “possession of a good appetite,” curiously enough). As a group, the students rated themselves as much better than their peers in sense of humor: only 1.4% rated themselves as below average, and 25% gave themselves the highest scale score. This study and others (e.g., Chapman & Gadfield, 1976; Fine, 1975; Lefcourt & Martin, 1986) illustrate the classic better-than-average effect in self-perceptions of humor.

Although people overall rate themselves as having a good sense of humor, their ratings are affected by many variables. Research on personality and self-ratings of global humor, for example, shows a role for extraversion and openness to experience. People higher in those traits are more likely to rate themselves as funny people (Beins & O’Toole, 2010), to see being funny as important to their identity (Silvia, Rodriguez, & Karwowski, 2020), and to say they are good at amusing their friends (i.e., endorsing an affiliative humor style; Plessen et al., 2020). Smaller but notable effects have been found for other traits. People high in conscientiousness, for example, report lower humor self-efficacy (confidence in making others laugh; Silvia et al., 2020).

Gender, a key variable in the psychology of humor (Greengross, 2020, Martin, 2014), shapes people’s views of how funny they are. American culture has a pervasive stereotype that “women aren’t funny” (Hooper et al., 2016, Mickes et al., 2012), which is perhaps best bookended by Christopher Hitchens’s (2007) infamous “Why Women Aren’t Funny” and Jilly Gagnon’s (2013) satirical “Reasons Women Aren’t Funny.” Research finds that women’s self-concepts reflect this cultural stereotype. In American and UK samples, for example, women have lower humor self-efficacy (Caldwell and Wojtach, 2020, Silvia et al., 2020) and view being a funny person as less central to their self-concept (Silvia et al., 2020). A meta-analysis of lab-based humor production studies found a small advantage for men (Greengross, Silvia, & Nusbaum, 2020), so these self-beliefs have parallels in behavioral contexts.

1.2. Judging global humor vs specific ideas

People are inclined to see themselves as “funny people,” but how do they view their “funny ideas”? The better-than-average effect has many moderators (Alicke and Govorun, 2005, Kruger, 1999), and the most relevant for the present work is *trait ambiguity*. Better-than-average effects are much larger when people judge qualities that are hard to pin down. People show stronger better-than-average effects for traits like morality than for intelligence, for example, because traits like morality are complex, vague, and difficult to measure, whereas traits like intelligence have relatively concrete benchmarks (Allison et al., 1989, Zell et al., 2020). For this reason,

better-than-average effects are much weaker, if they appear at all, when people judge concrete traits, specific behaviors, or tangible products. This disparity between judgments of abstract traits versus specific behaviors follows from the different ways that people represent trait concepts versus autobiographical knowledge of past behaviors (Klein and Loftus, 1993, Klein et al., 1992).

A trait like “sense of humor” has all the hallmarks of an ambiguous trait: it is abstract, subject to idiosyncratic definitions, hard to benchmark relative to other people, and can’t be disconfirmed by any single behavioral event (see Dunning, Meyerowitz, & Holzberg, 1989). Nevertheless, the better-than-average literature suggests that people’s ratings of specific behaviors, such as jokes they created or times when they tried to make others laugh, may be much more modest and realistic because the specific behaviors and ideas are more concrete and less ambiguous. People’s views of their specific attempts at humor wouldn’t necessarily show a self-serving pattern despite the well-established funnier-than-average effect for global humor traits.

Aside from the interesting question of whether people’s judgments of global and specific humor diverge, how people judge the funniness of their attempts at humor ties into some important theoretical issues in the psychology of creative thought. Theories of creativity point out that success in any creative domain—whether it is coming up with ideas for research studies, billboards, song lyrics, or one-liner jokes—requires not only generating ideas but also exercising sound judgment about which ones are worth sharing and which ones should never see the light of day (Sawyer, 2012, Weisberg, 2020). In all creative domains, people’s attempts to come up with good ideas are a mix of hits and misses (Simonton, 1999, Weisberg, 2020). The craft of creativity requires discerning the difference between a likely hit and a likely miss so that one can develop, refine, and share one’s best ideas (Cropley, 2006, Karwowski et al., 2020, Kozbelt, 2007, Silvia, 2008). Idea evaluation—judging, refining, and discarding one’s ideas—is thus a major theme in theories of creativity, from classic models to modern cognitive neuroscience (Beaty et al., 2016, Kleinmintz et al., 2019, Weisberg, 2020).

Humor, as a form of creative thought, shares this interplay of generation and evaluation. It is one thing to come up with a bunch of possibly funny jokes, but another entirely to weed out the likely duds, select the most promising ones, and then hone them to be even funnier. Understanding how people judge the funniness of their ideas can shed light on the links between humor and the larger creativity literature as well as on the effective use of humor in social interactions, which surely must require a reasonably discerning sense of whether an attempt to be funny is likely to work.

Who judges their own ideas as funny? Essentially nothing is known about how people judge their own humor attempts, but the broader literature on creative discernment offers some guidance for how to unpack and study the issue (Berg, 2019, Dean et al., 2006, Grohman et al., 2006, Kozbelt, 2007, Silvia, 2008). A typical creative discernment paradigm asks people to come up with creative ideas and then to rate or sort the creativity of their own responses. Independent judges then rate the creativity of each response, and discernment is indicated by the covariation of the participants’ and judges’ ratings.

For example, Silvia (2008) recruited a sample of 226 American college students and asked them to complete four divergent thinking tasks (unusual uses for a brick and a knife, and creative instances of things that are round and that make a noise). After generating their responses, people selected their two best responses for each task. Three judges then scored all the responses on a 5-point scale (1 = *not at all creative*, 5 = *very creative*). A multilevel latent-variable analysis found a significant within-person correlation between whether someone picked a response as a “top two” idea and the judges’ numerical ratings ($\beta = 0.31$)—the ideas that people saw as their best were indeed rated higher by the judges.

In a later study, Karwowski et al. (2020) recruited a sample of 500 Polish adults (ages 26–46) and asked them to complete three divergent thinking tasks (unusual uses for tape, a can, and a brick). For each task, participants rated the creativity of the set of responses on a pair of items (“My answers were creative” and “My answers were more original than those of my peers”), using a 0–100 scale. The sets of responses were also scored by three trained judges, who used 7-point scale (1 = *not creative at all*, 7 = *very creative*). Latent variable models found that the judges’ creativity ratings covaried significantly with people’s own ratings of their ideas ($\beta = 0.38$).

The body of work on creative discernment thus offers some methodological tools and empirical guidance for the related problem of self-rated humor. First, people show within-person variability in their ratings of their ideas—they rarely think all their ideas are great and instead give differentiated judgments (Silvia, 2008). Second, people have at least some insight into the creative effectiveness of their ideas. Research using many methods and contexts finds that people aren’t guessing at chance—there is at least modest covariation between self and judge ratings, a finding known as *creative discernment* (Silvia, 2008) and *creative metacognition* (Karwowski et al., 2020, Kaufman and Beghetto, 2013).¹ People can judge the creativity of their ideas at better-than-chance levels by applying metacognitive tools and judgment heuristics to weed out the most obviously bad ideas and identify relevant markers of idea quality (Puente-Díaz et al., 2021, Stemler and Kaufman, 2020).

1.3. The present research

In the present research, we explored people’s ratings of their own attempts to be funny. We focused on three issues: (1) Do personality traits and gender predict people’s views of the funniness of their ideas? Who is more self-critical?; (2) Are people discerning about their attempts at humor? Do their ratings covary with scores given by independent judges?; and (3) Do personality and gender moderate discernment? Do some people show better agreement with the judges about how funny their ideas are? This was an exploratory study with no specific predictions. To test these questions, we report findings from seven similar studies that we conducted on humor production (total $n = 1133$), all of which measured personality, gender, and humor performance. Instead of reporting all seven studies individually, however, we used statistical tools from meta-analysis to provide a concise and compact summary of our studies.

