

Time is a Funny Thing: Response Times and Humor Quality in a Creative Joke Production Task

Paul J. Silvia¹, Katherine N. Cotter², and Alexander P. Christensen³

¹ Department of Psychology, University of North Carolina at Greensboro, USA

² Positive Psychology Center, University of Pennsylvania, USA

³ Department of Psychology and Human Development, Vanderbilt University, USA

ABSTRACT

Generating creative ideas takes time: the first idea to come to mind is usually obvious, and people need time to shift strategies, enact executive processes, and evaluate and revise an idea. The present research explored the role of time in creative humor production tasks, which give people a prompt and ask them to create a funny response. A sample of 152 young adults completed four joke stems prompts. Their response times were recorded, and the responses were judged for humor quality (funniness) by six independent judges and by the participants themselves. Mixed-effect models found that, at the within-person level, response time's link to humor quality diverged for judges and participants. The judges' ratings of funniness predicted longer response times (relatively funnier responses took longer to create), but participants' self-ratings of their own responses predicted shorter response times (relatively funnier responses were created faster). Controlling for elaboration (quantified via word count of the response) diminished the effect of judge-rated humor but not participant-rated humor. Taken together, the results suggest that the role of time in humor generation is complex: judges may be weighting elabo-

ration more heavily when judging funniness, whereas participants may be weighting metacognitive cues like ease-of-generation when judging their own ideas.

KEYWORDS:

humor production, creativity, response time, openness to experience, intelligence

Article history:

Received: October 10, 2023

Received in revised form: December 12, 2023

Accepted: December 22, 2023

ISSN 2354-0036

DOI: 10.2478/ctra-2024-0001

Corresponding author at:

Paul J. Silvia

E-MAIL: p_silvia@uncg.edu

A time-honored way of measuring creative thought is to ask people to generate open-ended responses to a prompt—such as creating captions for images, unusual uses for paperclips, or drawings that incorporate a shape or squiggle—and then score their responses (Reiter-Palmon et al., 2019). Researchers who start using these methods quickly notice that many of the ideas in a huge pool of responses are pretty funny. It isn't surprising, then, that researchers interested in creative humor production (Ruch & Heintz, 2019)—the creation of new ideas that are intended to be funny (versus recalling old jokes from memory)—use a similar open-ended assessment strategy. People are presented with a context for humor and asked to generate a funny response, such as writing a funny caption for a photograph (Kozbelt & Nishioka, 2010) or single-panel cartoon (Koppel & Sechrest, 1970), completing a resume with funny entries based on a job-seeker's photo (Howrigan & MacDonald, 2008), creating a funny definition for a quirky pair of words like "yoga bank" (Nusbaum et al., 2017), or finishing a "set up" prompt with a funny ending (Shin et al., 2020). These responses are then rated by judges (and sometimes by participants themselves) for funniness (Nusbaum et al., 2017; Silvia et al., 2021).

When we view humor production as a subset of creativity more broadly, it isn't surprising to see that predictors of performance on humor tasks resemble the predictors of performance on creativity tasks like divergent thinking and metaphor production. For example, humor production is strongly linked to individual differences in openness to experience (Nusbaum & Silvia, 2017; Sutu et al., 2021), several factors of intelligence (Christensen et al., 2018; Greengross & Miller, 2011; Kellner & Benedek, 2017; Masten, 1986), and brain regions involved in creative thought (Amir & Biederman, 2016). At the same time, much more is known about creativity than about humor production. In particular, the psychology of divergent thinking is a rich, mature area that brings together many methods and schools of thought to understand the cognitive processes involved when people try to come up with creative ideas.

To get a foothold into the cognitive processes involved in generating funny ideas, we can borrow from established findings in the broader literature on creativity. Current models of creative thought—as captured with divergent thinking tasks—contend that it isn't simply a matter of spontaneous, bottom-up processes (Beaty & Kenett, 2023; Benedek et al., 2023; Gerver et al., 2023). Bottom-up processes are undoubtably central to creative thought, but on their own they tend to lead to obvious, close ideas. These low-level processes need to be guided, managed, or inhibited in different contexts via executive processes.

A key finding for our purposes is that creative ideas take time: in general, "first is worst" when it comes to clever ideas. When asked to generate unusual uses for an object, for example, people tend to start with obvious, close uses. These ideas are usually sparked by semantic similarity or by ease of retrieval from memory (Benedek et al., 2023; Gilhooly et al., 2007). With time, most people move past these obvious responses and begin generating responses that are novel, interesting, and eccentric. The temporal aspect of creative thought has been firmly established. At the study level, a recent meta-analysis of the divergent thinking literature found that giving participants longer time limits is associated with higher levels of originality (Said-Metwaly et al., 2020). At the

within-person level, the “serial order effect” illustrates that most people’s responses to a prompt grow more creative as time passes (Beaty & Silvia, 2012).

When people are spending more time on an open-ended creativity task, what are they doing that is making their responses more creative? There’s no single process at work, but rather a set of things that people might be doing that fosters creativity. Strategy use, for example, unfolds over time. Most people start a divergent thinking task with a simple “what do I remember?” strategy, in which they search memory for uses that they have observed (Benedek et al., 2023), and many people will eventually shift to more complex strategies that open up new possibilities (Gilhooly et al., 2007). Another key meta-cognitive process is idea evaluation. During the task, people may judge their ideas, discard the weaker ones, and then refine and elaborate the better ones (Finke et al., 1996). An emphasis on idea evaluation often appears as a “quality over quantity” strategy, in which people generate fewer responses across time but all the ideas tend to be good, even the earlier ones (Gonthier & Besançon, 2022).

