An *Aha!* Walks into a Bar: Joke Completion as a Form of Insight Problem Solving

Alexander H. Bower (ahbower@uci.edu) and Mark Steyvers (mark.steyvers@uci.edu)
Department of Cognitive Sciences
University of California, Irvine
Irvine, CA 92697 USA

Abstract

The present work introduces a new insight problem task: joke completion. We found that performance and magnitude of insight within it correlated with an established task: rebus puzzles. However, participants performed worse on and took longer in joke completion problems than in their rebus counterparts. Further, the distribution of reported insight was bimodal only for rebuses, as should be expected of an insight problem. In joke completion problems, both self-estimated and externally-rated joke funniness correlated with reported insight. Challenging the assumption of impasse, performance and insight decreased as a function of trial time for both problem types, with the best and most insightful solutions submitted within the first 20 seconds. While this is a preliminary study, we argue that it signals a promising direction for the problem solving, humor, and creativity literatures by providing a new approach to capture insight in a manner conducive to linguistic and cognitive modeling techniques.

Keywords: creativity; humor; insight; problem solving

Introduction

Insight, or the sudden flash of understanding following a seemingly impossible problem, is one of psychology's oldest and greatest mysteries (Köhler,1925; Sternberg & Davidson, 1995). These unpredictable moments of revelation permeate our daily lives and are believed to have facilitated some of history's greatest achievements (Hill & Kemp, 2018; Jarman, 2014). However, the cognitive basis of insight remains poorly understood. This is due in part to the limitations of classic insight problems (Ash, Cushen, & Wiley, 2009). These shortcomings include their familiarity with typical subject populations, their self-contained nature, and the relative scarcity of data yielded per experiment due to their inherent difficulty and length.

Insight research has experienced a boon in recent decades thanks to problem sets designed to address these concerns. Informally designated "contemporary" insight problems (Webb, Little, & Cropper, 2018), these include rebus puzzles (MacGregor & Cunningham, 2009), anagrams (Novick & Sherman, 2003), and compound remote associate (CRA) problems (Bowden & Jung-Beeman, 2003). These problems have several advantages over their classic counterparts: many can be completed in single experimental sessions, they can be solved with or without the presence of insight, they can be supplemented with neuroimaging techniques, they have large normative datasets, and they allow for easy collection of solution accuracy and time latency data.

However, if we are to understand the myriad of contexts in which insight occurs in the real world and gain a more comprehensive account of its nature, we must continue to develop instruments with these desirable traits. Thus, we propose a similar task to satisfy this call: joke completion. In joke completion problems, we present participants with a subject (e.g., TREE) and a joke stem (e.g., "Walks into a bar…") and prompt them to create a punchline that is a functional wordplay resolving both in the same context (i.e., a pun). For example, someone may produce the following: *A TREE walks into a bar and says, "Can I order a **root** beer?"*

We will further detail joke completion problems' character and advantages in the following sections. For now, however, we will explain why we believe this constitutes an insight problem task and how its use meaningfully contributes to the literature.

Humor, Insight, and the Joke Completion Task

The connection between humor and insight is well-established in the context of joke comprehension and appreciation (Gick & Lockhart, 1995; Suls, 1972). The process of "getting" a joke and solving an insight problem share fundamental similarities: initial puzzlement, the need to resolve conflicting schemas (i.e., incongruity; Attardo & Raskin, 1991), and a sudden representational shift accompanied by a feeling of surprise and pleasure (Canestrari et al., 2018; Kozbelt & Nishioka, 2010). Further, performance on insight problem solving tasks, such as the remote associates test (Mednick, 1962), has been found to correlate with humor production and comprehension (Sitton & Pierce, 2004). This link is further supported by recent neuroimaging studies which suggest an overlap in brain areas activated by insight and humor comprehension (Amir et al., 2015; Tian et al., 2017).

While this connection exists in the passive context of joke comprehension, the role of insight in its active counterpart of joke *production* is far less explored and understood. Existing research in the area has typically used the cartoon caption generation task (e.g., Kudrowitz, 2010). However, this tool and its variants have held a nigh-monopolistic position in humor production literature. While tasks similar to ours exist, they either lack a sufficient sample size and statistical reporting (Kudrowitz, 2010), are used exclusively in the context of comprehension/appreciation (Brownell et al., 1983), and/or do not covary production ability with the experience of insight (Nusbaum, Silvia, & Beaty, 2017). Our joke completion task is the first to our knowledge to have an adequate sample, collect original productions, explicitly assess insight, and be operationalized to allow for theory-driven data collection.