2. Method

2.1. Participants

We combined the studies in our line of research on humor production that measured gender, personality, and self-ratings of the funniness of one’s own ideas. One study was omitted because it manipulated the humor generation tasks in ways that complicate and obscure individual differences (Shin, Cotter, Christensen, & Silvia, 2020); all our other datasets were included. Table 1 describes each sample. Five samples had appeared in publications primarily about humor; two samples had unpublished humor data, collected during our early forays into humor research, that were collected as part of other projects (Diedrich et al., 2018, Nusbaum et al., 2015, Study 1). None of the prior publications analyzed or reported the self-ratings of humor.

Table 1. Summary of the samples.

| Sample | <i>n</i> | Personality Measure | Humor Tasks | Raters & Rating Scale | % Female | Age | Notes |
|---|----------|---------------------|----------------------------------|---|----------|----------------|---|
| 1. NCSA Study (NCSA; Diedrich et al., 2018) | 132 | BFI-10 | 2 (Jokes) | 3 raters (2 female), 1–5 scale, $\omega = 0.55$ | 42 | 20.10 (18, 49) | This sample has unusually high O scores, consistent with a sample of arts students. |
| 2. Physical Anhedonia (PHYS.AN; Nusbaum et al., 2015) | 185 | NEO-FFI | 3 (Jokes) | 2 raters (1 female), 1–5 scale, $\omega = 0.64$ | 70 | 19.65 (18, 32) | |
| 3. CHC & Humor Study (CHC; Christensen et al., 2018) | 212 | NEO FFI | 9 (Cartoons, Jokes, Definitions) | 3 raters (1 female), 1–5 scale, $\omega = 0.73$ | 85 | 19.11 (18, 48) | |
| 4. Ha Ha, 1 (PACA_S1; Nusbaum et al., 2017) | 142 | NEO FFI | 6 (Cartoons, Jokes) | 4 raters (2 female), 1–5 scale, $\omega = 0.71$ | 81 | 18.79 (18, 40) | The Resumes task reported in the article did not have self-ratings. |
| 5. Ha Ha, 2 (PACA_S2; Nusbaum et al., 2017) | 147 | NEO FFI | 6 (Cartoons, Jokes) | 5 raters (3 female), 1–5 scale, $\omega = 0.75$ | 72 | 19.07 (18, 32) | |
| 6. Ha Ha, 3 (PACA_S3; Nusbaum et al., 2017) | 129 | HEXACO-100 | 9 (Cartoons, Jokes, Definitions) | 2 raters (1 female), 1–5 scale, $\omega = 0.63$ | 69 | 18.49 (18, 24) | |
| 7. RWA & Humor (RWA; Silvia et al., 2021) | 186 | HEXACO-100 | 9 (Cartoons, Jokes, Definitions) | 3 raters (1 female), 0–2 scale, $\omega = 0.69$ | 78 | 19.07 (18, 53) | Only the 3 raters who scored all 9 tasks were included. |

Notes. For gender, the percent of participants identifying as female is reported. For age, the mean and min/max values are reported. For more details about the samples, see the original publications: NCSA (Diedrich et al., 2018), Physical Anhedonia (Nusbaum et al., 2015), CHC Humor (Christensen et al., 2018), Ha Ha 1–3 (Nusbaum et al., 2017, Studies 1–3), and RWA (Silvia et al., 2021). The first six samples had raters score responses on a 1–5 scale; the sixth used a 0–2 scale. The sample sizes may vary slightly from the original publications due to more stringent exclusion criteria. The raw data and input files are available at Open Science Framework (<https://osf.io/57nvg/>).

The participants across the seven samples consisted of 1133 adults who were native English speakers; 76% identified as female. Except for the first sample, which consisted of college students with a variety of arts majors enrolled at UNCSA (Sample 9 in Diedrich et al., 2018), all the participants were students enrolled in psychology courses at UNCG (Christensen et al., 2018, Nusbaum et al., 2015, Nusbaum et al., 2017, Silvia et al., 2021). Regarding power and planned sample size, we did not have a priori expectations for effect sizes, but from the beginning of our humor work we had planned to accumulate data on self-ratings in each study we ran until we had at least 500 people. The long-term interruption of our lab research due to the Covid-19 pandemic struck us as a natural stopping point to wrangle the data for analysis.

2.2. Procedure

Humor tasks. Each study followed a similar general procedure. The instructions explained that the study was about humor and how people come up with funny ideas. Just as divergent thinking tasks use “be creative” instructions (Nusbaum et al., 2014, Said-Metwaly et al., 2020), we encouraged people to “be funny” by aiming for funny responses. Humor production tasks present people with a prompt that sets up an opportunity to be funny (Ruch & Heintz, 2019). Three tasks, developed by Nusbaum, Silvia, and Beaty (2017) and available online (<https://osf.io/4s9p6/>), were used. Table 1 notes the tasks and number of prompts used in each study.

In the *cartoon captions task*, people saw a one-panel cartoon with the caption removed, and they were asked to write a funny caption for it. The cartoons depicted an astronaut on the moon speaking into a cell phone, a psychotherapist seated next to a king on a couch, and a man in an office holding a smoking gun next to a body on the floor. In the *joke stems task*, people read a scenario that set up a joke, such as eating something terrible that a friend cooked and then describing what it was like, giving honest feedback to a friend about their terrible singing, and trying to describe what it was like to sit in a painfully boring class. For each joke stem, people completed the set-up with a funny ending. Finally, in the *definitions task*, people were given a quirky noun-noun pair (e.g., *yoga bank*, *cereal bus*, *fruit jar*) and asked to give it a funny definition. In all cases, people gave only one response per prompt.

Self-ratings of humor. After giving their response, people were asked to rate how funny they thought it was. For each response, they completed a single item tailored to the task:

- “In your opinion, how funny is your caption?”
- “In your opinion, how funny is your joke?”
- “In your opinion, how funny is your definition?”

They responded to the item using a 5-point scale (1 = *not at all funny*, 5 = *very funny*). Their response was visible to them during the self-rating.

Judge ratings of humor. In all studies, all responses were subjectively scored for humor. The raters gave each response an overall, holistic funniness rating while unaware of other information about the participant (e.g., gender or personality scores) and unaware of the other raters’ scores. Each rater provided a score for each response from all participants. The number of raters, shown in Table 1, ranged from 2 to 5, and all samples were rated by both male and female raters. In most studies, the raters used a 5-point (1–5) funniness scale; in the most recent sample, the raters used a 3-point (0–2) funniness scale because Rasch rating-scale analyses found that 3 categories is probably optimal (Primi et al., 2019, Silvia et al., 2021). Reliability for the ratings was estimated with coefficient omega. To account for the faceted design (i.e., raters made many judgments per person), these were estimated using cluster-robust standard-error models, as described in the Results.

Measures of individual differences. Participants self-identified their gender as male (0) or female (1). Personality was assessed with one of three scales: (1) the NEO-FFI (Costa &

McCrae, 1992), which measures the five domains with 12 items each; (2) the HEXACO-100 (Lee & Ashton, 2018), which measures the six factors in the HEXACO model of personality with 16 items per factor; or (3) the BFI-10 (Rammstedt & John, 2007), which measures each factor with 2 items. The traits were scored using item averages.