Less is known about the cognitive processes at work when people try to come up with funny ideas (Sun et al., 2022), but there are some clear parallels with the broader creativity literature regarding the importance of executive and metacognitive processes. Focusing on the role of time, a few studies suggest that spending more time on a humor task is related to generating funnier ideas, using independent judges’ humor ratings as the criterion. Like divergent thinking tasks, humor production tasks show a serial order effect. Derks and Hervas (1988) showed 38 adult participants images from classic movies and asked them to generate picture captions that were “as funny as possible” (p. 37) for each one. The participants were asked to generate either 2 funny captions or 10 funny captions per image. The funniness of the captions increased steeply at first but then flattened over time, a profile resembling the serial-order effect in divergent thinking. In a study of creative fixation, 175 people were given a humor production task along with examples of funny responses, unfunny responses, or no examples (Shin et al., 2020). When people saw the bad examples of unfunny responses that they should avoid generating, they took longer to generate a response and created responses that were much funnier.

A study of humor production and time (Kozbelt & Nishioka, 2010) is a good jumping-off point for the present research. They recruited a small sample of 18 people who were selected via snowball sampling for being funny and extraverted. Among other things, the participants were shown 20 color photos and asked to create a single funny caption for each one. There was no time limit, and response times were recorded. A recommended time window, however, was suggested: people were asked to “try to take up to a minute” (p. 388) to create their funny caption. The results suggested a positive relationship between time spent generating a response and its funniness. At the picture level (ignoring the nesting of photos within participants), there was a small correlation ($r = .14$) between how long it took to create the response and how funny a group of judges found it. These findings are intriguing and suggestive for a role of time in creative humor production, but the small sample size and recommended time window for responding complicate the interpretation of a link between time and funniness.

THE PRESENT RESEARCH

In the present research, we examine the relationships between time and the production of humorous responses. Like past work, we were interested in the link between how long people spend attempting to create a funny response and the eventual humor quality of their responses. This project sought to go beyond past work in several key respects. First, the study used a multilevel design and evaluated the *within-person* relationships between response times and humor quality. The participants were asked to generate a single response to four different prompts, and each response was rated for funniness by independent judges. Because people will vary in the funniness of their responses and how long they took to create each one, the within-person variability in response times can be explored. Within-person relationships provide a person-level interpretation—when a person's responses are funnier than their own average level of funniness, did the response take them longer to generate?—and they yield effects that are unconfounded by between-person individual differences (Nezlek, 2012), such as traits like intelligence or openness to experience.

Second, we wanted to go beyond judges' ratings of humor and explore how time relates to both judges' ratings and to participants' own ratings of their responses. Creativity researchers have a long-standing interest in how well people can judge the creative quality and potential of their ideas. In lab tasks, people typically show some creative discernment, quantified as the relative covariation of ratings between participants and judges (Silvia, 2008). Humor researchers, too, have a long interest in people's beliefs about their humor abilities (Omwake, 1937), usually couched in the context of "funnier-than-average" effects (Beins & O'Toole, 2010) in which people tend to overstate their sense of humor and humor ability. Having both judges and participants rate the funniness of each response allows us to see if the main finding from past research—funnier responses take longer to generate—applies to both independent judges and to the participants to created the responses. We did not have any specific predictions about how self-ratings of humor would relate to response time, but recent work has found a small correlation between how funny participants and judges find a response (Silvia et al., 2021), so it wouldn't be surprising for self-rated and judge-rated humor to diverge.

Third, we wanted to explore some possible reasons why time and humor are related. A natural place to start is the elaborateness of the response. In the broader creativity literature, an idea's elaborateness influences how human raters and automated systems evaluate creativity (Dumas et al., 2021; Forthmann et al., 2019; Maio et al., 2020). Not surprisingly, relatively elaborate responses take more time to generate (Gonthier & Besançon, 2022), for both mundane reasons (longer responses take more time to write or type) and for interesting reasons. In particular, elaborate responses are often the result of a longer process of response evaluation and revision, where a solid idea's rough form is reworked and refined into an improved final response.

Unlike divergent thinking research, which has a long interest in text features that relate to rated creativity (e.g., elaboration and semantic distance), humor production research knows little about how text features relate to funniness ratings in structured humor tasks. In the present study,

we measured the length of each humor response (the number of words) as a rough proxy for its elaborateness. Using this measure of elaboration, we can explore links between response length and time (more elaborate responses should take longer to generate) as well as between response length and humor ratings by judges and participants. We had no specific predictions, but exploring elaboration seemed worthwhile in light of the lack of knowledge about how response features connect to rated humor quality.

METHOD

Participants

The sample consisted of 152 young adults (122 women, 26 men, and 4 who indicated another gender identity or declined to state any; age $M = 19.07$ years, $SD = 2.27$, range from 18 to 30) at the University of North Carolina at Greensboro, USA. All participants provided written informed consent and received credit toward a voluntary research option in their psychology courses. The research was approved by the UNCG Institutional Review Board (Protocol #14-0002, approval date August 23, 2019). The research was not preregistered, and the sample size was determined by practical guidelines (running until the end of an academic semester) and not an a priori power analysis.

Procedure

The data were collected in-person at an on-campus laboratory prior to the Covid-19 pandemic. The participants took part in sessions of up to eight people at individual computer stations in the lab. After completing the informed consent forms, they learned about the purpose of the study and were given instructions for completing the tasks and scales. All participants first completed the joke production tasks, then the intelligence tasks, and finally the self-report measures of personality.

Humor Tasks

To measure creative humor production, we used the joke stems task (Nusbaum et al., 2017). Designed to be intuitive and reasonably realistic for participants, this task uses a “set up, knock down” comedic structure. Participants were given four prompts that contained an emotional scenario that sets up a joke (see Table 1). The participants were then asked to complete the prompt by writing something funny; only one response is given per prompt, and there is no time limit. Each scenario involved situations and emotions that are common contexts for everyday humor.