There are three dominant parallels between traditionally-defined insight problems and joke completion:

1. Like other ill-defined problems, there is no clear mapping of the initial problem space nor an obvious, algorithmic solution path toward the goal (i.e., a punchline).
2. One must restructure initially incompatible problem elements (i.e., subject word and stem script) to form a new, compatible representation. This is expressed in the humor literature as the resolution of opposing scripts – or incongruity – in a joke.
3. To find a suitable solution, one must access semantically distant information to generate a surprising (i.e., non-obvious) target word. To do so, one must disregard high-frequent or irrelevant candidate words, thus overcoming an initial misdirection.

There are several practical and theoretical advantages to the joke completion task. Like other contemporary problems, they are short, easily administered, and varied. This task can also be adapted based on the nature and demands of the research question(s) of interest. For example, if one postulates that this task's difficulty varies based on the linguistic features of the subject and/or stem, this can be explored using techniques such as Latent Semantic Analysis (Landauer, Foltz, & Laham, 1998). Having normed databases of user-generated jokes may also inform the development of computational and cognitive models of humor (Kao, Levy, & Goodman, 2016). Thus, it is amenable to modern research techniques requiring numerous observations per condition.

One critical distinction of joke completion from other insight problems is the lack of a single proposed solution, instead featuring many solutions of varying quality. This feature of humorous production has been acknowledged and explored by researchers interested in divergent thinking and creativity (Derks & Hervas, 1988). Since contemporary problems are typically convergent, joke completion provides a new way to study the relative quality of solutions and their accompanying magnitudes of insight. Indeed, it is the ill-defined and open-ended nature of joke completion that makes it an arguably *better* proxy for the kind of insightful and creative problem solving experienced in the real world, opposed to problems relegated to experimental settings (e.g., CRAs, rebus puzzles).

The Present Study

We conducted a preliminary investigation into the validity of joke completion as an insight problem task. To do so, we evaluated how well it exhibits known features of such problems. We further calibrated it by comparing participants' solution behavior, the frequency of insight, and its general phenomenology with corresponding characteristics in an established insight problem task: rebus puzzles (MacGregor & Cunningham, 2009). Rebus puzzles combine visual, spatial, verbal and/or numerical clues to produce a common phrase. To retrieve this phrase, one must break assumptions of normal reading and interpretation. For example, the solution to "PUNISHMENT is "capital punishment."

Rebus puzzles were selected for a few reasons. Practically, rebuses have large data sets, require less time to solve than classic insight problems, are easy to administer and score, and allow for the collection of many data points in a single experimental session (Cunningham et al., 2009; MacGregor & Cunningham, 2009; Salvi et al., 2016; Threadgold et al., 2018). Additionally, the first author has observed that solving rebuses elicits a qualitatively similar *aha!* response to getting a joke (e.g., laughing, groaning). However, we emphasize that the focus of this study is not on the rebus task. We expect that any insight problem with a central verbal component should correlate with the joke completion task.

First, we examined the relationship between dimensions in the joke completion and rebus puzzle tasks. If joke completion constitutes an insight problem, we anticipated a similar frequency and distribution of reported insight between the tasks. Specifically, we expected a bimodal, "all-or-none" distribution signifying that insight largely either has or has not occurred (Smith & Kounios, 1996). We also compared overall performance and average time spent per trial between tasks.

Second, we examined trends and relationships within the joke completion task, itself. In accordance with previous studies, we predicted that if this is an insight task, self-estimated funniness (i.e., confidence) and externally-rated funniness (i.e., performance) would correlate with reported insight (Danek & Salvi, 2018; Salvi et al., 2016). We also tested the assumption of impasse – namely, that insight would increase with trial time, indicating an overcoming of these mental barriers (Ohlsson, 1992). We performed corresponding analyses for the rebus puzzle task to evaluate similarities.