Table 1 notes which sample used which scales. One sample used the BFI-10, two samples used the HEXACO-100, and four samples used the NEO-FFI. We should note that the one sample that used the brief BFI-10—arts students enrolled at UNCSA—was the most distinctive one. As an arts university, UNCSA attracts a much different student population than UNCG, one that has a history of creative achievements and that is unusually high in openness to experience (Diedrich et al., 2018). Things unique to the BFI-10 thus cannot be separated from the sample. Of these scales, only the HEXACO-100 affords facet scales. Interested readers can find facet-level findings for the two samples using the HEXACO-100 in online supplemental material (<https://osf.io/57nvg/>). Finally, these three personality scales draw upon different underlying models of personality traits, and their distinctive and overlapping features have been extensively discussed (Ashton et al., 2014, Ashton and Lee, 2019, Johnson, 1994, Miller et al., 2011). We do not quantitatively contrast the scales in the Results, but we note issues unique to the scales and their different senses of the traits when relevant.

3. Results

The effect sizes were analyzed in R 4.0 (R Core Team, 2020) using *meta* (Schwarzer, 2020). Effects for gender were expressed as the standardized mean difference (Cohen's *d*); effects for personality traits were expressed in the *r* metric. Correlations were analyzed using *z*-transformed values that were back-transformed to *r*. The meta-analysis weighted the effect sizes using the inverse variance method and estimated tau using maximum likelihood. Both fixed-effects and random-effects models were estimated. A random-effects model is generally preferred, but in many cases the software constrained I^2 to zero, which can happen when an analysis has a small and homogenous set of effects. We thus report the fixed-effects models in the text, but both the fixed and random models are depicted in the figures and any salient differences are noted in the text. The data and R files are available on Open Science Framework (<https://osf.io/57nvg/>).

Calculating the effect sizes is relatively more complicated for these samples because (1) people gave between 2 and 9 self-ratings, one for each humor item, and (2) the judges' ratings of humor introduce a facet into the data (Primi et al., 2019). For the effects of gender and personality on self-ratings, the effect sizes were estimated using cluster-robust standard-error (CR-SE) models (McNeish et al., 2017, Wu and Kwok, 2012), a design-based approach to nested data that affords standardized estimates corrected for clustering, in Mplus 8.4 using maximum likelihood with robust standard errors. These models were also used to estimate omega reliability for the raters.

For effect sizes involving the judges' ratings—whether people's self-ratings correlated with the judges, and whether individual differences moderated the correlation—we used multilevel models. Because the judges' scores are highly skewed and ordinal (see Nusbaum et al., 2017), they were treated as categorical indicators of a latent humor-score variable. This latent variable was an outcome at Levels 1 and 2. Participants' self-rated humor was a group-mean centered

predictor at Level 1. Its effect thus represents how much the latent humor score changes as a person’s rating changes relative to their own mean self-rating—i.e., a within-person relationship between self-appraised funniness and judge-rated funniness. This model yields for each participant an estimated intercept (how funny the judges rated their ideas) and a slope (the covariation between the person’s ratings and the judges’ ratings). Gender and personality traits were grand-mean centered variables at Level 2, and the correlations involving the intercepts, slopes, and Level 2 factors were estimated. To obtain standardized effects (r and d) in a multilevel framework, we used Bayesian Markov Chain Monte Carlo estimation in Mplus 8.4, with at least 5000 iterations of Gibbs sampling, a potential scale reduction criterion of 0.05, and thinning to every 10th step.

3.1. How did people rate their ideas?

We first explored the distribution of self-ratings to see what people thought about the funniness of their ideas. Fig. 1 shows the distribution of all humor self-ratings in the seven samples. People were generally modest: the mode was 3 on the 1–5 scale, and low scores (1, 2) were more common than high scores (4, 5). This distribution suggests that the pervasive “funnier-than-average” bias found for global self-ratings of humor doesn’t appear when people are asked to rate their concrete attempts at humor.

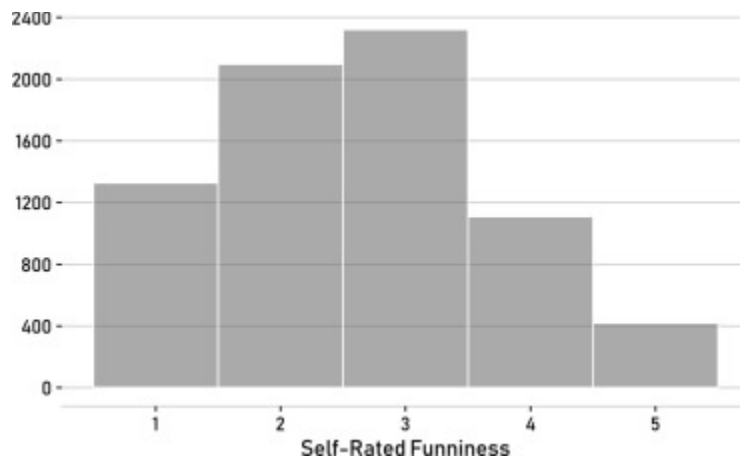


Fig. 1. Distribution of self-rated funniness across all responses.

3.2. Who rated their ideas as funny?

What predicted between-person variation in self-rated humor? Gender was an important factor (see Fig. 2). Compared to men, women rated their responses as less funny, $d = -0.28$ 95% CI $[-0.37, -0.19]$, $k = 7$. Women thus appeared to be more critical of their humor responses, consistent with past research on women’s lower humor self-efficacy (Silvia et al., 2020) and humor performance (Greengross et al., 2020).

| Source | SMD (95% CI) |
|---|---------------------|
| 6.PACA_S3 | -0.22 [-0.54; 0.10] |
| 4.PACA_S1 | -0.16 [-0.44; 0.12] |
| 2.PHYS_AN | -0.14 [-0.59; 0.31] |
| 1.NCSA | -0.10 [-0.60; 0.40] |
| 3.CHC | -0.05 [-0.30; 0.19] |
| 5.PACA_S2 | 0.09 [-0.24; 0.42] |
| 7.RWA | 0.18 [-0.07; 0.43] |
| Total (fixed effect) | -0.03 [-0.15; 0.08] |
| Total (random effects) | -0.03 [-0.15; 0.08] |
| Heterogeneity: $\chi^2_6 = 5.80$ ($P = .45$), $I^2 = 0\%$ | |

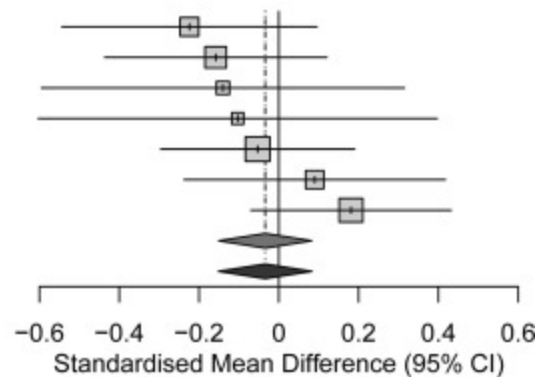


Fig. 2. Gender and the self-rated funniness of one's ideas. **Note.** Negative effect sizes indicate that women gave lower self-ratings than men.

For personality and self-ratings of funniness, two of the six traits—extraversion and openness to experience—had effect sizes with confidence intervals excluding zero (see Fig. 3). People rated their ideas as being funnier when they were more extraverted ($r = 0.12$ [0.07, 0.18], $k = 7$) and more open to experience ($r = 0.09$ [0.03, 0.15], $k = 7$); the effect sizes were small in magnitude. The effects for agreeableness, conscientiousness, honesty-humility, and neuroticism were non-significant and very small.

