# If You're Funny and You Know It:
# Personality, Gender, and People's Ratings of Their Attempts at Humor

Paul J. Silvia

University of North Carolina at Greensboro

Gil Greengross

Aberystwyth University

Katherine N. Cotter, Alexander P. Christensen

University of Pennsylvania

Jeffrey M. Gredlein

University of North Carolina School of the Arts

Preprint: October 8, 2020

## Author Note

Paul J. Silvia, Department of Psychology, University of North Carolina at Greensboro, USA; Gil Greengross, Department of Psychology, Aberystwyth University, Wales; Katherine N. Cotter, Positive Psychology Center, University of Pennsylvania, USA; Alexander P. Christensen, Penn Center for Neuroaesthetics, Department of Neurology, University of Pennsylvania, USA; Jeffrey M. Gredlein, Division of Liberal Arts, University of North Carolina School of the Arts, USA.

The data and research materials are available at Open Science Framework (https://osf.io/57nvg/). The authors report no conflicts of interest. This project was not preregistered.

Please address correspondence to Paul J. Silvia, Department of Psychology, University of North Carolina at Greensboro, Greensboro, NC, 20402-6170, p_silvia@uncg.edu.

**Paul J. Silvia:** Conceptualization, Formal Analysis, Methodology, Writing-Original draft preparation. **Gil Greengross:** Conceptualization, Formal Analysis, Writing-Reviewing and Editing. **Katherine N. Cotter:** Conceptualization, Investigation, Methodology, Writing-Reviewing and Editing. **Alexander P. Christensen:** Conceptualization, Investigation, Methodology, Writing-Reviewing and Editing. **Jeffrey M. Gredlein:** Investigation, Writing-Reviewing and Editing.

**Abstract**

In seven studies ($n$ = 1,133), adults tried to create funny ideas and then rated the funniness of their responses, which were also independently rated by judges. People were relatively modest and self-critical about their ideas. Extraversion ($r$ = .12 [.07, .18], $k$ =7) and openness to experience ($r$ = .09 [.03, .15], $k$ = 7) predicted rating one's responses as funnier; women rated their responses as less funny ($d$ = -.28 [-.37, -.19], $k$ = 7). The within-person correlation between self and judge ratings was small but significant ($r$ = .13 [.07, .19], $k$ = 7), so people had some insight into their ideas' funniness.

Keywords: humor; comedy; creativity; personality; gender; discernment

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 are overly positive and self-serving.

In her 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 (Fine, 1975; Lefcourt & Martin, 1986) illustrate the classic "better-than-average effect" (Alicke & Govorun, 2005) in self-perceptions of humor.

Research on personality and self-ratings of global humor 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 et al., 2020), and to say they are good at amusing others and making them laugh (i.e., 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., 2020), 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 & Wojtach, 2020; Silvia et al., 2020) and view being a funny person as more peripheral to their self-concept (Silvia et al., 2020). A meta-analysis of lab-based humor production studies found a small advantage for men (Greengross et al., 2020), so these self-beliefs have parallels in behavioral contexts.

But what about judgments of specific attempts to be funny? People are inclined to see themselves as "funny people," but how do they view their concrete attempts at humor? In all creative domains, people's attempts to come up with good ideas are a mix of hits and misses (Simonton, 1999; Weisberg, 2020). Although one can get more hits by cranking out attempts, 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 (Karwowski et al., 2020; Kozbelt, 2007; Silvia, 2009). The dialectic between generation and evaluation—coming up with ideas versus judging, refining, and discarding them—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).

Who finds 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 unpacking the issue (Berg, 2019; Dean et al., 2006; Grohman et al., 2006; Kozbelt, 2007; Silvia, 2009). First, people show within-person variability in their ratings of their ideas—they rarely think all their ideas are great. 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, 2009) and *creative metacognition*

(Karwowski et al., 2020).[1]

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? To test these questions, we pooled data from seven similar studies on humor production (total $n$ = 1,133), all of which measured personality, gender, and humor performance. Instead of reporting all seven studies individually, we used meta-analysis methods to provide a distilled, compact analysis of who finds their own ideas funny.