Methods

Participants

One hundred and ten participants (n = 57 female) were recruited through Amazon Mechanical Turk. The age distribution was 21-29 (n = 11), 30-39 (n = 41), 40-49 (n = 28), 50-59 (n = 19), and 60+ (n = 11). Participants were compensated with $4 per hour, with typical participation time lasting between 30-40 mins. Eight participants were excluded from the study; five due to repeat participation, one due to missing data, one due to low-effort responses, and one due to being located outside of the U.S.

Materials

For the joke completion task, we constructed three joke stems pulled from popular culture and improvisational comedy games: 1) "A _____ walks into a bar…"; 2) "Waiter, there's a _____ in my soup!"; and 3) "_____ , I'm breaking up with you…". The blank spaces were occupied by different subjects that were locked for each stem. Subjects for the first joke stem were presented in the following order: TREE, DOCTOR, CAR, PIRATE, COMPUTER, and ARTIST.

Subjects for the second stem were presented in the following order: BIRD, SANDWICH, LAWYER, CAT, ASTRONAUT, and PENCIL. Subjects for the third and final joke stem were presented in the following order: OCEAN, GHOST, CLOCK, BOOK, BANK, and GUITAR. Subjects were selected on the basis of being nouns and having a diverse representation of living and nonliving things, tools, food, and occupations. In all, participants completed 18 joke completion trials. Examples of participant-generated jokes are presented in the Appendix.

For the rebus task, we adopted the 24 puzzles presented in MacGregor and Cunningham (2009). We scored performance based on their suggested solutions.

Procedure

Participation took place over Qualtrics. For the joke completion task, participants were told that they were going to write a series of punning jokes based on one of the joke stems, and that they must produce a punchline that is a wordplay of each trial's given subject. They were given three example jokes for each stem.

Each joke completion trial lasted 90 seconds. A countdown timer on the page alerted participants to how much time remained. Following the expiration of this time or their submission of a response, they were brought to a new page in which they were asked to rate how funny other people would find their joke on a Likert scale from 1 ("not funny at all") to 5 ("very funny"). They were also instructed to rate the level of insight that they experienced when they came up with the punchline on a scale from 1 ("no insight at all") to 5 ("complete insight"). "Insight" was described to them as follows: "INSIGHT is when a solution suddenly and unexpectedly pops into your head, accompanied by a strong burst of positive emotion - the *aha!* moment." Additionally, they were told that if they did not experience any insight at all, a rating of "1" was acceptable. Lastly, they completed a practice trial using the subject DOG before proceeding to the experiment proper.

In the second phase of the experiment, participants were prompted to rate the previous 10 participants' jokes for each respective stem/subject on the same 1-5 Likert scale, except now they were told to rate the jokes according to how funny they themselves found them to be. This phase typically comprised 180 jokes (18 jokes over 10 former participants), though it varied based on how many punchlines were left unanswered. This staggering was done due to the demands of a companion study collected alongside the present one. The stem/subject blocks were randomized to control for order effects. To be consistent with correct/incorrect solution rates in rebus scoring, joke submissions were designated as being "correct" if their average funniness rating was greater than or equal to 2.6 (surpassing the Likert median).

Next, participants completed the rebus puzzle task. They were told that these problems contain verbal and visual cues that form a familiar phrase and were given the following example: three of the word "SECRET" stacked vertically, with the top one circled. They were also given the target

solution: "Top Secret." Each rebus puzzle trial lasted 60 seconds. A countdown timer on the page altered participants to how much time remained. Following the expiration of this time or their submission of a response, they were brought to a new page in which they were asked to rate the level of insight they experienced when and if they solved the problem, using the same 5-point Likert scale and definition from the joke completion task

Results

We first tested whether performance, trial time, and rate of reported insight differed between joke completion and rebus puzzles. A paired samples t-test revealed that, on average, participants performed significantly worse on the joke completion task (M = 0.381, SD = 0.231) than on the rebus puzzle task (M = 0.474, SD = 0.213), $t(99) = 3.758$, $p < .001$, $BF_{10} = 68.94$. (Interestingly, there was no corresponding difference in reported difficulty between the two tasks, $t(99) = 3.078$, $p = 0.003$, $BF_{10} = 9.080$). Participants also spent significantly more time in joke completion trials (M = 37.25, SD = 14.177) than in rebus puzzle trials (M = 21.40, SD = 7.576), $t(99) = 11.94$, $p < .001$, $BF_{10} = 6.882e+17$. However, there was no significant difference in average reported insight between joke completion (M = 2.717, SD = 0.777) and rebus puzzles (M = 2.820, SD = 0.745), $t(99) = -2.62$, $p = 0.210$, $BF_{10} = 0.239$.