3.3. Discernment: correlation of self-ratings and judges' ratings

How discerning were people in their self-ratings? An analysis of the within-person correlations between self-ratings and the judges' scores found a small, positive effect, $r = 0.13$ [0.07, 0.19], $k = 7$ (see Fig. 4). Because these are within-person correlations, their interpretation is unconfounded by between-person differences (e.g., humor ability or self-critical tendencies). Relative to their average rating, the ideas that people rated more highly were also likely to receive higher ratings from the judges. The effect size was small in magnitude, so self and judge ratings were significantly but weakly related.

Overall, then, people's self-ratings had a small within-person correlation with the judges' ratings. What moderated the strength of this relationship? One possibility is that people who were funnier (i.e., had higher ratings from the judges) had a stronger self-judge correlation. To appraise this possibility, we analyzed the correlation between the intercept and slope (i.e., the random intercept reflecting judges' ratings and the distribution of random slopes reflecting self-judge relatedness). The results were weak and inconsistent. As Fig. 5 shows, although the confidence intervals for the fixed-effect model excluded zero ($r = 0.08$ [0.02, 0.13], $k = 7$), the effect sizes were highly variable, as reflected in the random-effects model ($r = 0.10$ [-0.13, 0.32], $k = 7$). Given the homogeneity of the participants, tasks, and judges across the seven samples, we conclude that there's a lack of evidence for funnier people being more discerning.

Finally, did gender or personality moderate the correlation between self and judge ratings of humor? No evidence was found for moderation by gender ($d = -0.03$ [-0.15, 0.08], $k = 7$), as Fig. 6 shows. Although women rated their ideas as less funny, their ratings were not more or less tightly correlated with the judges' ratings. Women and men were thus equally discerning about the funniness of their ideas.

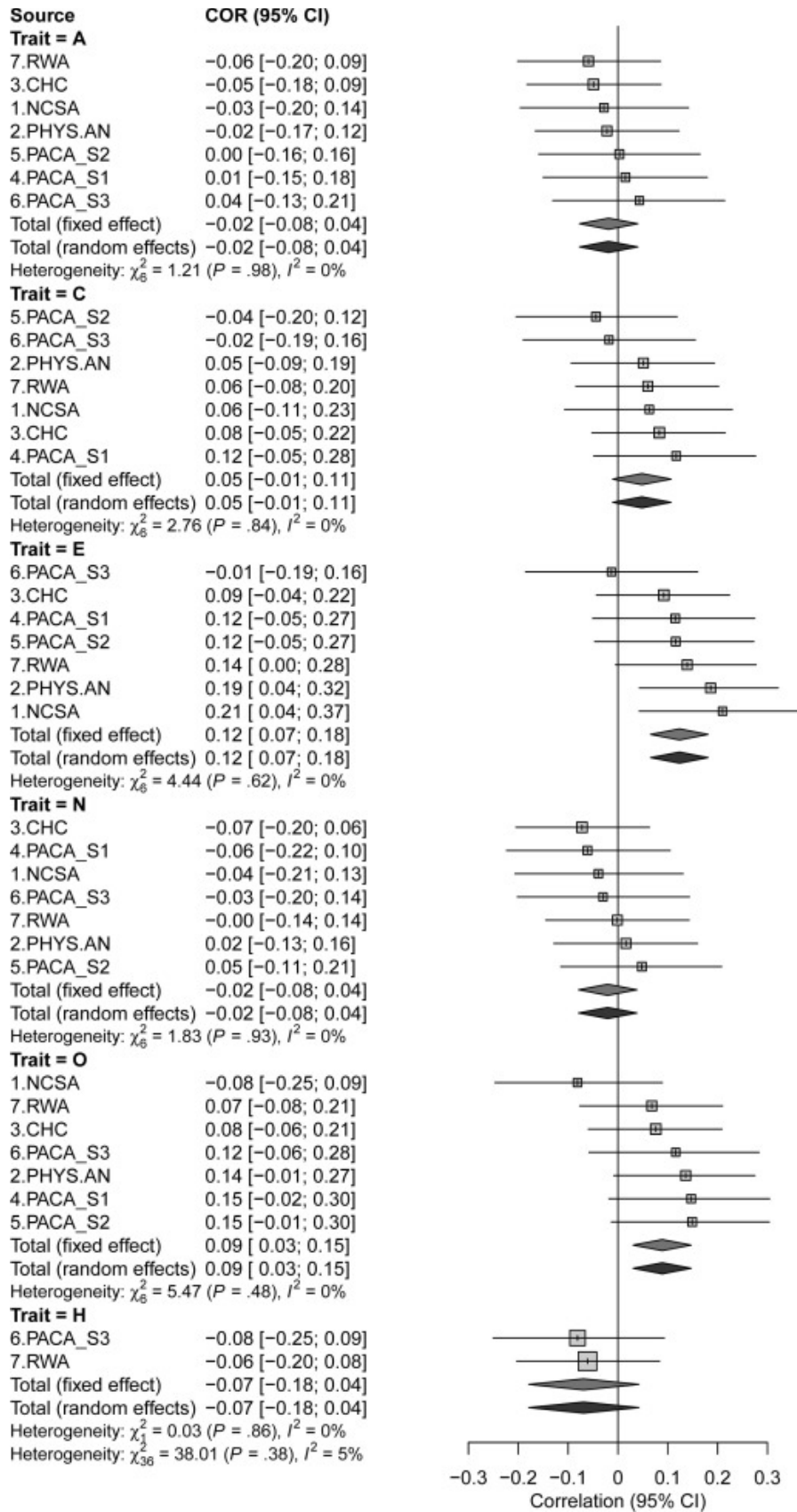


Fig. 3. Personality traits and the self-rated funniness of one's ideas.

| Source | COR (95% CI) |
|---|---------------------|
| 1.NCSA | -0.01 [-0.18; 0.15] |
| 2.PHYS.AN | 0.02 [-0.12; 0.17] |
| 6.PACA_S3 | 0.09 [-0.08; 0.26] |
| 3.CHC | 0.11 [-0.03; 0.24] |
| 5.PACA_S2 | 0.19 [0.03; 0.34] |
| 4.PACA_S1 | 0.20 [0.04; 0.36] |
| 7.RWA | 0.28 [0.14; 0.41] |
| Total (fixed effect) | 0.13 [0.07; 0.19] |
| Total (random effects) | 0.13 [0.06; 0.20] |
| Heterogeneity: $\chi^2_6 = 11.06$ ($P = .09$), $I^2 = 46\%$ | |

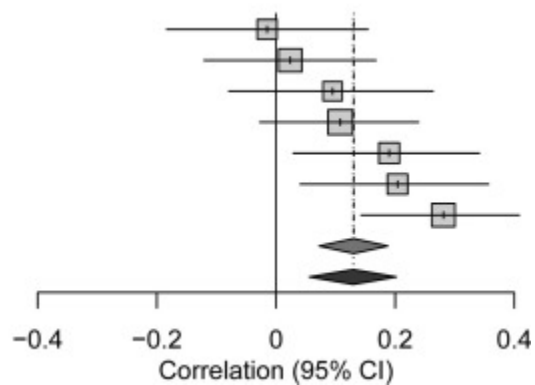


Fig. 4. Correlation between self-rated funniness and the judges' ratings.

| Source | COR (95% CI) |
|--|----------------------|
| 7.RWA | -0.28 [-0.41; -0.14] |
| 5.PACA_S2 | -0.18 [-0.34; -0.02] |
| 1.NCSA | 0.00 [-0.17; 0.17] |
| 3.CHC | 0.08 [-0.06; 0.21] |
| 2.PHYS.AN | 0.15 [0.00; 0.29] |
| 4.PACA_S1 | 0.16 [-0.01; 0.31] |
| 6.PACA_S3 | 0.65 [0.54; 0.74] |
| Total (fixed effect) | 0.08 [0.02; 0.13] |
| Total (random effects) | 0.10 [-0.14; 0.32] |
| Heterogeneity: $\chi^2_6 = 98.45$ ($P < .001$), $I^2 = 94\%$ | |