## Method

### 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 et al., 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.

The participants across the seven samples consisted of 1,133 adults who were native English speakers; 76% identified as female. Except for the first sample, which consisted of

[1] Creative discernment is usually described in terms of covariation between self-ratings and judges' ratings, not in terms of the "accuracy" of self-ratings. Sociocultural models of creativity point out that the creative effectiveness of an idea is heavily influenced by the audience and a constellation of social and cultural processes (Sawyer, 2012; Sternberg, 2006), so there is not an objective criterion for "accuracy" when it comes to creativity and humor (see Silvia, 2009, for a discussion).

college students with a variety of arts majors enrolled at University of North Carolina School of the Arts (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, 2017; Silvia et al., in press). 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.

**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 and explained in detail elsewhere (Nusbaum et al., 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. In the *joke stems task*, people read a scenario that set up a joke (e.g., eating something terrible that a friend cooked and then describing what it was like) and then completed it 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. In most studies, the raters used a 5-point (1-5) scale; the most recent sample used a 3-point (0-2) scale because Rasch analyses found that judges rarely separate more than 3 categories (Primi et al., 2019; Silvia et al., in press).

**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.

## 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 conducted. In many cases, the software constrained $I^2$ to zero, which can happen when an analysis has a small and homogenous set of effects. The

results from the fixed-effects models are reported in the text. Findings from the random-effects models are presented in the figures and called out in the text when they vary notably. The data and R files are available on Open Science Framework (https://osf.io/57nvg/).

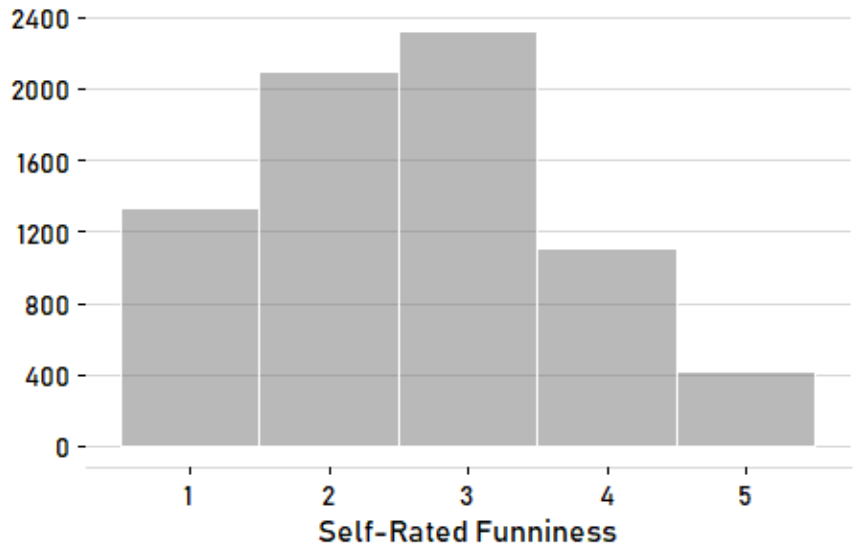
Calculating the effect sizes is relatively more complicated for these samples because (1) people gave between 2 and 9 self-ratings, one for each humor item, and (2) the judges' ratings of humor introduce a facet into the data (Primi et al., 2019). For the effects of gender and personality on self-ratings, the effect sizes were estimated using cluster-robust standard-error models (McNeish et al., 2017; Wu & 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.

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 .05, and thinning to every 10$^{th}$ step.

**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. Figure 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). Although not too much should be made of absolute points on a observed rating scale (Bond et al., 2020), 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.