Table 1. The Joke Stems Task

Prompt Name	Scenario
Gross Food	Imagine that your friend invites you over and cooks dinner—and the food is totally disgusting. Later, when describing it to another friend, you say, “Wow, that food was so bad. . .”
Boring Class	Imagine that one of your classes this semester is unbelievably boring, and you’re trying to describe just how boring it is to one of your best friends. So you say, “Seriously, that class is so boring. . .”
Bad Singing	Imagine that one of your friends wants your opinion on how well she sings. She sings a minute or two to demonstrate her voice, and you cringe—she might be the worst singer you’ve ever heard. When she asks, “So, how was it?” you decide to be honest, so you say, “To be honest, listening to that. . .”
Dirty Car	Imagine that a new friend is picking you up so you can go shopping. You haven’t seen their car before, so when it drives up you’re shocked—it might be the dirtiest, messiest car you’ve ever seen. When your friend smiles and says, “Sorry, it’s a little dirty,” you laugh and say, “A little? Seriously, this car. . .”

Just as creativity research explicitly instructs participants to “be creative” (Harrington, 1975; Nusbaum et al., 2014; Said-Metwaly et al., 2020), we explicitly instructed participants to “be funny.” The task was described as a humor task with the goal of creating funny ideas. The participants were encouraged to feel free to express their own sense of humor. Because college students’ humor can lean toward perverse, profane, and quirky styles, the instructions gave permission for non-normative responses by telling them that it was fine for their responses to be “weird, silly, dirty, ironic, bizarre, or whatever,” so long as the responses were funny.

The four different joke stem prompts were shown in a randomized order for each participant. Participants read the prompt followed by a reminder that the goal was to be funny. Unlike unusual uses tasks, which usually ask for many responses within a fixed time window, this joke task asked people to generate a single response per prompt with no time constraint. They could spend as much time as they wished to generate a response, and no guidance or target time was implied (cf. Kozbelt & Nishioka, 2010). A text box on the same screen allowed participants to type in their response. The text box allowed edits and revisions, so participants could rework their response until they decided to click a button that submitted it. The software recorded the response time (in ms) for each prompt, which was measured from the time that the prompt appeared to when the participant clicked submit.

Participant and Rater Judgments of Humor

After submitting their response to a prompt, participants were asked to rate the humor quality of their response. The next screen showed the response that the participant just submitted and asked, “In your opinion, how funny is your joke?” The participants judged their jokes with a 5-point scale (1 = *not at all funny*, 5 = *very funny*).

After data collection was complete, the responses were anonymized and rated for humor quality by a set of judges. Six people (three women, three men) rated the responses independently of each other, and they were unaware of a participant’s scores on other variables as well as

a participant's responses to the other prompts. The raters judged each response on a three-point scale ranging from 0 (*not funny*) to 2 (*very funny*). The specific instructions to the raters are provided in the online supplementary material. Because raters can be relatively lenient or severe when judging humor, we used many-facet Rasch models (Eckes, 2011; Primi et al., 2019; Wang & Long, 2022) to estimate latent "rated humor quality" scores that are adjusted for judges' severity. The models were estimated using the R package TAM (Robitzsch et al., 2022). The expected a posteriori (EAP) reliability of overall individual differences in rated humor quality (adjusting for the relative difficulty of the four prompts and the relative severity of the six raters) was .78. EAP reliabilities for each individual prompt were .59 (bad singing), .63 (gross food), .64 (dirty car), and .69 (boring class).

Measures of Individual Differences

Cognitive abilities were measured with a group of tasks used in past research on humor (Christensen et al., 2018) and creative thinking (Frith et al., 2021). Fluid intelligence (Gf) was assessed with three tasks: a letter sets task (4 minutes, 15 items) presents sets of four letters that follow a rule and asks people to pick the set that breaks the rule (Ekstrom et al., 1976); a series completion task (3 minutes, 13 items) asks people to pick a shape that completes an evolving series of shapes (Cattell & Cattell, 1961); and a paper folding task (3 minutes, 10 items) asks people to imagine what a square piece of paper would look like after being folded, punched with holes, and then unfolded (Ekstrom et al., 1976). Crystallized intelligence (Gc) was measured with advanced (18 items) and extended-range (24 items) vocabulary tasks (Ekstrom et al., 1976).

Personality traits were measured with the NEO-FFI 3 (McCrae & Costa, 2007), a 60-item scale that measures the five main factors of personality (12 items per factor). Participants responded using a 5-point scale (1 = *strongly disagree*, 5 = *strongly agree*). Internal consistency was good for neuroticism ($\alpha = .78$), extraversion ($\alpha = .82$), openness to experience ($\alpha = .74$), agreeableness ($\alpha = .76$), and conscientiousness ($\alpha = .85$).

RESULTS

Data Reduction and Analysis Approach

The data and R input files are available at Open Science Framework (<https://osf.io/k9tdr/>). Response times were expressed in seconds. Response length was expressed as the total number of words after cleaning the text for spelling, punctuation, contractions, and abbreviations. One response (a single entry for the "gross food" prompt) was treated as missing due to a probable software error. The three Gf tasks were standardized and then averaged to create an overall Gf score; the total number of correct vocabulary items served as the Gc score. The 12 items for each of the NEO personality trait scales were averaged to create overall scores. The data were analyzed in R 4.3 (R Core Team, 2023). Because each participant generated four responses (one for each humor prompt), the

design has a multilevel structure. The multilevel models were explored and estimated using the packages *misty* (Yanagida, 2023) and *lme4* (Bates et al., 2015).