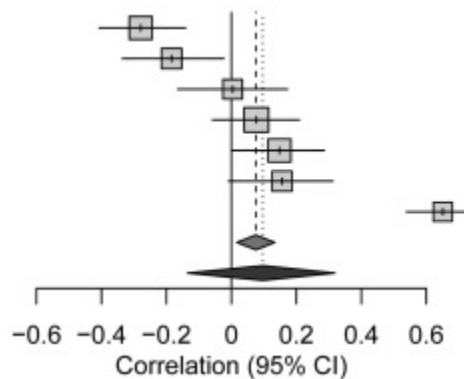


Fig. 5. Correlation between overall judge-rated funniness and self-judge covariation.

| Source | SMD (95% CI) |
|---|---------------------|
| 6.PACA_S3 | -0.22 [-0.54; 0.10] |
| 4.PACA_S1 | -0.16 [-0.44; 0.12] |
| 2.PHYS.AN | -0.14 [-0.59; 0.31] |
| 1.NCSA | -0.10 [-0.60; 0.40] |
| 3.CHC | -0.05 [-0.30; 0.19] |
| 5.PACA_S2 | 0.09 [-0.24; 0.42] |
| 7.RWA | 0.18 [-0.07; 0.43] |
| Total (fixed effect) | -0.03 [-0.15; 0.08] |
| Total (random effects) | -0.03 [-0.15; 0.08] |
| Heterogeneity: $\chi^2_6 = 5.80$ ($P = .45$), $I^2 = 0\%$ | |

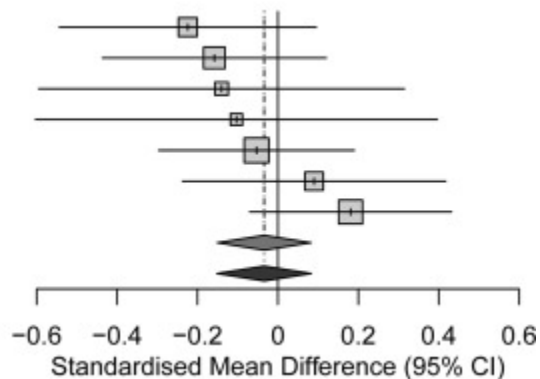


Fig. 6. Effects of gender on self-judge covariation.

Likewise, for personality traits we found at most weak and inconclusive evidence for moderation. The only personality trait worth noting was extraversion. As Fig. 7 shows, the fixed effects model found a non-zero effect ($r = -0.07$ [-0.13, -0.01], $k = 7$). People higher in extraversion were less discerning—their self-ratings were less strongly associated with the judges' ratings. But the effects were highly variable, and the effect size included zero in the random effects model ($r = -0.05$ [-0.18, 0.08], $k = 7$), so we would interpret this effect as at most food-for-thought for future research.

| Source | COR (95% CI) |
|--|----------------------|
| Trait = A | |
| 7.RWA | -0.19 [-0.32; -0.04] |
| 5.PACA_S2 | 0.00 [-0.16; 0.16] |
| 4.PACA_S1 | 0.05 [-0.12; 0.21] |
| 1.NCSA | 0.07 [-0.10; 0.24] |
| 3.CHC | 0.10 [-0.04; 0.23] |
| 2.PHYS.AN | 0.11 [-0.03; 0.25] |
| 6.PACA_S3 | 0.15 [-0.02; 0.32] |
| Total (fixed effect) | 0.04 [-0.02; 0.10] |
| Total (random effects) | 0.04 [-0.04; 0.12] |
| Heterogeneity: $\chi^2_6 = 13.18$ ($P = .04$), $I^2 = 54\%$ | |
| Trait = C | |
| 2.PHYS.AN | -0.22 [-0.35; -0.08] |
| 7.RWA | -0.19 [-0.33; -0.05] |
| 3.CHC | -0.07 [-0.21; 0.06] |
| 4.PACA_S1 | 0.02 [-0.15; 0.18] |
| 6.PACA_S3 | 0.06 [-0.11; 0.23] |
| 1.NCSA | 0.10 [-0.07; 0.26] |
| 5.PACA_S2 | 0.23 [0.07; 0.38] |
| Total (fixed effect) | -0.03 [-0.09; 0.03] |
| Total (random effects) | -0.02 [-0.13; 0.10] |
| Heterogeneity: $\chi^2_6 = 25.92$ ($P < .001$), $I^2 = 77\%$ | |
| Trait = E | |
| 3.CHC | -0.30 [-0.42; -0.17] |
| 7.RWA | -0.20 [-0.34; -0.06] |
| 2.PHYS.AN | -0.04 [-0.18; 0.11] |
| 5.PACA_S2 | -0.02 [-0.18; 0.14] |
| 6.PACA_S3 | -0.02 [-0.19; 0.15] |
| 4.PACA_S1 | -0.01 [-0.17; 0.16] |
| 1.NCSA | 0.29 [0.13; 0.44] |
| Total (fixed effect) | -0.07 [-0.13; -0.01] |
| Total (random effects) | -0.05 [-0.18; 0.08] |
| Heterogeneity: $\chi^2_6 = 34.3$ ($P < .001$), $I^2 = 83\%$ | |
| Trait = N | |
| 2.PHYS.AN | -0.11 [-0.25; 0.04] |
| 6.PACA_S3 | -0.07 [-0.24; 0.10] |
| 5.PACA_S2 | -0.06 [-0.22; 0.10] |
| 7.RWA | -0.05 [-0.19; 0.10] |
| 3.CHC | 0.02 [-0.11; 0.16] |
| 4.PACA_S1 | 0.05 [-0.12; 0.21] |
| 1.NCSA | 0.08 [-0.09; 0.25] |
| Total (fixed effect) | -0.02 [-0.08; 0.04] |
| Total (random effects) | -0.02 [-0.08; 0.04] |
| Heterogeneity: $\chi^2_6 = 4.72$ ($P = .58$), $I^2 = 0\%$ | |
| Trait = O | |
| 3.CHC | -0.19 [-0.32; -0.06] |
| 5.PACA_S2 | -0.18 [-0.33; -0.01] |
| 7.RWA | -0.07 [-0.21; 0.08] |
| 1.NCSA | -0.01 [-0.18; 0.15] |
| 4.PACA_S1 | 0.01 [-0.15; 0.18] |
| 2.PHYS.AN | 0.14 [0.00; 0.28] |
| 6.PACA_S3 | 0.24 [0.07; 0.40] |
| Total (fixed effect) | -0.02 [-0.08; 0.04] |
| Total (random effects) | -0.01 [-0.12; 0.10] |
| Heterogeneity: $\chi^2_6 = 24.29$ ($P < .001$), $I^2 = 75\%$ | |
| Trait = H | |
| 7.RWA | 0.01 [-0.14; 0.15] |
| 6.PACA_S3 | 0.05 [-0.12; 0.22] |
| Total (fixed effect) | 0.02 [-0.09; 0.13] |
| Total (random effects) | 0.02 [-0.09; 0.13] |
| Heterogeneity: $\chi^2_1 = 0.14$ ($P = .70$), $I^2 = 0\%$ | |
| Heterogeneity: $\chi^2_{36} = 109.57$ ($P < .001$), $I^2 = 67\%$ | |