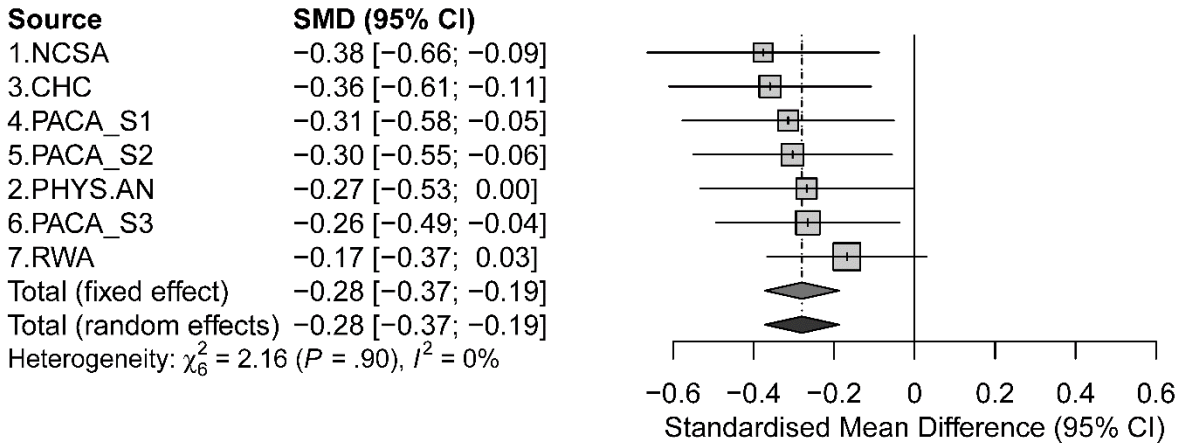
**Figure 1.** Distribution of self-rated funniness across all responses.




**Who Rated Their Ideas as Funny?**

What predicted between-person variation in self-rated humor? Gender was an important factor (see Figure 2). Compared to men, women rated their responses as less funny, $d$ = -.28 95% CI [-.37, -.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).

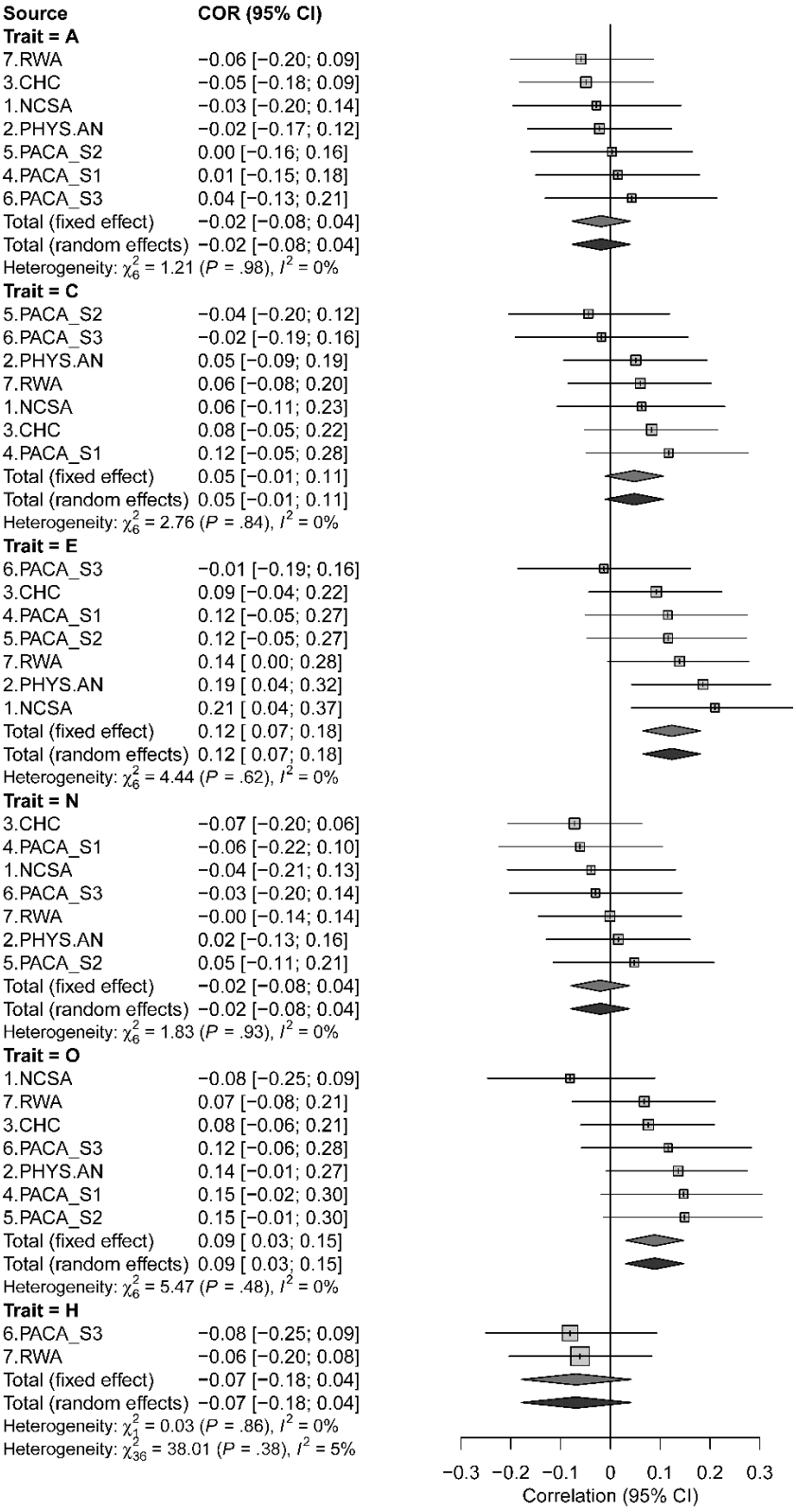
**Figure 2.** Gender and the self-rated funniness of one's ideas.