Table 2. Descriptive Statistics

Variable	Mean	Median	SD	Min	Max	<i>n</i>
RT: Car	55.69	45.45	32.69	14.27	218.68	152
RT: Class	60.43	50.11	47.13	12.09	461.12	152
RT: Food	48.37	41.89	27.74	15.65	221.25	151
RT: Singing	61.84	46.64	48.35	12.20	361.52	152
Words: Car	7.70	7.00	4.31	1.00	25.00	152
Words: Class	9.24	8.00	4.43	4.00	35.00	152
Words: Food	8.37	7.00	3.60	3.00	25.00	151
Words: Singing	8.53	7.50	5.10	2.00	39.00	152
Judge Humor: Car	0.00	-1.20	1.54	-1.28	3.96	152
Judge Humor: Class	0.01	0.02	1.75	-1.54	3.80	152
Judge Humor: Food	0.01	-1.28	1.75	-1.28	3.72	151
Judge Humor: Singing	0.00	0.02	1.14	-1.00	3.38	152
Self Humor: Car	2.82	3.00	1.07	1.00	5.00	152
Self Humor: Class	2.81	3.00	1.22	1.00	5.00	152
Self Humor: Food	2.82	3.00	1.02	1.00	5.00	151
Self Humor: Singing	2.76	3.00	1.16	1.00	5.00	152
Neuroticism	3.28	3.29	0.60	1.67	4.75	152
Extraversion	3.37	3.33	0.58	2.00	4.92	152
Openness to Experience	3.57	3.54	0.52	2.42	4.83	152
Agreeableness	3.73	3.83	0.52	2.33	4.83	152
Conscientiousness	3.54	3.58	0.55	2.00	4.67	152
Fluid Intelligence (Gf)	0.00	0.09	1.00	-3.34	2.51	152
Crystallized Intelligence (Gc)	0.00	-0.17	1.00	-2.12	3.33	151

Note. Car, Class, Food, and Singing refer to scores for the Dirty Car, Boring Class, Gross Food, and Bad Singing joke stem prompts shown in Table 1. RT = response time; Words = word count for the response; Judge Humor = IRT-based funniness based on the judges' ratings; Self Humor = participants' own ratings of the funniness of their response.

Exploring Response Times and Word Counts

Table 2 shows descriptive statistics for each joke stem and for individual differences in personality and intelligence. As expected, there was broad variability in how long people took to generate responses to the joke stems (in seconds) and the elaborateness of their responses (in words). Figure 1 displays the skewed and lumpy distributions of response time and word count for each joke stem prompt. All four prompts were similar in their central tendencies and shapes, so they appeared to behave similarly.

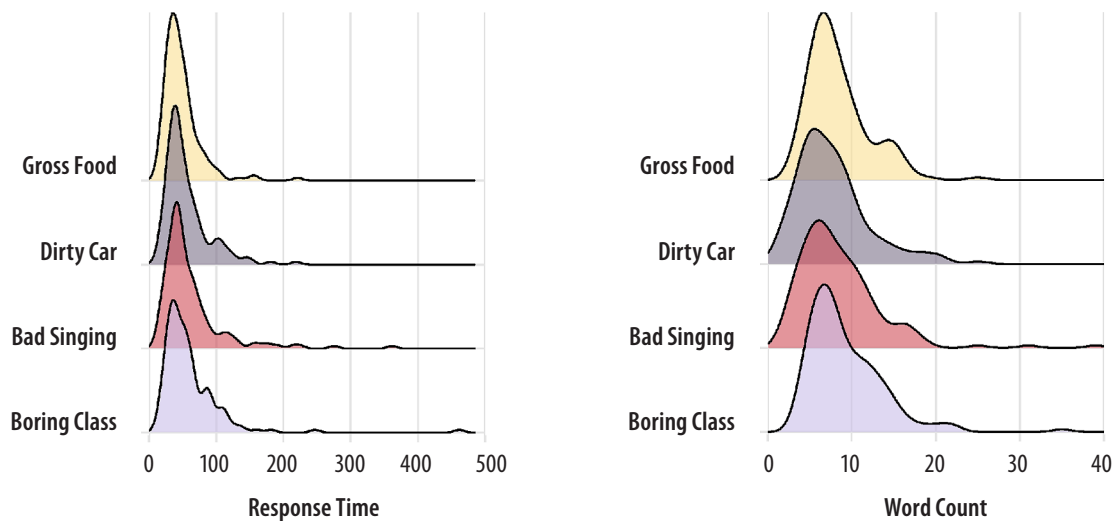


Figure 1. Distributions of Response Times and Word Counts for Each Item in the Joke Stem Task

Note. Response times are in seconds.

Funniness ratings by both judges and participants, shown in Figure 2, showed similar variability. The judges' scores (expressed as IRT-based theta scores) were notably skewed: overall, the judges didn't find many responses to be very funny. The participants' scores (given on the original 5-point scale) were relatively normal, with ratings at both the floor and ceiling of the scale. It's important to note that many participants did rate some of their joke responses as relatively unfunny. Although people overall do tend to see themselves as "funnier than average," they can be quite critical and discerning when judging the funniness of their specific attempts at humor (Silvia et al., 2021).

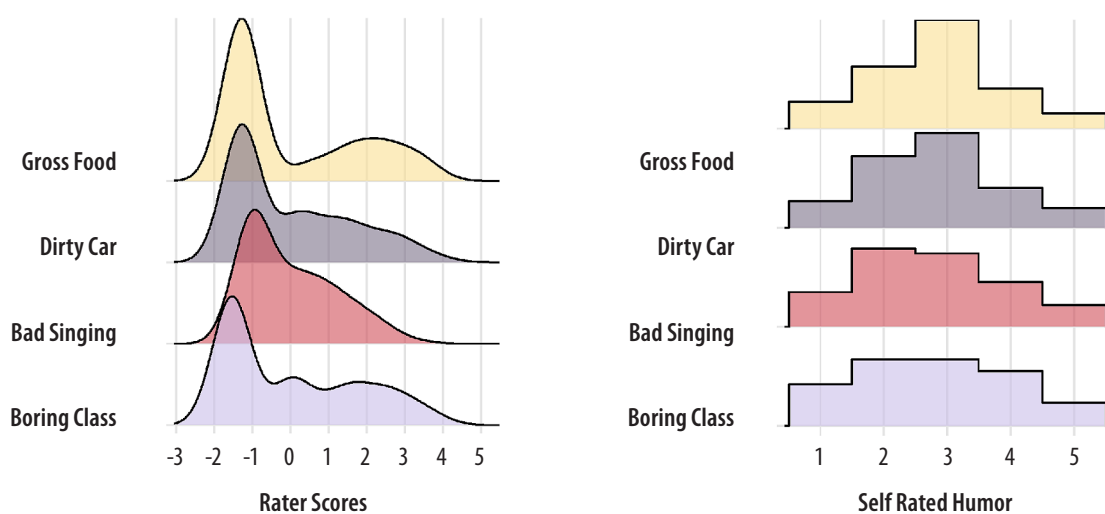


Figure 2. Distributions of Joke Funniness Ratings by Judges and Participants

Note. Judges' ratings are IRT-based scores that correct for judge severity; participants' ratings are on a 1–5 scale.