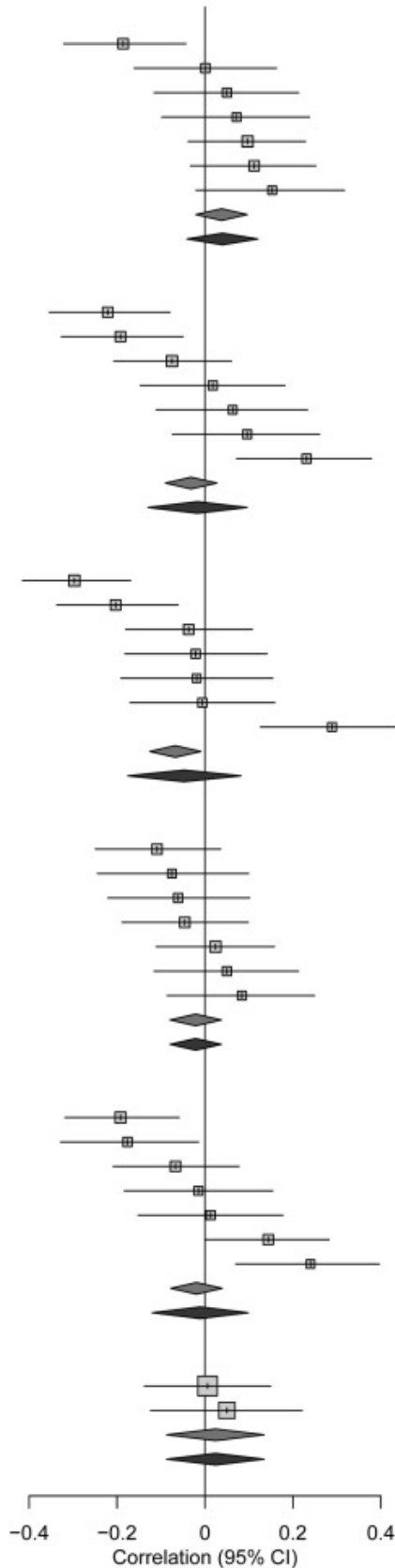


Fig. 7. Effects of personality traits on self-judge covariation.

4. Discussion

Being funny is hard. Humor production, like creative thought more generally, is lumpy and uneven, a mix of hits and misses. Even hilarious people generate a lot of duds, so effective humor requires judging one's ideas to decide which jokes are "keepers" and which ones need more work. In the present research, we explored people's ratings of the funniness of their attempts at humor. Overall, people were relatively modest and self-critical in their ratings. In contrast to the funnier-than-average effect found for global humor traits, like "having a sense of humor" or "being a funny person," people were more circumspect about the funniness of their specific responses to the humor prompts.

Although people were modest on average, the variability in self-rated funniness was associated with several factors. People higher in extraversion and openness to experience rated their ideas as funnier. Both traits are pervasive in humor research, and both are associated with viewing oneself as a funny person (Silvia et al., 2020) and using humor in everyday life (Heintz, 2017). For gender, men's self-ratings were higher than women's, consistent with a large literature on lower humor self-efficacy in women (Caldwell and Wojtach, 2020, Hooper et al., 2016, Silvia et al., 2020) and an edge for men in lab-based humor tasks (Greengross et al., 2020). For both personality and gender, the effects were generally small in magnitude.

The sample showed at least some discernment about the funniness of their ideas. The within-person correlation between a person's self-ratings and the judges' rating was small but positive. When people rated an idea as funnier than average, relative to their own average rating, it was likely to get a higher funniness rating from the judges. This finding, as a within-person effect, is not confounded by between-person third variables (e.g., humor ability or self-efficacy), so it is compelling evidence that people have at least some ability to evaluate the comedic effectiveness of their ideas. The effect size was small in magnitude, which is consistent with several possibilities. One explanation recognizes that there's a wide range of ways that people can try to be funny and enormous individual differences in what people find funny (Plessen et al., 2020). In this view, people have some insight into the funniness of their ideas, but this insight is relatively fragile—a notion consistent with most people's experience of wayward attempts at humor in everyday life. Alternatively, the small effect size for discernment could reflect unique qualities of the humor production context. Instead of creating a possibly funny idea in the context of an ongoing interaction or via non-verbal methods, participants tried to create a funny verbal response on the spot for an unknown audience. The small effect size could thus reflect the complexity of the task or the psychological distance between the participants and the eventual audience.

The participants were discerning, but variability in discernment was unrelated to the moderators measured in these studies. Many creativity studies have found that some people show more insight into the creative quality of their ideas (e.g., Benedek et al., 2016, Grohman et al., 2006, Karwowski et al., 2020, Steele et al., 2018). People high in divergent thinking, for example, are good at coming up with original ideas and at selecting which of their ideas are best (Grohman et al., 2006). Likewise, people high in openness to experience generate ideas that are much more creative and give self-ratings of their ideas that covary much more strongly with

judges' ratings (Silvia, 2008). This “double threat” effect—being better at both generating and evaluating ideas—has often appeared in creativity research but did not occur for humor in the present study. Variation in the within-person correlation between self-rated and judge-rated humor was not significantly predicted by personality traits (Fig. 7), gender (Fig. 6), or receiving higher humor scores from the judges (Fig. 5). At most, there was a suggestion that people higher in extraversion might be less discerning.

In light of the large sample size, we would conclude that there's a lack of evidence for personality and gender differences in humor discernment, at least for the personality traits assessed in these studies. This finding may speak to the distinctiveness of humor production ability from humor appreciation, which plays different roles in social interaction and interpersonal attraction (Greengross, 2014, Greengross et al., 2020). It may also reflect limits in the method. When participants are asked to output creative ideas—such as poems, metaphors, unusual uses, or jokes—they tend to hold back ideas that they don't think are worth writing down. This unseen filtering of ideas is a limitation to behavioral studies using production tasks, and it might be especially acute in humor tasks, which ask people for their single best idea. In future work, deeper insight into individual differences in humor discernment can be gained by extending these methods. The study of divergent thinking offers some useful models, such as “think aloud” paradigms that can illuminate the shifting strategies involved in creative ideation (Gilhooly, Fioratou, Anthony, & Wynn, 2007) or cognitive neuroscience methods that focus on the time course of generative and evaluative cognitive processes (Beaty et al., 2016, Beaty et al., 2017).

Regarding limits on generality, the samples we recruited were relatively narrow in age and cultural background. Although recruiting from a regional public university as well as a specialized arts university expands the sample somewhat, the participants were nevertheless all college students living in the Southeastern USA and predominantly young and female. We should also note that pooling the seven samples via meta-analysis methods gives insight into study-to-study sampling variation, but the stability of the effect sizes from the individual studies are nevertheless influenced by study-level sample sizes (Schönbrodt & Perugini, 2013), which ranged from 129 to 212 participants (see Table 1) and are thus somewhat modest for individual differences research.

In addition, the studies used a cluster of humor tasks that have performed well in past work, but some of them have been used only by our lab, so extending these findings to other ways of measuring humor (Ruch & Heintz, 2019) could be worthwhile. In particular, these tasks emphasize verbal wit expressed through writing, a key form of humor but one that might favor abilities and traits linked to verbal creativity and crystallized intelligence, such as openness to experience. Tasks that include other forms of humor might show different personality profiles. Studies with these and similar tasks, for example, consistently find that openness to experience has much larger effects on humor production than other traits (e.g., Nusbaum, 2015, Nusbaum and Silvia, 2017, Nusbaum et al., 2017, Sutu et al., 2020), but tasks that involve non-verbal skill (e.g., delivering jokes vs writing them), public performance, interpersonal charisma, or behavioral disinhibition might reveal larger effects for other traits.