**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 Figure 3). The effects for agreeableness, conscientiousness, honesty-humility, and neuroticism were non-significant. People rated their ideas as being funnier when they were more extraverted ($r$ = .12 [.07, .18], $k$ =7) and more open to experience ($r$ = .09 [.03, .15], $k$ = 7); the effect sizes were small in magnitude.

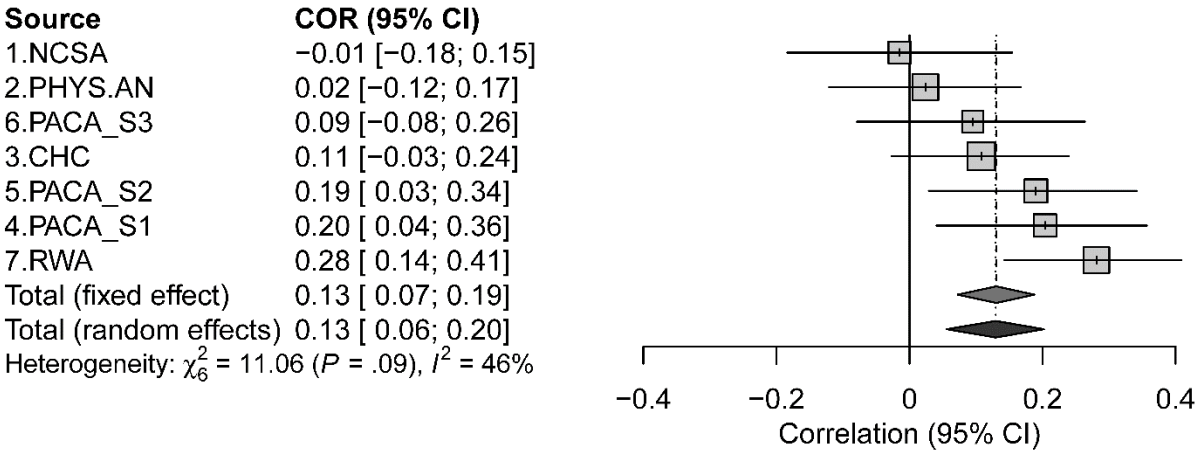
**Figure 3.** Personality traits and the self-rated funniness of one's ideas.



**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, $r$ = .13 [.07, .19], $k$ = 7 (see Figure 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.