Because each participant completed four joke stems (one response per joke stem), we can explore variability at both the between-person and within-person levels. Table 3 displays multilevel correlations at both the between-person (above diagonal) and within-person (below diagonal) levels. Intraclass correlations showed marked variability at the within-person level for response time ($ICC = .43$), word count ($ICC = .38$), judge-rated humor ($ICC = .27$), and self-rated humor ($ICC = .33$). The within-person variability across the four tasks is noteworthy, and the simple correlations suggest that multilevel regression models of within-person relationships are likely to be fruitful.

Table 3. Multilevel Correlations: Within-person (Below) and Between-Person (Above)

Variable	1. RT	2. Words	3. Judge-rated Humor	4. Self-rated Humor
1. RT	1	.79	.55	.13
2. Words	.34	1	.64	.17
3. Judge-rated Humor	.11	.23	1	.20
4. Self-rated Humor	-.22	.13	.15	1

Note. Within-person correlations (r) are below the diagonal; between-person correlations are above the diagonal.

Multilevel Predictors of Response Time

What variables predicted taking more or less time to come up with a funny idea? The first set of multilevel models explored response times for the joke stems task. The within-person predictors (judge ratings and self-ratings) were within-person centered, so they represent deviations around each participant's own mean. The between-person predictors were standardized and hence grand-mean centered. The regression coefficients are unstandardized. Because of the small within-person sample size (only four joke responses per person), the slopes were estimated as fixed effects to facilitate convergence.

When creative production tasks have no time limit or word constraints, variables like response time and word count can cause problems for mixed-model assumptions because of distant, outlying values and skewed distributions. Model residuals and diagnostics were evaluated, and robustness checks were conducted by estimating key models using alternate frameworks, particularly robust mixed models using *robustlmm* (Koller, 2016) and quantile mixed models at the median using *lqmm* (Geraci, 2014). The model results were not appreciably different for robust and median-quantile models (the patterns of slopes and their significance levels didn't change), so the conventional mixed models are reported here throughout. For the curious, the OSF archive presents R code for the robust and quantile mixed models (<https://osf.io/k9tdr/>).

Within-person predictors. Our first model explored the within-person predictors of response time. The results showed that judges' ratings and participants' ratings of a joke's funniness both significantly predicted response time, but in opposing directions. Response times were longer for the jokes that judges rated as relatively funnier ($b = 3.21$ [1.17, 5.26], $SE = 1.04$, $p = .002$), but they were

shorter for the jokes that the participants themselves rated as relatively funnier ($b = -8.00$ $[-11.01, -5.00]$, $SE = 1.53$, $p < .001$).¹ Stated differently, within the set of a typical person's responses, jokes that the judges found relatively funnier took more time to generate, but the jokes that the participants found relatively funnier took less time to generate.

Intelligence. Fluid and crystallized intelligence were then added to the multilevel model as between-person predictors. Fluid intelligence didn't predict response times to the joke task ($b = -.16$ $[-4.90, 4.59]$, $SE = 2.42$, $p = .948$), but crystallized intelligence did ($b = 7.87$ $[3.11, 12.63]$, $SE = 2.42$, $p < .001$). People higher in crystallized intelligence took significantly longer to generate their responses to the joke stems.

Personality. Finally, we explored how personality predicted response time by replacing the intelligence variables with the five NEO factors. Significant effects appeared for openness to experience ($b = 6.81$ $[2.07, 11.55]$, $SE = 2.41$, $p = .005$), agreeableness ($b = 6.38$ $[1.50, 11.25]$, $SE = 2.48$, $p = .010$), and conscientiousness ($b = -6.95$ $[-11.98, -1.92]$, $SE = 2.56$, $p = .007$), but not for neuroticism ($b = -2.55$ $[-7.62, 2.51]$, $SE = 2.58$, $p = .323$) or extraversion ($b = -1.22$ $[-6.17, 3.72]$, $SE = 2.52$, $p = .627$). Overall, people took more time to generate their jokes when they were higher in openness to experience, higher in agreeableness, and lower in conscientiousness.

Why Did Judges' and Participants' Ratings Have Opposing Relationships?

The within-person analyses found that longer responses were rated as relatively funnier by the judges but as relatively less funny by the participants. What might explain this divergence? As noted earlier, the elaborateness of verbal responses is known to influence creativity ratings, so it's possible that judges and participants vary in their use of elaboration as an informative cue when rating a response's funniness.

We explored the role of elaboration by including the length of the joke response, measured as the number of words, as a within-person centered predictor. The multilevel model included word count, judge ratings, and self ratings as predictors, and it found significant within-level effects for the number of words ($b = 3.18$ $[2.43, 3.93]$, $SE = .38$, $p < .001$) and for self-rated funniness ($b = -9.16$ $[-11.97, -6.34]$, $SE = 1.43$, $p < .001$). Longer response times during the joke task were thus associated with creating a relatively longer joke and, as before, with rating one's own joke as relatively less funny. The effect of judge-rated funniness, however, was positive but no longer significant when word count was included ($b = 1.42$ $[-0.54, 3.37]$, $SE = 1.00$, $p = .155$). Whereas the effect of self-rated funniness was independent of word count, the effect of judge-rated funniness was not. To explore this further, an additional model included the interactions between the three predictors; no interactions were significant. Overall, this pattern is consistent with the notion that judges' impressions of a response's funniness were relatively more informed by the response's elaboration.

¹ Models that included a within-person interaction term found that judge-rated and self-rated humor didn't interactively predict response times or word counts or change the patterns of main effects.

