UNIVERSITY OF CALIFORNIA,
IRVINE

Insights into Problem Solving, Algorithm Aversion, and Theory of Mind

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Psychology

by

Alexander Harrison Bower

Dissertation Committee:
Professor Mark Steyvers, Chair
Professor Zygmunt Pizlo
Professor Sara Mednick

2022

DEDICATION

To my dad, who always believed in me.

To my wife: My love, my guide, my partner. You are the reason I am here today.

To every other first-generation college student who felt that academia is a world beyond them. Let this prove otherwise.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Thank you to my dissertation committee for their time and wisdom. Their guidance helped make this work possible.

Thank you to my advisor, Mark Steyvers, for his mentorship. His encouragement, tutelage, and support have empowered me to be the scientist I am.

Thank you to my former advisor, Bill Batchelder, who sadly passed away during my graduate journey. His curiosity, passion, and intellectual rigor remain a constant inspiration.

Thank you to Daniel Mann, who helped guide my pedagogical development and fostered my love for teaching.

Thank you to Jerry Rudmann, who showed me the impact a good teacher can have on their students and community.

Thank you to my MADLAB colleagues and to my collaborators at UC Santa Barbara for their valuable insights and support.

Thank you to my cohort, who helped me through the most challenging moments. I'm glad we had each other.

Thank you to my wife, Erica Heinrich, for her countless hours of wisdom and support throughout this process.

I would also like to thank and acknowledge the Irvine Initiative in AI, Law, and Society for their funding, which supported the work conducted for Chapters 3 and 4.

Lastly, thank you to the Cognitive Science Society for expressed permission to include Chapters 1 and 2 of my dissertation, the contents of which were originally published in the *41st* and *42nd Annual Conference of the Cognitive Science Society*, respectively. I also thank Nature Publishing Group for permission to include Chapter 3 of my dissertation, the contents of which were originally published in *Scientific Reports*. These works are all covered under the Creative Commons CC-BY license.

VITA

Alexander Harrison Bower

EDUCATION

Doctor of Philosophy in Psychology **2022**
University of California, Irvine *Irvine, CA*

Master of Arts in Psychology **2020**
University of California, Irvine *Irvine, CA*

Bachelor of Arts in Psychology **2014**
University of California, Irvine *Irvine, CA*

RESEARCH EXPERIENCE

Graduate Student Researcher **2015–2021**
University of California, Irvine *Irvine, CA*

Laboratory Manager **2014-2015**
University of California, Irvine *Irvine, CA*

Research Assistant **2013-2015**
University of California, Irvine *Irvine, CA*

TEACHING EXPERIENCE

Instructor of Record **2018–2020**
University of California, Irvine *Irvine, CA*

Course: Psychology Fundamentals 9B / Psychology and Social Behavior 11B

Teaching Assistant **2015–2020**
University of California, Irvine *Irvine, CA*

REFEREED JOURNAL PUBLICATIONS

Perceptions of AI engaging in human expression **2021**
Scientific Reports

Authors: **Bower, A. H.** & Steyvers, M.

Words of wisdom: Brief textual descriptions accurately convey domain knowledge **in prep**
Authors: **Bower, A. H.**, Han, N., Eckstein, M., & Steyvers, M.

REFEREED CONFERENCE PUBLICATIONS

An aha! walks into a bar: Joke completion as a form of insight problem solving **2020**
42nd Annual Conference of the Cognitive Science Society

Authors: **Bower, A. H.** & Steyvers, M.

An insight into language: Lexical and morphological effects in compound remote associate problem solving **2019**
41st Annual Conference of the Cognitive Science Society

Authors: **Bower, A. H.**, Burton, A., Steyvers, M., & Batchelder, W. H.

BOOK CHAPTERS

Teaching in the moment: Lessons from improv **2021**
Teaching Gradually: Practical Pedagogy for Graduate Students, by Graduate Students - Stylus Publishing

Author: **Bower, A. H.**

TALKS

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. Talk presented at 41st Annual Conference of the Cognitive Science Society, Montreal, QC.