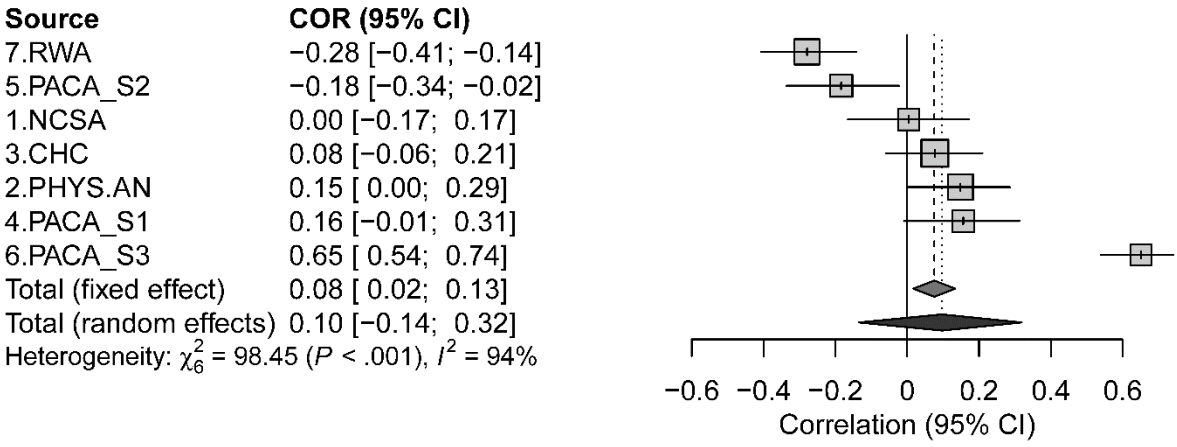
**Figure 4.** Correlation between self-rated funniness and the judges' ratings.




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 Figure 5 shows, although the confidence intervals for the fixed-effect model excluded zero ($r$ = .08 [.02, .13], $k$ = 7), the effect

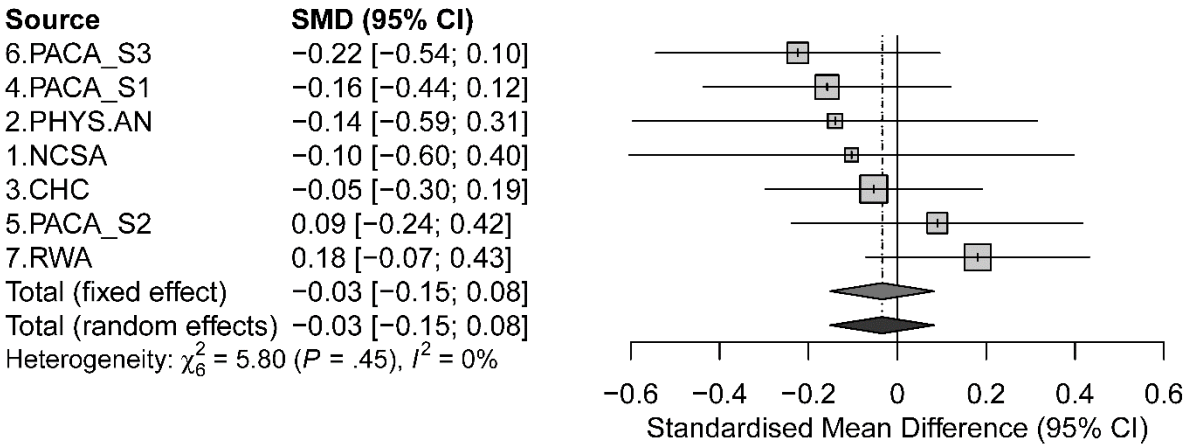
sizes were highly variable, as reflected in the random-effects model ($r$ = .10 [-.13, .32], $k$ = 7). Given the homogeneity of the participants, tasks, and judges across the seven samples, we conclude that there's a lack of evidence for funnier people being more discerning.