DISCUSSION

Do funnier ideas take more time to create? The broader literature on creative thought shows that better ideas often take longer for many reasons, from the tendency for the first idea to be obvious to the application of slower, controlled processes that aid in selecting and refining a promising idea (Beatty & Kenett, 2023; Finke et al., 1996; Gilhooly et al., 2007). The present research examined the role of time in creative humor production and extended past work in several ways. The limited work in this area suggests that humor creation tasks resemble divergent thinking tasks: humor shows a serial order effect, with later ideas being funnier (Derks & Hervás, 1988), and judges' ratings of humor quality correlate positively with how long it took to create the idea, albeit weakly and in a small sample (Kozbelt & Nishioka, 2010).

In the present project, people completed four prompts for a joke stems task, and their response time was recorded. Each response was then rated for funniness by independent judges and by the participants themselves. Having several responses per person afforded within-person analyses of links between time and humor quality. The results both extend and complicate the literature on relationships between funniness and time (Derks & Hervás, 1988; Kozbelt & Nishioka, 2010). As in past work, judges' ratings of humor quality significantly predicted response times. Because this is a within-person effect, it means that for a typical person, their relatively funnier responses took relatively longer to generate than their relatively unfunny responses. In a broad way, this finding replicates past work and is consistent with the literature on how developing creative ideas usually takes time. The participants' own ratings of their ideas' funniness, however, diverged from the judges' ratings. When participants rated an idea as relatively funnier, that idea was generated significantly faster.

At the between-person level, response times were largely predicted by individual differences in traits that predict humor quality as rated by judges. In particular, people who were higher in openness to experience and crystallized intelligence, two well-known predictors of humor ability (Kellner & Benedek, 2017; Nusbaum & Silvia, 2017; Sutur et al., 2021), took more time on average to generate their responses.

Understanding Judge Ratings and Self Ratings

Why would the judges' ratings predict slower response times and the participants' ratings predict faster response times? Clearly, these two groups are approaching their ratings from different perspectives. The judges saw all the responses from all participants before assigning their scores; the participants, of course, saw only their own four responses. Unlike the judges, who had access only to the finished products, the participants had to generate responses and thus had access to rich inner experiences during the task.

Two main reasons for the divergence come to mind. One possibility is a diverging role of elaboration in humor judgments. People can use a wide array of informational cues when deciding if they find something funny, and one factor is the relative complexity of the response. Divergent

thinking research has found that responses rated as creative tend to be more elaborate (Dumas et al., 2021; Maio et al., 2020), and that longer response times are associated with more elaborate responses (Gonthier & Besançon, 2022). It appears that the link between judges' ratings of humor and response time are accounted for in part by the fact that more elaborate humor responses (indicated by word count) took longer to generate and were seen as funnier. Because judges seem to appreciate the relatively longer responses, the predictive effect of judges' ratings on response time diminishes when word count is included in the model.

Another explanation, grounded in creative metacognition (Lebuda & Benedek, 2023; Puente-Díaz, 2023), is that participants are using cognitive fluency and the ease of response generation as cues for judging the quality of their ideas. One reason why the "first is worst" strategy is so entrenched is because the first thing that comes to mind often comes to mind so easily. The subjective experience of fluent, effortless idea generation can influence people's confidence in their creative abilities (Puente-Díaz & Cavazos-Arroyo, 2022b) and bias people's impressions of the idea's merit (Puente-Díaz & Cavazos-Arroyo, 2022a). Recent work, for example, found that people who experienced their ideas as easy to generate were more likely to overestimate the creativity of their ideas (Puente-Díaz et al., 2021).

In our study, it's possible that the effect of cognitive fluency on self-rated quality explains our findings. Given the literature on ease-of-generation effects, a similar process might be at work with humor generation: people are more likely to see their idea as funnier if it came to mind relatively quickly. Such responses might be subjectively experienced as "snappy" and apt for the prompt, leading people to stop the generation process and enter the response. Our study lacks the data needed to examine a biasing effect of fluency, but this strikes us as an especially promising direction for future research by adapting the paradigms for studying ease-of-generation in divergent thinking (e.g., Puente-Díaz et al., 2021) for humor production tasks.

Future Directions and Limitations

Broadly speaking, the role of time in creative thought is studied using two paradigms. One approach, and the one used here, asks for a single best response and measures how long it takes people to generate it. Another approach asks for many responses and tracks how the ideas and their qualities change over time; this is the method used in the classic "serial order effect" in divergent thinking (Beaty & Silvia, 2012). Both approaches offer different views of the creative process. When people must create only one good response, variation in time can catch different strategies, like people who stop and settle for the first thing that comes to mind, and can offer insight into the revision and reworking of ideas. When people create many ideas, changes in the ideas across time can illuminate how people dynamically switch strategies and explore knowledge regions. To date, there's some evidence for a serial order effect in humor production (Derks & Hervas, 1988), much like with divergent thinking, and this approach is a fertile direction for future research. By asking participants to generate many responses to the same prompt, for example, one could explore not only how the quality of the ideas changes over time but also if the ideas become more or less elab-

orate; if the ideas cluster in clumps, reflecting similar semantic domains or humor strategies; and if some people are adopting a slow-but-steady strategy that favors quality over quantity (Gonthier & Besançon, 2022). It would be particularly interesting to see if participants' judgments of their ideas show a serial order effect that parallels the judges' ratings.

The present work has some limitations that should be kept in mind and that can motivate future work. For one, the data collection methods afforded only a global measure of response time—from when the prompt started to when people clicked to submit their response—so it is hard to unpack what is happening within that span of time. As an example of other methods that could lend some insight, Kozbelt and Nishioka (2010) used software to measure the time participants spent pre-typing (from the onset of the prompt to when they clicked any key) and time spent typing. This method allowed them to distinguish between time spent reading, thinking, and planning from the time spent generating and reworking the response. In future work, software for logging keystrokes could shed light on the complex process of generation, deletion, and revision that goes into coming up with a single good idea.