Bower, A. H. & Batchelder, W. H. (2018). The effects of compound candidate frequencies in remote associate problem solving. Third Year Talk at the Cognitive Sciences Department, UC Irvine.

Bower, A. H., Cross, M. P., Pressman, S. D. (2014). The connections between facial expressions and heart rate variability. Talk presented at 21st Annual UC Irvine Undergraduate Research Symposium, Irvine, CA.

Bower, A. H. (2012). Effects of probabilistic judgment and conditional reasoning on paranormal belief. Talk presented at 5th Annual Southern California Psychology Conference for Students, Saddleback College, Mission Viejo, CA.

POSTERS

Bower, A. H. Steyvers, M. (2021). The funny thing about algorithm aversion: Investigating bias toward AI humor. 43nd Annual Conference of the Cognitive Science Society, Vienna, AT.

Bower, A. H. Steyvers, M. (2020). An aha! walks into a bar: Joke completion as a form of insight problem solving. 42nd Annual Conference of the Cognitive Science Society, Toronto, ON.

Cross, M. P., Cores, S. M. E., **Bower, A. H.**, Pressman, S. D. (2014). What types of social relationship variables are associated with respiratory sinus arrhythmia? Annual Scientific Meeting of the American Psychosomatic Society, San Francisco, CA.

AWARDS & HONORS

Most Promising Future Faculty Member (Monetary Award) **2020**
UC Irvine, Division of Teaching Excellence and Innovation
Pedagogical Fellow Honoree **2018**
UC Irvine, Division of Teaching Excellence and Innovation
Latin Honor: Summa cum Laude **2014**
UC Irvine
Psi Beta Outstanding Research Recognition Award **2012**
Irvine Valley College
Winner of Worth Publishers / Psi Beta Research Paper Competition **2012**
Irvine Valley College
Best Presentation **2012**
5th Annual Southern California Psychology Conference for Students
Psi Beta's Exemplary Research Presentation Award **2012**
Irvine Valley College

GRANTS AND FELLOWSHIPS

Irvine Initiative in AI, Law, and Society Fellowship **2021-2022**
UC Irvine
Teaching-as-Research Fellowship **2020**
Center for Integration of Research, Teaching, and Learning
Pedagogical Fellowship **2016-2019**
UC Irvine, Division of Teaching Excellence and Innovation

Associate Dean Fellowship **2018**
UC Irvine, School of Social Sciences
Undergraduate Research Opportunities Program **2013-2014**
UC Irvine

TEACHING CERTIFICATIONS

Summer Remote Teaching Institute Certification **2020**
UC Irvine, Division of Teaching Excellence and Innovation
UCI DTEI Certificate of Teaching Excellence **2018**
UC Irvine, Division of Teaching Excellence and Innovation
UCI CIRTL Certificate – Scholar Level **2017**
UC Irvine, Division of Teaching Excellence and Innovation

PEDAGOGICAL WORKSHOPS (DESIGNED/CONDUCTED)

Course Design Certificate Program **2019**
UC Irvine, Division of Teaching Excellence and Innovation
How people learn: Novice vs. expert: Things to consider about our students and strategies for facilitating learning **2019**
UC Irvine, Division of Teaching Excellence and Innovation
International TA Training **2017**
UC Irvine, Division of Teaching Excellence and Innovation
TA Professional Development Program **2017, 2019**
UC Irvine, Division of Teaching Excellence and Innovation

MENTORSHIP

Student: Francesca Sassoon **2020**
Project Title: Developing meditation and exercise interventions to mitigate stress and improve health outcomes for e-sports athletes.
UC Irvine, UROP Program
Student: Alexes Starkweather **2020**
Project Title: The effect of humor on long-term memory performance.
UC Irvine, UROP Program
Student: Louis Garcia **2019**
Project Title: The effect of sequential presentation in rebus puzzles.
UC Irvine, UROP Program