Our seven studies used a cluster of self-report scales—the NEO-FFI, BFI-10, and HEXACO-100—to measure personality traits, and these scales reflect different underlying models of the traits. In some cases, the differences in trait concepts are stark, such as the contrasting views of agreeableness in the HEXACO, Big Five, and Five Factor models (Ashton et al., 2014, Miller et al., 2011). Other differences, while less dramatic, are nevertheless notable, such as different senses of the meaning of emotionality/neuroticism (Ashton et al., 2014) and the wide variation in the meaning and breadth of openness to experience, particularly the relative weight of imaginative, intellectual, and unconventional components (Christensen, Cotter, & Silvia, 2019). On the one hand, using a broad set of scales provides some breadth and generality to the findings; on the other hand, pooling different trait models can obscure the differences between them. We did not have enough studies with each scale to meaningfully compare them, especially because the only sample to use the BFI was relatively distinctive (a sample of arts students; see Table 1), and the predominantly homogeneous effect sizes suggest that such comparisons would be unlikely to be fruitful. Nevertheless, focused examinations of these traits, particularly at the facet level, would be a natural next step for the study of personality and self-ratings of humor.

In future work, it would be interesting to explore how people forecast how funny different audiences—such as close friends, family members, similar strangers, and people in general—would find their ideas. This kind of judgment requires shifting from what one finds personally funny and taking a detached, outside perspective on one’s ideas. Such a skill is obviously crucial to effectively using humor in interpersonal situations, as anyone who failed to “read the room” before uncorking a dud knows all too well. It is possible that different predictors—perhaps traits connected to social skills, emotional intelligence, and perspective taking—would be relevant to people’s expectations of audience reactions.

Open practices

Open Data: The data and R files are available at Open Science Framework (<https://osf.io/57nvg/>).

CRedit authorship contribution statement

Paul J. Silvia: Conceptualization, Formal analysis, Methodology, Writing - original draft. **Gil Greengross:** Conceptualization, Formal analysis, Writing - review & editing. **Katherine N. Cotter:** Conceptualization, Investigation, Methodology, Writing - review & editing. **Alexander P. Christensen:** Conceptualization, Investigation, Methodology, Writing - review & editing. **Jeffrey M. Gredlein:** Investigation, Writing - review & editing.

References

- Alicke, M. D., & Govorun, O. (2005). The better-than-average effect. In M. D. Alicke, D. A. Dunning, & J. I. Krueger (Eds.), *The self in social judgment* (pp. 85–106). Psychology Press.
- Allison, S. T., Messick, D. M., & Goethals, G. R. (1989). On being better but not smarter than others: The Muhammad Ali effect. *Social Cognition*, 7(3), 275–295.

- Ashton, M. C., & Lee, K. (2019). How well do Big Five measures capture HEXACO scale variance? *Journal of Personality Assessment*, 101, 567–573.
- Ashton, M. C., Lee, K., & De Vries, R. E. (2014). The HEXACO Honesty-Humility, Agreeableness, and Emotionality factors: A review of research and theory. *Personality and Social Psychology Review*, 18(2), 139–152.
- Beaty, R. E., Benedek, M., Silvia, P. J., & Schacter, D. L. (2016). Creative cognition and brain network dynamics. *Trends in Cognitive Sciences*, 20, 87–95.
- Beaty, R. E., Christensen, A. P., Benedek, M., Silvia, P. J., & Schacter, D. L. (2017). Creative constraints: Brain activity and network dynamics underlying semantic interference during idea production. *Neuroimage*, 148, 189–196.
- Beins, B. C., & O’Toole, S. M. (2010). Searching for the sense of humor: Stereotypes of ourselves and others. *Europe’s Journal of Psychology*, 6(3), 267–287.
- Benedek, M., Nordtvedt, N., Jauk, E., Koschmieder, C., Pretsch, J., Krammer, G., & Neubauer, A. C. (2016). Assessment of creativity evaluation skills: A psychometric investigation in prospective teachers. *Thinking Skills and Creativity*, 21, 75–84.
- Berg, J. M. (2019). When silver is gold: Forecasting the potential creativity of initial ideas. *Organizational Behavior and Human Decision Processes*, 154, 96–117.
- Caldwell, T. L., & Wojtach, P. (2020). Men are funnier than women under a condition of low self-efficacy but women are funnier than men under a condition of high self-efficacy. *Sex Roles*, 83, 338–352.
- Chapman, A. J., & Gadfield, N. J. (1976). Is sexual humor sexist?. *Journal of Communication*, 26(3), 141–153.
- Christensen, A. P., Cotter, K. N., & Silvia, P. J. (2019). Reopening openness to experience: A network analysis of four openness to experience inventories. *Journal of Personality Assessment*, 101(6), 574–588.
- Christensen, A. P., Silvia, P. J., Nusbaum, E. C., & Beaty, R. E. (2018). Clever people: Intelligence and humor production ability. *Psychology of Aesthetics, Creativity, and the Arts*, 12, 136–143.
- Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.
- Cropley, A. (2006). In praise of convergent thinking. *Creativity Research Journal*, 18, 391–404.
- Dean, D. L., Hender, J., Rodgers, T., & Santanen, E. (2006). Identifying good ideas: Constructs and scales for idea evaluation. *Journal of Association for Information Systems*, 7(10), 646–699.
- Diedrich, J., Jauk, E., Silvia, P. J., Gredlein, J. M., Neubauer, A. C., & Benedek, M. (2018). Assessment of real-life creativity: The Inventory of Creative Activities and Achievements (ICAA). *Psychology of Aesthetics, Creativity, and the Arts*, 12, 304–316.

- Dunning, D., Meyerowitz, J. A., & Holzberg, A. D. (1989). Ambiguity and self-evaluation: The role of idiosyncratic trait definitions in self-serving assessments of ability. *Journal of Personality and Social Psychology*, 57(6), 1082–1090.
- Fine, G. A. (1975). Components of perceived sense of humor ratings of self and other. *Psychological Reports*, 36(3), 793–794.
- Gagnon, J. (2013, August 5). Reasons women aren't funny. *McSweeney's Internet Tendency*.
- Gilhooly, K. J., Fioratou, E. E., Anthony, S. H., & Wynn, V. V. (2007). Divergent thinking: Strategies and executive involvement in generating novel uses for familiar objects. *British Journal of Psychology*, 98, 611–625.
- Greengross, G. (2014). Male production of humor produced by sexually selected psychological adaptations. In V. A. Weekes-Shackelford & T. K. Shackelford (Eds.), *Evolutionary perspectives on human sexual psychology and behavior* (pp. 173–196). Springer.
- Greengross, G. (2020). Sex and gender differences in humor. Introduction and overview. *Humor*.
- Greengross, G., Silvia, P. J., & Nusbaum, E. C. (2020). Sex differences in humor production ability: A meta-analysis. *Journal of Research in Personality*, 84, 103886.
- Grohman, M., Wodniecka, Z., & Kłusak, M. (2006). Divergent thinking and evaluation skills: Do they always go together? *Journal of Creative Behavior*, 40, 125–145.
- Heintz, S. (2017). Putting a spotlight on daily humor behaviors: Dimensionality and relationships with personality, subjective well-being, and humor styles. *Personality and Individual Differences*, 104, 407–412.
- Hitchens, C. (2007). Why women aren't funny. *Vanity Fair*, 557, 54–59.
- Hooper, J., Sharpe, D., & Roberts, S. G. B. (2016). Are men funnier than women, or do we just think they are? *Translational Issues in Psychological Science*, 2, 54–62.
- Johnson, J. A. (1994). Clarification of factor five with the help of the AB5C model. *European Journal of Personality*, 8(4), 311–334.
- Karwowski, M., Czerwonka, M., & Kaufman, J. C. (2020). Does intelligence strengthen creative metacognition? *Psychology of Aesthetics, Creativity, and the Arts*, 14, 353–360.
- Kaufman, J. C., & Beghetto, R. A. (2013). In praise of Clark Kent: Creative metacognition and the importance of teaching kids when (not) to be creative. *Roeper Review*, 35, 155–165.
- Klein, S. B., & Loftus, J. (1993). The mental representation of trait and autobiographical knowledge about the self. In T. K. Srull & R. S. Wyer (Eds.), *Advances in social cognition* (Vol. 5, pp. 1–49). Erlbaum.
- Klein, S. B., Loftus, J., Traflet, J. G., & Fuhrman, R. W. (1992). Use of exemplars and abstractions in trait judgments: A model of trait knowledge about the self and others. *Journal of Personality and Social Psychology*, 63, 739–753.