Next, we evaluated the relationship between performance in the two tasks. There was a significant positive correlation between individual funniness (M = 2.38, SD = 0.38) and rebus solution rate (M = 0.47, SD = 0.21), $r = 0.408$, $p < .001$, $BF_{10} > 100$. Linear regression revealed that an individual's mean funniness score can significantly predict their performance on the rebus task, $F(1,98) = 19.58$, $p < .001$, $BF_{10} > 100$. We further confirmed these results by discretizing jokes as funny or not funny by a mean 2.6 rating threshold, where this effect persisted, $r = 0.386$, $p < .001$, $BF_{10} > 100$. These relationships are demonstrated in Figure 1. Together, there was a positive relationship between performance on each task, indicating that individuals who produced funnier jokes also tended to be better rebus puzzle solvers.
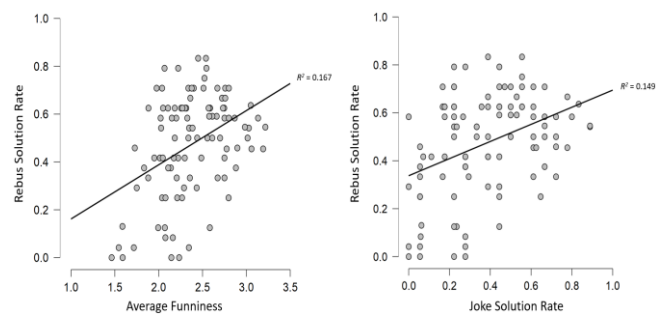

Figure 1: (Left) Average funniness and (right) discretized solution rate in joke completion problems by solution rate in rebus puzzles.

We also explored the distribution of reported insight between the two tasks to test the all-or-none hypothesis. Specifically, we expected a bimodal distribution of reported insight in both tasks, with the highest densities at 1 ("no insight at all") and 5 ("total insight"). While we observed this for rebus puzzles, with the highest frequencies at 1, or "no insight at all" (n = 771, or 32.1% of all cases) and 5, or "total insight" (n = 566, or 23.6% of all cases), we saw the opposite trend for joke completion, with the lowest frequencies for 5 (n = 127, or 7.1%) and 1 (n = 314, or 17.4%), respectively. This suggests that, while participants largely experienced no or complete insight in rebuses, completing jokes did not elicit the same bimodal response pattern. These results are depicted in Figure 2.
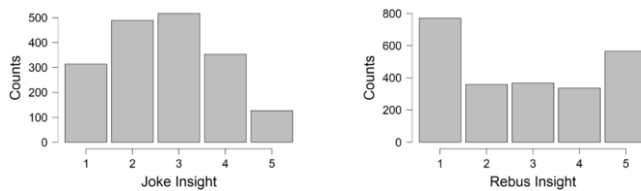

Figure 2: Frequency of reported insight in (left) joke completion and (right) rebus puzzles.

Next, we examined if performance on the joke completion task correlated with reported insight. There was a significant positive correlation between rated joke funniness and self-reported insight, $r_s = 0.218$, $p < .001$, $BF_{10} = 1.335e+17$, indicating that joke funniness increased with magnitude of insight. There was also a significant positive correlation between estimated joke funniness and self-reported insight, $r_s = 0.716$, $p < .001$, $BF_{10} > 100$, indicating that higher estimates of how funny jokes would be perceived to be (i.e., confidence) coincided with higher reports of insight.

These trends were echoed in rebus puzzles. There was a significant positive correlation between rebus accuracy and self-reported insight, $r_s = 0.616$, $p < .001$, $BF_{10} > 100$, indicating that magnitude of reported insight increased with successful performance on rebus puzzles. Similarly, a paired samples t-test revealed that there was a higher degree of insight reported alongside correct rebus solutions (n = 1130, M = 3.843, SD = 1.262) than incorrect rebus solutions (n = 1270, M = 1.909, SD = 1.232), $t(1129) = 37.84$, $p < .001$, $BF_{10} = 9.207e+198$.