ABSTRACT OF THE DISSERTATION

Insights into Problem Solving, Algorithm Aversion, and Theory of Mind

By

Alexander Harrison Bower

Doctor of Philosophy in Psychology

University of California, Irvine, 2022

Professor Mark Steyvers, Chair

This dissertation explores several important topics in the cognitive sciences: Insight, algorithm aversion, and theory of mind. First, I tackle the challenge of understanding insight, or the "aha!" experience. In Chapter 1, I use a well-defined class of problems (compound remote associates: Bowden and Jung-Beeman, 2003; Mednick, 1962) to test if various lexical and morphological properties affect solution retrieval and the likelihood of insight. While performance is only affected by one property (familiarity), other findings contest popular assumptions about insight. Namely, the reported magnitude of insight decreases with trial time (challenging the impasse hypothesis) and increases with the number of cues solved (challenging the all-or-none hypothesis). In Chapter 2, I introduce a new insight problem task: Joke completion. I find that performance and magnitude of insight within it correlate with an established task (rebus puzzles: MacGregor and Cunningham, 2009), though the distribution of reported insight is not bimodal, as was expected. Further, self-estimated and externally-rated joke funniness correlate with reported insight. Lastly, performance and reported insight decrease with trial time, again refuting impasse. In Chapter 3, I shift focus to a more recent concern: Algorithm aversion. As AI becomes an integral part of our daily lives, it is crucial to identify and anticipate biases regarding it. Since aversion mostly occurs in subjective contexts, I test whether people find jokes less humorous if they believe an AI created them. When joke source is ambiguous, people exhibit bias toward jokes they

identify as human-created. However, this bias disappears when the (purported) source of jokes is stated. This demonstrates that such biases are weaker than proposed and are dependent on framing. In Chapter 4, I conclude by exploring another perennial topic: Theory of mind. Namely, I examine whether a few words provide an accurate estimate of another person's domain knowledge. This was done by having one group of people ("informants") describe images depicting various domains (e.g., video games, astronomy), then having a second group ("evaluators") make pairwise comparisons between these informants regarding who they believe is more knowledgeable, based on these descriptions. Strikingly, evaluators perform above chance at identifying the more knowledgeable informants when only one description is available (around seven words, on average). Further, the most knowledgeable informants produce the most specific facts and the most knowledgeable evaluators are the most sensitive to false information. However, less knowledgeable evaluators treat specific statements interchangeably, regardless of their factuality. These results show the inferential power a mere few words hold.

0.1 Introduction

How and why do the solutions to once impossible problems seemingly emerge out of nowhere, resulting in an "aha!" experience? Is it possible to predict the likelihood of such occurrences? Why do humans have biases against artificial intelligence (AI) systems, and can such biases be overcome? Can we know the extent of someone else's knowledge through limited interaction with them? These are some of the critical questions that this dissertation explores, reflecting some of the oldest and most challenging mysteries in the cognitive sciences.

In Chapter 1, I examine how different lexical properties affect cognitive processes involved in a popular class of insight problems: Compound remote associates (CRAs: Bowden and Jung-Beeman, 2003; Mednick, 1962). These properties are familiarity, lexeme meaning dominance, and semantic transparency. I find that a higher proportion of problems are solved when they are presented beginning with the most familiar word cues, but not when they begin with right-headed dominant or the most semantically transparent cues. Further, I find that participants focus their efforts disproportionately on the first and last cues, that subjective ratings of insight decrease as trial times elapses, and that the magnitude of reported insight increases with the number of cues successfully solved. This suggests that participants can monitor their progress in such problems. These results contest longstanding assumptions of requisite periods of impasse and the absence of incremental progress in insightful problem solving.

In Chapter 2, I introduce a new insight problem task: Joke completion. I find that individual performance and magnitude of insight within it correlate with an established insight problem: Rebus puzzles (MacGregor and Cunningham, 2009). However, participants perform worse at and take longer in joke completion problems than in their rebus counterparts. Further, the distribution of reported insight is bimodal only for rebuses, as should be expected of an insight problem. In joke completion problems, both self-estimated and externally-rated

joke funniness correlate with reported insight. Challenging the assumption of impasse, performance and insight decrease as a function of trial time for both problem types, with the best and most insightful solutions being submitted within the first 20 seconds. While this is a preliminary study, I 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.