- Kleinmintz, O. M., Ivancovsky, T., & Shamay-Tsoory, S. G. (2019). The two-fold model of creativity: The neural underpinnings of the generation and evaluation of creative ideas. *Current Opinion in Behavioral Sciences*, 27, 131–138.
- Kozbelt, A. (2007). A quantitative analysis of Beethoven as self-critic: Implications for psychological theories of musical creativity. *Psychology of Music*, 35, 144–168.
- Kruger, J. (1999). Lake Woebegon be gone! The “below-average effect” and the egocentric nature of comparative ability judgments. *Journal of Personality and Social Psychology*, 77, 221–232.
- Lee, K., & Ashton, M. C. (2018). Psychometric properties of the HEXACO-100. *Assessment*, 25, 543–556.
- Lefcourt, H. M., & Martin, R. A. (1986). *Humor and life stress: Antidote to adversity*. Springer.
- Martin, R. A. (2014). Humor and gender: An overview of psychological research. In D. Chiaro & R. Baccolini (Eds.), *Gender and humor: Interdisciplinary and international perspectives*. Mouton de Gruyter.
- McNeish, D., Stapleton, L. M., & Silverman, R. D. (2017). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods*, 22, 114–140.
- Mickes, L., Walker, D. E., Parris, J. L., Mankoff, R., & Christenfeld, N. J. (2012). Who’s funny: Gender stereotypes, humor production, and memory bias. *Psychonomic Bulletin & Review*, 19, 108–112.
- Miller, J. D., Gaughan, E. T., Maples, J., & Price, J. (2011). A comparison of agreeableness scores from the Big Five Inventory and the NEO PI-R: Consequences for the study of narcissism and psychopathy. *Assessment*, 18(3), 335–339.
- Nusbaum, E. C. (2015). A meta-analysis of individual differences in humor production and personality (Doctoral dissertation. University of North Carolina at Greensboro).
- Nusbaum, E. C., & Silvia, P. J. (2017). What are funny people like? Exploring the crossroads of humor ability and openness to experience. In G. J. Feist, R. Reiter-Palmon, & J. C. Kaufman (Eds.), *Cambridge handbook of creativity and personality research* (pp. 294–322). Cambridge University Press.
- Nusbaum, E. C., Silvia, P. J., & Beaty, R. E. (2014). Ready, set, create: What instructing people to “be creative” reveals about the meaning and mechanisms of divergent thinking. *Psychology of Aesthetics, Creativity, and the Arts*, 8, 423–432.
- Nusbaum, E. C., Silvia, P. J., & Beaty, R. E. (2017). Ha ha? Assessing individual differences in humor production ability. *Psychology of Aesthetics, Arts, and Creativity*, 11, 231–241.
- Nusbaum, E. C., Silvia, P. J., Beaty, R. E., Burgin, C. J., & Kwapil, T. R. (2015). Turn that racket down! Physical anhedonia and diminished pleasure from music. *Empirical Studies of the Arts*, 33, 228–243.
- Omwake, L. (1937). A study of sense of humor: Its relation to sex, age, and personal characteristics. *Journal of Applied Psychology*, 21, 688–704.

- Plessen, C. Y., Franken, F. R., Ster, C., Schmid, R. R., Wolfmayr, C., Mayer, A. M., ... Maierwieser, R. J. (2020). Humor styles and personality: A systematic review and meta-analysis on the relations between humor styles and the Big Five personality traits. *Personality and Individual Differences*, 154 109676.
- Primi, R., Silvia, P. J., Jauk, E., & Benedek, M. (2019). Applying many-facet Rasch modeling in the assessment of creativity. *Psychology of Aesthetics, Creativity, and the Arts*, 13, 176–186.
- Puente-Díaz, R., Cavazos-Arroyo, J., & Vargas-Barrera, F. (2021). Metacognitive feelings as a source of information in the evaluation and selection of creative ideas. *Thinking Skills and Creativity*, 39 100767.
- R Core Team (2020). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Available from <https://www.Rproject.org>.
- Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, 41, 203–212.
- Ruch, W., & Heintz, S. (2019). Humor production and creativity: Overview and recommendations. In S. R. Luria, J. Baer, & J. C. Kaufman (Eds.), *Creativity and humor* (pp. 1–42). Academic Press.
- Said-Metwaly, S., Fernández-Castilla, B., Kyndt, E., & Van den Noortgate, W. (2020). Testing conditions and creative performance: Meta-analyses of the impact of time limits and instructions. *Psychology of Aesthetics, Creativity, and the Arts*, 14, 15–38.
- Sawyer, R. K. (2012). *Explaining creativity: The science of human innovation* (2nd ed.). Oxford University Press.
- Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, 47(5), 609–612.
- Schwarzer, G. (2020). meta: General package for meta-analysis. R package version 4.14.0. Downloaded from <https://CRAN.R-project.org/package=meta>.
- Shin, H., Cotter, K. N., Christensen, A. P., & Silvia, P. J. (2020). Creative fixation is no laughing matter: The effects of funny and unfunny examples on humor production. *Journal of Creative Behavior*, 54, 487–494.
- Silvia, P. J. (2008). Discernment and creativity: How well can people identify their most creative ideas?. *Psychology of Aesthetics, Creativity, and the Arts*, 2, 139–146.
- Silvia, P. J., Christensen, A. P., & Cotter, K. N. (2021). Right-wing authoritarians aren't very funny: RWA, personality, and creative humor production. *Personality and Individual Differences*, 170 110421.
- Silvia, P. J., Rodriguez, R. M., & Karwowski, M. (2020, August 23). *Funny selves: Development of the Humor Efficacy and Identity Short Scales (HEISS)*. Unpublished preprint. <https://doi.org/10.31234/osf.io/p3mza>.

- Simonton, D. K. (1999). Creativity as blind variation and selective retention: Is the creative process Darwinian?. *Psychological Inquiry*, 309–328.
- Steele, L. M., Johnson, G., & Medeiros, K. E. (2018). Looking beyond the generation of creative ideas: Confidence in evaluating ideas predicts creative outcomes. *Personality and Individual Differences*, 125, 21–29.
- Stemler, S. E., & Kaufman, J. C. (2020). Are creative people better than others at recognizing creative work?. *Thinking Skills and Creativity*, 38 100727.
- Sternberg, R. J. (2006). The nature of creativity. *Creativity Research Journal*, 18, 87–98.
- Sutu, A., Phetmisy, C. N., & Damian, R. I. (2020). Open to laugh: The role of openness to experience in humor production ability. *Psychology of Aesthetics, Creativity, and the Arts*.
- Weisberg, R. W. (2020). *Rethinking creativity: Inside-the-box thinking as the basis for innovation*. Cambridge University Press.
- Wu, J. Y., & Kwok, O. M. (2012). Using SEM to analyze complex survey data: A comparison between design-based single-level and model-based multilevel approaches. *Structural Equation Modeling*, 16–35.
- Zell, E., Strickhouser, J. E., Sedikides, C., & Alicke, M. D. (2020). The better-than-average effect in comparative self-evaluation: A comprehensive review and meta-analysis. *Psychological Bulletin*, 146(2), 118–149.