Next, we examined if there was a period of impasse preceding responses – specifically, when they were funny/correct and corresponded with reported insight. Most responses for joke completion problems were submitted between 10-20 seconds into each trial (n = 421, or 24% of all cases). This period also saw the highest averages for joke funniness (M = 2.541, SD = 0.700) and reported insight (M = 2.903, SD = 1.197). Further, there was a significant negative correlation between joke trial time and self-reported insight, $r_s = -0.180$, $p < .001$, $BF_{10} = 5.138e+8$. This indicates that both performance and insight in joke completion problems decreased as trial time elapsed. These trends are illustrated in

Figures 3 and 4 (left). Thus, responses submitted earlier in trials were more likely to be funny and to accompany greater feelings of insight.
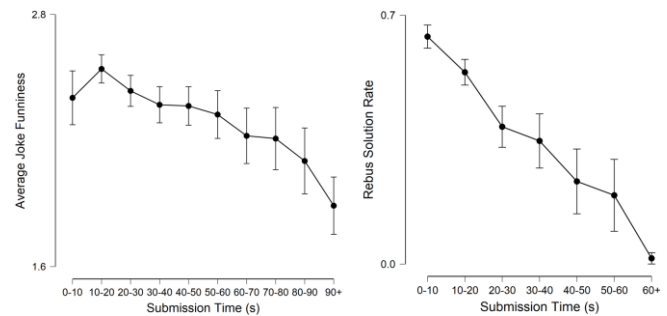

Figure 3: (Left) Average joke funniness and (right) rebus solution rate within each 10-second trial time interval (error bars denote 95% confidence intervals).

This trend is once again reflected in the rebus task. Most responses to rebus puzzles were submitted within the first 10 seconds of each trial (n = 841, or 35% of all cases). This period also saw the highest averages for correct responses (M = 0.640, SD = 0.481), and reported insight (M = 3.309, SD = 1.472). Further, there was a significant negative correlation between rebus trial time and self-reported insight, $r_s = -0.363$, $p < .001$, $BF_{10} = 2.974e+101$, as well as rebus accuracy, $r_s = -0.345$, $p < .001$, $BF_{10} = 2.101e+82$. This indicates that performance and reported insight in rebus puzzles decreased as trial time elapsed. This is illustrated in Figures 3 and 4 (right).
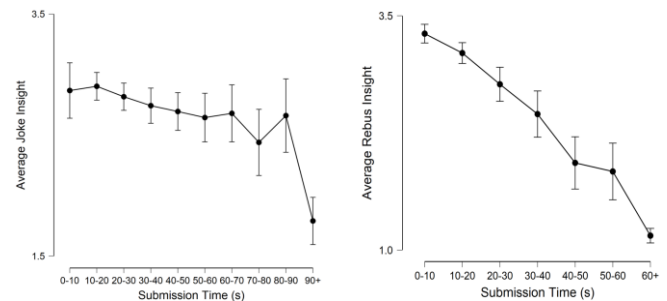

Figure 4: Average insight ratings for (left) jokes and (right) rebus puzzles within each 10-second trial time interval (error bars denote 95% confidence intervals).

Discussion

We conducted a preliminary investigation into the validity of joke completion as an insight problem. We found that, while participants spent more time in and performed significantly worse on the joke completion task, there was no difference in the magnitude of reported insight between it and the rebus puzzle task. Further, we found a positive relationship between performance on the two tasks.

Atypical of such problems, the distribution of reported insight was only bimodal for rebus puzzles and not joke completion problems. In fact, the latter saw the lowest

densities at "none" or "complete" insight – an inversion of what is to be expected in an insight problem. It is possible that, due to joke completion problems not converging on a single solution as rebuses do, participants chose the first available, rather than optimal, solution that came to mind. Future experiments may control for this by having participants generate several solutions and either rank them or identify a single "best" solution. This will also test a previous finding that humorous productions increase in funniness by output order (Derks & Hervas, 1988). It is unclear, however, if this increase would covary with magnitude of insight.