**Figure 5.** Correlation between overall judge-rated funniness and self–judge covariation.



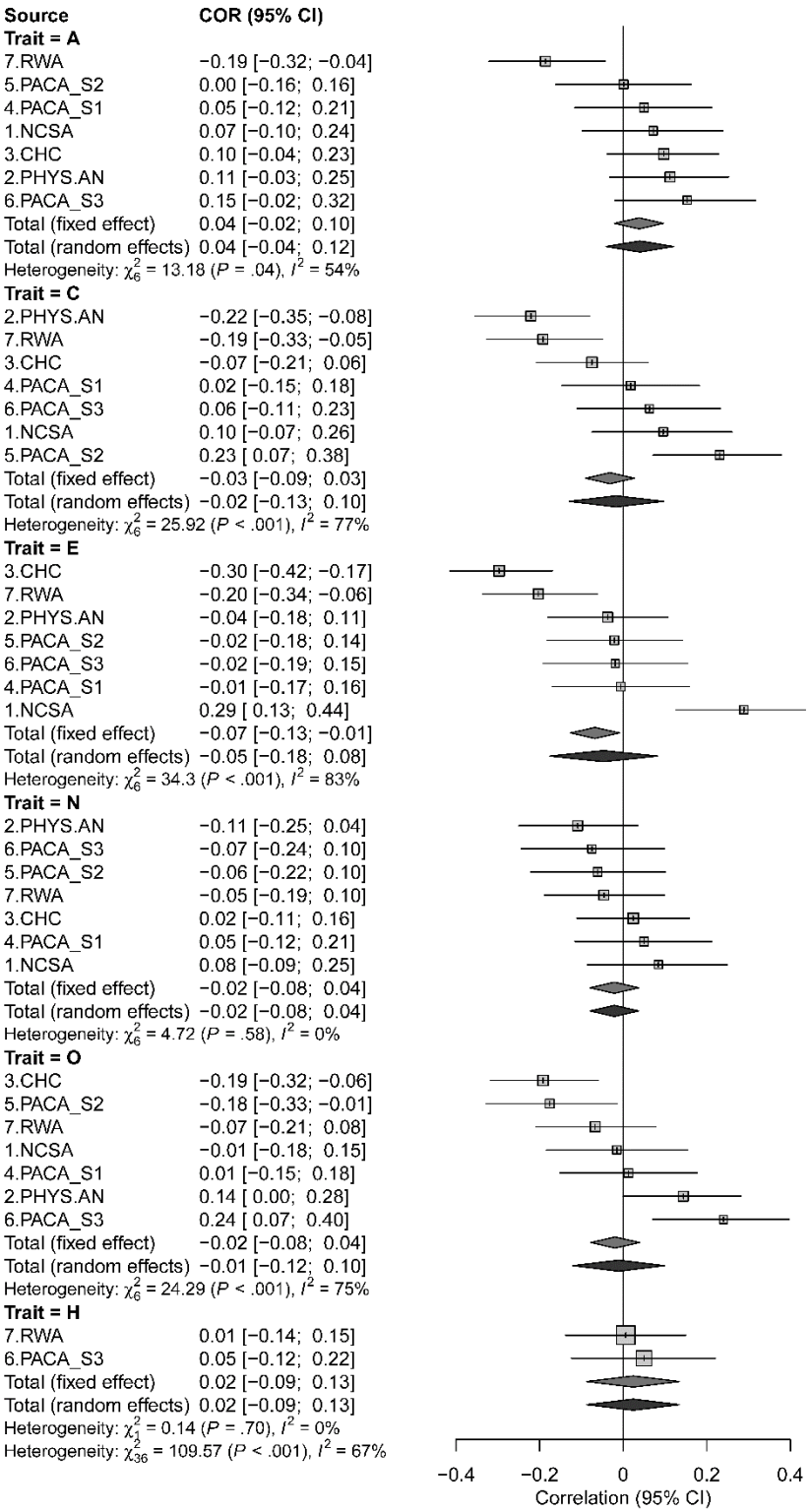

Finally, did gender or personality moderate the correlation between self and judge ratings of humor? No evidence was found for moderation by gender ($d$ = -.03 [-.15, .08], $k$ = 7), as Figure 6 shows. Although women rated their ideas as less funny, their ratings were not more or less tightly correlated with the judges' ratings. Women and men were thus equally discerning about the funniness of their ideas.