In addition, the humor task that we used probably has distinctive features relative to other humor tasks that are relevant to our results. Humor assessment is not as well developed as other areas of creativity assessment, but there are several open-ended tasks available. The most common task presents an image and asks for funny captions or descriptions (Greengross & Miller, 2011; Howrigan & MacDonald, 2008); another task asks people to generate definitions for noun-noun pairs (Nusbaum et al., 2014). These tasks each have a unique structure that likely draw upon different kinds of knowledge and strategies. For example, based on our intuitions as researchers who use these tasks, we suspect that some tasks are more likely to provoke a quick, obvious response that immediately jumps to mind. For many cartoon caption prompts, for example, there's usually no obvious, dominant response that is implied by the cartoon; generating a response usually feels slow and challenging because people must develop a model of what is happening in the cartoon and then use it as a frame for humor. For the joke stems prompts, in contrast, the "set up, knock down" structures and salient emotions (e.g., boredom, disgust) probably evoke dominant responses more quickly. These tasks thus likely have different sources of interference that need to be managed, different strategies for generating something funny, and thus different metacognitive processes, particular processes connected to the experience of ease-of-generation. These notions are mostly speculative, but it seems likely that different humor tasks pose different creative challenges and could thus show different relationships with time and metacognitive processes.

More generally, the present work highlights the need to understand how people judge the funniness of a response to an open-ended humor task. Divergent thinking research has a rich history of studying the psychology of rater behavior (Long & Wang, 2022), such as the cognitive processes that affect ratings (e.g., Forthmann et al., 2017) and relevant individual differences (Kaufman et al., 2013). The present study, along with past research (Silvia et al., 2021), shows that the ratings of judges and participants covary weakly and can diverge, so it is worth studying how both judges and participants approach the quirky task of evaluating the funniness of an idea that was gener-

ated on the spot in response to a structured prompt (vs other contexts, such as improvisational humor during unstructured interactions). It seems likely that personality traits and abilities of the judges are important, but at this point we could only speculate about rater features that shape their funniness judgments.

References

- Amir, O., & Biederman, I. (2016). The neural correlates of humor creativity. *Frontiers in Human Neuroscience*, 10. <https://www.frontiersin.org/articles/10.3389/fnhum.2016.00597>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Beaty, R.E., & Kenett, Y.N. (2023). Associative thinking at the core of creativity. *Trends in Cognitive Sciences*, 27(7), 671–683. <https://doi.org/10.1016/j.tics.2023.04.004>
- Beaty, R.E., & Silvia, P.J. (2012). Why do ideas get more creative across time? An executive interpretation of the serial order effect in divergent thinking tasks. *Psychology of Aesthetics, Creativity, and the Arts*, 6, 309–319.
- 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), Article 3. <https://doi.org/10.5964/ejop.v6i3.216>
- Benedek, M., Beaty, R.E., Schacter, D.L., & Kenett, Y.N. (2023). The role of memory in creative ideation. *Nature Reviews Psychology*, 2(4), 246–257.
- Cattell, R.B., & Cattell, A.K.S. (1961). *Measuring intelligence with the Culture Fair Tests*. Hogrefe.
- 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(2), 136–143. <https://doi.org/10.1037/aca0000109>
- Derks, P., & Hervas, D. (1988). Creativity in humor production: Quantity and quality in divergent thinking. *Bulletin of the Psychonomic Society*, 26(1), 37–39. <https://doi.org/10.3758/BF03334854>
- Dumas, D., Organisciak, P., Maio, S., & Doherty, M. (2021). Four text-mining methods for measuring elaboration. *The Journal of Creative Behavior*, 55(2), 517–531. <https://doi.org/10.1002/jocb.471>
- Eckes, T. (2011). *Introduction to many-facet Rasch measurement*. Peter Lang D. <https://doi.org/10.3726/978-3-653-04844-5>
- Ekstrom, R.B., French, J.W., Harman, H.H., & Dermen, D. (1976). *Manual for Kit of Factor-Referenced Cognitive Tests*. Educational Testing Service.
- Finke, R.A., Ward, T.B., & Smith, S.M. (1996). *Creative cognition: Theory, research, and applications*. MIT Press.
- Forthmann, B., Holling, H., Zandi, N., Gerwig, A., Çelik, P., Storme, M., & Lubart, T. (2017). Missing creativity: The effect of cognitive workload on rater (dis-) agreement in subjective divergent-thinking scores. *Thinking Skills and Creativity*, 23, 129–139.
- Forthmann, B., Oyebade, O., Ojo, A., Günther, F., & Holling, H. (2019). Application of latent semantic analysis to divergent thinking is biased by elaboration. *The Journal of Creative Behavior*, 53(4), 559–575. <https://doi.org/10.1002/jocb.240>
- Frith, E., Elbich, D.B., Christensen, A.P., Rosenberg, M.D., Chen, Q., Kane, M.J., Silvia, P.J., Seli, P., & Beaty, R.E. (2021). Intelligence and creativity share a common cognitive and neural basis. *Journal of Experimental Psychology: General*, 150(4), 609–632. <https://doi.org/10.1037/xge0000958>
- Geraci, M. (2014). Linear quantile mixed models: The lqmm Package for Laplace quantile regression. *Journal of Statistical Software*, 57, 1–29. <https://doi.org/10.18637/jss.v057.i13>
- Gerver, C.R., Griffin, J.W., Dennis, N.A., & Beaty, R.E. (2023). Memory and creativity: A meta-analytic examination of the relationship between memory systems and creative cognition. *Psychonomic Bulletin & Review*. <https://doi.org/10.3758/s13423-023-02303-4>