We found that both tasks saw quick submission times, with most responses entered within 0-10 seconds for rebuses and 10-20 seconds for joke completion. Further, submissions made during these intervals tended to be the most correct and coincide with the highest ratings of insight. There was also a downward trend for insight in both tasks, with average ratings falling as trial time progressed. This suggests that the best and most insightful responses were submitted early on, contradicting the impasse hypothesis. This has been found previously for CRA problems (Bower et al., 2019). It should be noted that these results fall within the typical 0-20 second solution latency window for rebuses (Salvi et al., 2016; Threadgold et al., 2018). It seems that this behavioral impasse may be relegated to classic insight problems. If impasse is denoted by task difficulty and increased trial latencies, the joke completion task arguably captures it better than rebuses and other contemporary problems. Another explanation for this disparity in response time is that, while solutions may have been reached earlier in jokes, submissions took longer because participants had to type out longer strings of words than they did for rebus solutions. Future studies using the joke completion task should evaluate the onset and duration of typing to tease this potential confound.

It is possible that these quick, high-quality submissions are better accounted for by general fluid intelligence (Cattell, 1963) than specific insight problem solving ability. Indeed, performance on insight ability constructs have been shown to correlate with fluid reasoning (Davidson, 1995; Gilhooly & Murphy, 2005; Paulewicz, Chuderski, & Nęcka, 2007) and working memory (Chuderski & Jastrzębsk, 2018) measures. However, the heterogenous nature of insight problems and their mental demands must again be acknowledged. Future studies exploring joke completion specifically as an insight task may benefit from evaluating performance between it and fluid reasoning tasks to extend our knowledge of its demand characteristics and describe individual differences in ability.

There are some limitations to this study. As with other insight research, it is possible that participants used confidence as the sole heuristic for reporting insight while avoiding other criteria – even if they were provided in our definition (e.g., suddenness, pleasure) (Danek & Salvi, 2018). This accuracy effect seems supported by the fact that confidence correlates strongly with reported insight experiences (Danek & Wiley, 2017; Webb, Little, & Cropper, 2016). However, even when confidence is mentioned as a

requisite of insight, this effect persists (Hedne, Norman, & Metcalfe, 2016; Salvi et al., 2016). Thus, it appears that explicitly defining confidence as a criterion is not a driving confound in these findings. Further, high confidence can sometimes accompany incorrect solutions (Danek & Wiley, 2017). In the future, however, it would be prudent to collect individual ratings to account for each dimension of the insight experience (e.g., pleasure, surprise, and suddenness).

Another limitation is the number of joke completion trials used in the present study. Future work should employ more problems to evaluate and potentially strengthen the usefulness and generalizability of this task. Joke completion should also be compared to a wider range of insight problems, such as CRAs and anagrams, to better validate it and understand its relationship with such existing measures.

Future work should also explore why certain joke solutions are perceived as funnier than others, as well as evaluate individual differences in humor production ability related to this task. As the present work captures participant responses, exploring the data rendered here may provide a key insight into the cognitive underpinnings of what makes something funny. This may be pursued through lexical analysis and modeling. Since this task may generate large amounts of behavioral data, it is amenable to such theory-driven approaches. This data can also be used to further test longstanding theories of humor (e.g., General Theory of Verbal Humor; Attardo & Raskin, 1991).

Lastly, we reiterate that our findings were correlational in nature. Future work should explore and describe the specific mechanisms underlying these observed relationships. Doing so has the potential to advance process-based theories for both insight and creative cognition research, in general. However, we believe that the present study lays a strong foundation for such efforts.

Conclusion

Like other contemporary problems, the joke completion task is easy to administer, has the potential to yield robust data sets, and embodies many of insight's requisite features. Further, we argue that it better approximates the kind of insightful problem solving encountered in daily life than in these other problems. It also significantly contributes to the problem solving, creativity, and humor literatures by providing a novel instrument through which to study humorous expression and insightful versus analytic problem solving behavior. The results and features of such problems are also amenable to linguistic analysis and cognitive modeling approaches, allowing investigators to make specific behavioral predictions based on them. While these problems should be further explored, they provide a promising avenue through which to study one of cognitive science's most elusive mysteries.

References

Amir, O., Biederman, I., Wang, Z., & Xu, X. (2013). Ha Ha! Versus Aha! A direct comparison of humor to

nonhumorous insight for determining the neural correlates of mirth. *Cerebral Cortex*, *25*(5), 1405-1413.