**Figure 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 Figure 7 shows, the fixed effects model found a non-zero effect ($r$ = -.07 [-.13, -.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 ($r$ = -.05 [-.18, .08], $k$ = 7), so we would interpret this effect as at most food-for-thought for future research.

**Figure 7.** Effects of personality traits on self–judge covariation.



## 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 & 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 the wide range of ways that people can try to be funny and the enormous individual differences in what people find funny (Plessen et al., 2020).

The participants were discerning, but variability in discernment was unsystematic. Many creativity studies have found that some people show more insight into the creative quality of their ideas (e.g., Benedek et al., 2016; 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, 2009). 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 (Figure 7), gender (Figure 6), or receiving higher humor scores from the judges (Figure 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.

Regarding limits on generality, the sample is 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. 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 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/).

## 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.

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.

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.

Bond, T. G., Yan, Z., & Heine, M. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences* (4th ed.). Routledge.

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.

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.

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.

Gagnon, J. (2013, August 5). Reasons women aren't funny. *McSweeney's Internet Tendency*. https://www.mcsweeneys.net/articles/reasons-women-arent-funny.

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. (2020). Sex and gender differences in humor: Introduction and overview. *Humor*.

Greengross, G., & Miller, G. F. (2009). The Big Five personality traits of professional comedians compared to amateur comedians, comedy writers, and college students. *Personality and Individual Differences*, *47*, 79-83.

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.

Karwowski, M., Czerwonka, M., & Kaufman, J. C. (2020). Does intelligence strengthen creative metacognition? *Psychology of Aesthetics, Creativity, and the Arts, 14*, 353–360.

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.

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.

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.

R Core Team (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Available from https://www.R-project.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.

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., 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.

Sternberg, R. J. (2006). The nature of creativity. *Creativity Research Journal, 18,* 87–98.

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.

Table 1

**Summary of the Samples**

| Sample | *n* | % Female | Age | Humor Tasks | Raters & Scale | Personality Measure | Notes |
|---|---|---|---|---|---|---|---|
| 1. NCSA (Diedrich et al., 2018) | 132 | 42 | 20.10 (18, 49) | 2 (Jokes) | 3 raters (1-5) | BFI-10 | Unusually high O scores, consistent with a sample of arts students. |
| 2. Physical Anhedonia (Nusbaum et al., 2015) | 185 | 70 | 19.65 (18, 32) | 3 (Jokes) | 2 raters (1-5) | NEO-FFI | |
| 3. CHC Humor (Christensen et al., 2018) | 212 | 85 | 19.11 (18, 48) | 9 (Cartoons, Jokes, Definitions) | 3 raters (1-5) | NEO FFI | |
| 4. Ha Ha, 1 (Nusbaum et al., 2017) | 142 | 81 | 18.79 (18, 40) | 6 (Cartoons, Jokes) | 4 raters (1-5) | NEO FFI | The Resumes task reported in the article did not have self-ratings. |
| 5. Ha Ha, 2 | 147 | 72 | 19.07 (18, 32) | 6 (Cartoons, | 5 raters (1-5) | NEO FFI | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| (Nusbaum et al., 2017) | | | | Jokes) | | | |
| 6. Ha Ha, 3 (Nusbaum et al., 2017) | 129 | 69 | 18.49 (18, 24) | 9 (Cartoons, Jokes, Definitions) | 2 raters (1-5) | HEXACO-100 | |
| 7. RWA Humor (Silvia et al., in press) | 186 | 78 | 19.07 (18, 53) | 9 (Cartoons, Jokes, Definitions) | 3 raters (0-2) | HEXACO-100 | Only the 3 raters who scored all 9 tasks were included. |

**Notes.** For gender, the percent 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., in press). 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/).