- Gilhooly, K.J., Fioratou, E., Anthony, S.H., & Wynn, V. (2007). Divergent thinking: Strategies and executive involvement in generating novel uses for familiar objects. *British Journal of Psychology*, 98(4), 611–625. <https://doi.org/10.1111/j.2044-8295.2007.tb00467.x>
- Gonthier, C., & Besançon, M. (2022). It is not always better to have more ideas: Serial order and the trade-off between fluency and elaboration in divergent thinking tasks. *Psychology of Aesthetics, Creativity, and the Arts*. <https://doi.org/10.1037/aca0000485>
- Greengross, G., & Miller, G. (2011). Humor ability reveals intelligence, predicts mating success, and is higher in males. *Intelligence*, 39(4), 188–192. <https://doi.org/10.1016/j.intell.2011.03.006>
- Harrington, D.M. (1975). Effects of explicit instructions to “be creative” on the psychological meaning of divergent thinking test scores¹. *Journal of Personality*, 43(3), 434–454. <https://doi.org/10.1111/j.1467-6494.1975.tb00715.x>
- Howrigan, D. P., & MacDonald, K. B. (2008). Humor as a mental fitness indicator. *Evolutionary Psychology*, 6(4), 147470490800600411. <https://doi.org/10.1177/147470490800600411>
- Kaufman, J.C., Baer, J., Cropley, D.H., Reiter-Palmon, R., & Sinnett, S. (2013). Furious activity vs. understanding: How much expertise is needed to evaluate creative work? *Psychology of Aesthetics, Creativity, and the Arts*, 7(4), 332–340.
- Kellner, R., & Benedek, M. (2017). The role of creative potential and intelligence for humor production. *Psychology of Aesthetics, Creativity, and the Arts*, 11, 52–58.
- Koller, M. (2016). robustlmm: An R package for robust estimation of linear mixed-effects models. *Journal of Statistical Software*, 75, 1–24. <https://doi.org/10.18637/jss.v075.i06>
- Koppel, M.A., & Sechrest, L. (1970). A multitrait-multimethod matrix analysis of sense of humor. *Educational and Psychological Measurement*, 30(1), 77–85. <https://doi.org/10.1177/001316447003000107>
- Kozbelt, A., & Nishioka, K. (2010). Humor comprehension, humor production, and insight: An exploratory study. *Humor*, 23(3), 375–401. <https://doi.org/10.1515/humr.2010.017>
- Lebuda, I., & Benedek, M. (2023). A systematic framework of creative metacognition. *Physics of Life Reviews*, 46, 161–181.
- Long, H., & Wang, J. (2022). Dissecting reliability and validity evidence of subjective creativity assessment: A literature review. *Educational Psychology Review*, 34(3), 1399–1443.
- Maio, S., Dumas, D., Organisciak, P., & Runco, M. (2020). Is the reliability of objective originality scores confounded by elaboration? *Creativity Research Journal*, 32(3), 201–205. <https://doi.org/10.1080/10400419.2020.1818492>
- Masten, A.S. (1986). Humor and competence in school-aged children. *Child Development*, 57(2), 461–473. <https://doi.org/10.2307/1130601>
- McCrae, R.R., & Costa, Jr., Paul T. (2007). Brief versions of the NEO-PI-3. *Journal of Individual Differences*, 28(3), 116–128. <https://doi.org/10.1027/1614-0001.28.3.116>
- Nezlek, J. (2012). *Diary methods for social and personality psychology*. SAGE Publications Ltd. <https://doi.org/10.4135/9781446287903>
- 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.
- Omwake, L. (1937). A study of sense of humor: Its relation to sex, age, and personal characteristics. *Journal of Applied Psychology*, 21(6), 688–704. <https://doi.org/10.1037/h0055199>

- 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(2), 176–186. <https://doi.org/10.1037/aca0000230>
- Puente-Díaz, R. (2023). Metacognitive feelings as a source of information for the creative process: A conceptual exploration. *Journal of Intelligence*, 11(3), 49.
- Puente-Díaz, R., & Cavazos-Arroyo, J. (2022a). Creative self-efficacy and metacognitive feelings as sources of information when generating, evaluating, and selecting creative ideas: A metacognitive perspective. *The Journal of Creative Behavior*, 56(4), 647–658. <https://doi.org/10.1002/jocb.557>
- Puente-Díaz, R., & Cavazos-Arroyo, J. (2022b). Evaluative self-efficacy and its potential role in the evaluation and selection of ideas: A metacognitive perspective. *Creativity: Theories—Research—Applications*, 9(2), 29–42. <https://doi.org/10.2478/ctra-2022-0010>
- Puente-Díaz, R., Cavazos-Arroyo, J., & Puerta-Sierra, L. (2021). Idea generation, selection, and evaluation: A metacognitive approach. *The Journal of Creative Behavior*, 55(4), 1015–1027. <https://doi.org/10.1002/jocb.505>
- R Core Team. (2023). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. <https://www.R-project.org>
- Reiter-Palmon, R., Forthmann, B., & Barbot, B. (2019). Scoring divergent thinking tests: A review and systematic framework. *Psychology of Aesthetics, Creativity, and the Arts*, 13(2), 144–152.
- Robitzsch, A., Kiefer, T., & Wu, M. (2022). *TAM: Test Analysis Modules*. <https://CRAN.R-project.org/package=TAM>
- 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(1), 15–38. <https://doi.org/10.1037/aca0000244>
- 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., Greengross, G., Cotter, K.N., Christensen, A.P., & Gredlein, J.M. (2021). 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>
- Sun, C., Wang, M., Zhou, X., Liu, G., & Zhou, Z. (2022). Dynamic cognitive processes of humor generation: Activation and inhibition of information. *Psychological Research*. <https://doi.org/10.1007/s00426-022-01752-1>
- Sutu, A., Phetmisy, C.N., & Damian, R.I. (2021). Open to laugh: The role of openness to experience in humor production ability. *Psychology of Aesthetics, Creativity, and the Arts*, 15, 401–411. <https://doi.org/10.1037/aca0000298>
- Wang, J., & Long, H. (2022). Reexamining subjective creativity assessments in science tasks: An application of the rater-mediated assessment framework and many-facet Rasch model. *Psychology of Aesthetics, Creativity, and the Arts*. <https://doi.org/10.1037/aca0000470>
- Yanagida, T. (2023). *misty: Miscellaneous functions "T. Yanagida."* <https://CRAN.R-project.org/package=misty>