Ash, I. K., Cushen, P. J., & Wiley, J. (2009). Obstacles in investigating the role of restructuring in insightful problem solving. *The Journal of Problem Solving*, *2*(2), 6-41.

Attardo, S., & Raskin, V. (1991). Script theory revis(it)ed: Joke similarity and joke representation model. *Humor*, *4*(3), 293-348.

Bowden, E. M., & Jung-Beeman, M. (2003). Normative data for 144 compound remote associate problems. *Behavior Research Methods, Instruments, & Computers*, *35*(4), 634-639.

Bower, A. H., Burton, A., Steyvers, M., & Batchelder, W. H. (2019). An insight into language: Lexical and morphological effects in compound remote associate problem solving. In *Proceedings from the 41st annual conference of the cognitive science society* (pp. 166-173). Montreal, QC: Cognitive Science Society.

Brownell, H. H., Michel, D., Powelson, J., & Gardner, H. (1983). Surprise but not coherence: Sensitivity to verbal humor in right-hemisphere patients. *Brain and Language*, *18*(1), 20-27.

Canestrari, C., Branchini, E., Bianchi, I., Savardi, U., & Burro, R. (2018). Pleasures of the mind: What makes jokes and insight problems enjoyable. *Frontiers in Psychology*, *8*, 2297.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, *54*(1), 1-22.

Chuderski, A., & Jastrzębski, J. (2018). Much ado about aha!: Insight problem solving is strongly related to working memory capacity and reasoning ability. *Journal of Experimental Psychology: General*, *147*(2), 257-281.

Cunningham, J. B., MacGregor, J. N., Gibb, J., & Haar, J. (2009). Categories of insight and their correlates: An exploration of relationships among classic-type insight problems, rebus puzzles, remote associates and esoteric analogies. *The Journal of Creative Behavior*, *43*(4), 262-280.

Danek, A. H., & Salvi, C. (2018). Moment of truth: Why Aha! experiences are correct. *The Journal of Creative Behavior*, *0*, 1-3.

Danek, A. H., & Wiley, J. (2017). What about false insights? Deconstructing the Aha! experience along its multiple dimensions for correct and incorrect solutions separately. *Frontiers in Psychology*, *7*, 2077.

Davidson, J. E. (1995). The suddennes of insight. In R. J. Sternberg & J. E. Davidson (Eds.), *The nature of insight*. MIT Press.

Derks, P., & Hervas, D. (1988). Creativity in humor production: Quantity and quality in divergent thinking. *Bulletin of the Psychonomic Society*, *26*(1), 37-39.

Gick, M. L., & Lockhart, R. S. (1995). Cognitive and affective components of insight. In R. J. Sternberg & J. E. Davidson (Eds.), *The nature of insight*. MIT Press.

Gilhooly, K. J., & Murphy, P. (2005). Differentiating insight from non-insight problems. *Thinking & Reasoning*, *11*(3), 279-302.

Hedne, M. R., Norman, E., & Metcalfe, J. (2016). Intuitive feelings of warmth and confidence in insight and noninsight problem solving of magic tricks. *Frontiers in Psychology*, *7*, 1314.

Hill, G., & Kemp, S. M. (2018). Uh-oh! What have we missed? A qualitative investigation into everyday insight experience. *The Journal of Creative Behavior*, *52*, 201-211.

Jarman, M. S. (2014). Quantifying the qualitative: Measuring the insight experience. *Creativity Research Journal*, *26*(3), 276-288.

Kao, J. T., Levy, R., & Goodman, N. D. (2016). A computational model of linguistic humor in puns. *Cognitive Science*, *40*(5), 1270-1285.

Köhler, W. (1925). *The mentality of apes*. Routledge & Kegan Paul.

Kozbelt, A., and Nishioka, K. (2010). Humor comprehension, humor production, and insight: an exploratory study. *Humor*, *23*(3), 375–401.

Kudrowitz, B. M. (2010). *Haha and aha! Creativity, idea generation, improvisational humor, and product design.* [Doctoral dissertation, Massachusetts Institute of Technology]. DSpace@MIT.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2-3), 259-284.