In Chapter 3, my research shifts focus toward an increasingly vital endeavor: Understanding human biases toward AI systems. Though humans should defer to the superior judgement of AI in an increasing number of domains, certain biases prevent us from doing so. Understanding when and why these biases occur represents a central challenge for human-computer interaction. One proposed source of such bias is task subjectivity. I test this hypothesis by having both real and purported AI engage in one of the most subjective expressions possible: Humor. Across two experiments, I address the following questions: Will people rate jokes as less funny if they believe an AI created them? When asked to rate jokes and guess their likeliest source, participants evaluate jokes that they attribute to humans as the funniest and those to AI as the least funny. However, when these same jokes are explicitly framed as either human or AI-created, there is no such difference in ratings. My findings demonstrate that user attitudes toward AI are more malleable than once thought - even when they (seemingly) attempt the most fundamental of human expressions.

Finally, in Chapter 4 my focus shifts once more to explore the topic of theory of mind. Namely, whether a mere few words accurately reveal the extent of someone's domain knowledge. I do this by having one group of people ("informants") answer trivia questions and describe various images belonging to a category (e.g., cartoons, Japanese cuisine). A second group of people ("evaluators") then decide who is more knowledgeable within pairs of informants, based on their image descriptions. We find that evaluators generally perform above chance at identifying the most knowledgeable informants (65% with only one descrip-

tion available; seven words, on average). The most knowledgeable informants produce the most specific facts, while the most knowledgeable evaluators are the most adept at identifying false information. Less knowledgeable evaluators tend to treat correct and incorrect statements interchangeably, though there is a low base rate of incorrect statements, overall. Together, these findings demonstrate the power a few words hold when inferring others' domain knowledge.

While the work herein may seem somewhat dissonant, it is unified by two common themes. First is the prominence of evaluating language - both in problem solving and in discerning the abilities, intentions, and mental states of others (our so-called *theory of mind*). Second is the use of jokes as stimuli to examine the complexities of natural language processing and subjective judgment.

As my graduate career is founded in the study of insight and this topic constitutes the majority of my empirical work (Chapters 1 and 2), the following subsections of the introduction are dedicated to outlining the history of its research, its challenges, and the most promising avenues for theoretic progress. For my work exploring algorithm bias (Chapter 3) and theory of mind (Chapter 4), relevant backgrounds are detailed within each chapter's respective introduction.

0.1.1 New(er) Approaches to Demystifying Insight

One of the oldest and most famous accounts of insight concerns the ancient Greek polymath Archimedes. His cousin, King Hiero, suspected that a gold crown he had commissioned was, in fact, partly silver. Pressured by rumors, Hiero contracted Archimedes to determine if this was the case.

While pondering this issue during one of his daily baths, Archimedes noticed that the water

he sunk into flowed over the edges. His interest piqued by this seemingly mundane detail, he submerged himself further, pondering that the water being displaced was equal to his body's volume. Then... *aha!* As if from nowhere, the solution emerged: Since silver weighs less than gold, a mixed-element crown would have to be bulkier to match the volume of one made from pure gold – thus, the former would displace more water than the latter. He leaped from the bath, exclaiming "Eureka! Eureka!" (translated as "I have found it! I have found it!"). In his excitement, Archimedes took to the streets unclothed, introducing two senses of the term "flash of insight."

While Archimedes might be the only person to commit this *faux pas*, he is far from the only one to experience a burst of insight. These sudden, revelatory leaps from utter loss to complete understanding have sparked our collective imagination and heralded some of history's greatest technical and artistic triumphs.