MacGregor, J. N., & Cunningham, J. B. (2009). The effects of number and level of restructuring in insight problem solving. *The Journal of Problem Solving*, *2*(2), 130-140.

Mednick, S. A. (1962). The associative basis of the creative process. *Psychological Review*, *69*(3), 220-232.

Novick, L. R., & Sherman, S. J. (2003). On the nature of insight solutions: Evidence from skill differences in anagram solution. *The Quarterly Journal of Experimental Psychology Section A*, *56*(2), 351-382.

Nusbaum, E. C., Silvia, P. J., & Beaty, R. E. (2017). Ha ha? Assessing individual differences in humor production ability. *Psychology of Aesthetics, Creativity, and the Arts*, *11*(2), 231.

Ohlsson, S. (1992). Information-processing explanations of insight and related phenomena. In M. T. Keane & K. J. Gilhooly (Eds.), *Advances in the psychology of thinking* (Vol. 1). Harvester Wheatsheaf.

Paulewicz, B., Chuderski, A., & Nęcka, E. (2007). Insight problem solving, fluid intelligence, and executive control: A structural equation modeling approach. In *Proceedings of the 2nd European cognitive science conference* (pp. 586-591).

Salvi, C., Costantini, G., Bricolo, E., Perugini, M., & Beeman, M. (2016). Validation of Italian rebus puzzles and compound remote associate problems. *Behavior Research Methods*, *48*(2), 664-685.

Sitton, S. C., & Pierce, E. R. (2004). Synesthesia, creativity and puns. *Psychological Reports*, *95*(2), 577-580.

Smith, R. W., & Kounios, J. (1996). Sudden insight: All-or-none processing revealed by speed–accuracy decomposition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1443-1462.

Sternberg, R. J., & Davidson, J. E. (1995). *The nature of insight.* MIT Press.

Suls, J. M. (1972). A two-stage model for the appreciation of jokes and cartoons: An information-processing analysis. In J. H. Goldstein & P. E. McGhee (Eds.), *The psychology of humor: Theoretical perspectives and empirical issues*. New York: Academic Press.

Threadgold, E., Marsh, J. E., & Ball, L. J. (2018). Normative Data for 84 English Rebus Puzzles. *Frontiers in Psychology*, *9*, 2513.

Tian, F., Hou, Y., Zhu, W., Dietrich, A., Zhang, Q., Yang, W., ... & Cao, G. (2017). Getting the Joke: Insight during Humor Comprehension–Evidence from an fMRI Study. *Frontiers in Psychology*, *8*, 1835.

Webb, M. E., Little, D. R., & Cropper, S. J. (2016). Insight is not in the problem: Investigating insight in problem solving across task types. *Frontiers in Psychology*, *7*, 1424.

Webb, M. E., Little, D. R., & Cropper, S. J. (2018). Once more with feeling: Normative data for the aha experience in insight and noninsight problems. *Behavior Research Methods*, *50*(5), 2035-2056.

Appendix

Sample responses from the joke completion task with their respective mean funniness ratings.

Prompt	Mean funniness (1-2)	Mean funniness (3-4)
A TREE walks into a bar…	Are you doing OAKAY? (1.9)	And asks for a tall one (3)
A DOCTOR walks into a bar…	I'll take a shot of whiskey, please. (1.9)	and nurses a drink for a while. (3.7)
A COMPUTER walks into a bar…	Bartender: Hey it's great to C you (1.6)	The bartender says "this is a bar, you can't crash here." (3.4)
Waiter, there's an ASTRONAUT in my soup!	Don't worry… they'll blast off by tonight. (1.9)	That's our launch special. (3.1)
Waiter, there's a LAWYER in my soup!	Waiter: I'll sue what I can do about it. (1.9)	We have a pretty low bar for quality here sir. (3)
Waiter, there's a CAT in my soup!	Meow you can have a new pet. (1.8)	You've gotta be kitten me! (3.1)
GUITAR, I'm breaking up with you…	That is off key! (1.9)	You're just stringing me along. (3.7)
OCEAN, I'm breaking up with you…	I'll wave bye! (1.9)	I feel too tide down (3.4)
BOOK, I'm breaking up with you…	Way to write me off (1.9)	You judge people by their cover (3.9)