These moments are not always so groundbreaking, however. Anecdotal evidence suggests that the experience of insight is common and occurs in a variety of contexts and across occupational backgrounds, such as in the fields of IT, healthcare, and business management (Jarman, 2014). It can also occur in different domains, such as when solving intellectual, practical, or personal problems (Hill and Kemp, 2018). Further, these experiences are not relegated to experimental settings and occur frequently in the natural world, such as at night, at work, in the shower, at home, when it is quiet, during transport, while exercising, or when in nature (Ovington et al., 2018).

Yet, despite its role in human achievement, prominence in everyday life, and being one of our discipline's oldest interests (Köhler, 1925), the cognitive basis of insight remains poorly understood. This has endowed it with an almost mystic status (Bowden et al., 2005), dividing researchers and theorists. For instance, it remains hotly debated whether insight constitutes its own special process, independent of conventional problem solving behavior (Bowden et al., 2005). More fundamentally, there is no consensus as to whether insight is a cognitive process

or an epiphenomenon (Chronicle et al., 2004; Sternberg and Davidson, 1995). Indeed, there is no single, operational definition by which researchers use to capture insight! Given these obstacles, it should be unsurprising that this field is both riddled with challenge and rife with opportunity.

Fortunately, researchers have made great strides in recent years toward better understanding this famously elusive topic. We have developed clever paradigms (Bowden and Jung-Beeman, 2003; MacGregor and Cunningham, 2008), explored its affective and behavioral components (Danek et al., 2014, 2016; Danek and Wiley, 2017; Ovington et al., 2018; Webb et al., 2016), and evaluated the domain-specific knowledge and processes employed in different problem sets (Cunningham et al., 2009; MacGregor and Cunningham, 2009; Webb et al., 2016). It is in following these trends that my work has tackled the problem of understanding insight. Before we proceed, however, I will provide a more comprehensive view of insight, as well as a brief history of its study.

0.1.1.1 What is Insight?

Broadly, insight can be described as the sudden, certain realization of a solution to a once difficult problem (Metcalfe and Wiebe, 1987; Sternberg and Davidson, 1995). It often requires restructuring how one views a problem (Ash et al., 2009) and follows a period of impasse, or a halt in progress after all solution paths are seemingly exhausted (Ohlsson, 1992). Problems used to evoke insight often employ some initial misdirection (Cunningham et al., 2009) and necessitate ignoring incorrect, high-frequency candidate solutions. Further, solvers often have little to no conscious access to the cognitive processes preceding such moments. This differs from so-called "analytic" problem solving, which involves the deliberate use of strategies and operations in service of gradually reaching a clearly-defined goal state (Fleck and Weisberg, 2013; Newell et al., 1972). An example of this is the Tower of Hanoi (e.g., Kotovsky et al., 1985). Last and perhaps most central to the insight experience is the

aha! moment, or the strong, sudden burst of emotion and sense of certainty after reaching a solution (Gick and Lockhart, 1995; Shen et al., 2016). While insight has been described in holistic terms (e.g., Jung-Beeman et al., 2004), there has been recent interest in evaluating its phenomenological components, such as confidence, impasse, surprise, and pleasure (Danek et al., 2014; Webb et al., 2018). (Note: There is converging evidence against some of these assumptions, such as insightful ends precluding strategic means, or the necessity of impasse – both of which are discussed later).

Historically, insight has proven difficult to study (Ash et al., 2009). Traditional methods relied upon so-called "classic" or "pure" insight problems (Cunningham et al., 2009). These include the Duncker candle problem (Duncker and Lees, 1945), the nine-dot problem (Figure 1; Burnham and Davis, 1969), and riddle-type verbal problems (e.g., Metcalfe and Wiebe, 1987). Such problems stem from Gestalt psychology (see Ohlsson, 1984) and are prefaced on the need to restructure a problem space into a more appropriate – though initially imperceptible - representation (Ash and Wiley, 2006; Gilhooly and Murphy, 2005; Weisberg, 1995). For example, the nine-dot problem presents solvers with a 3x3 matrix of dots and tasks them to connect all of the dots with four straight lines without lifting the pencil from the paper. To solve this, one must break free of the self-imposed constraint that all lines must remain inside the boundary of the dots and, quite literally, think outside-the-box.

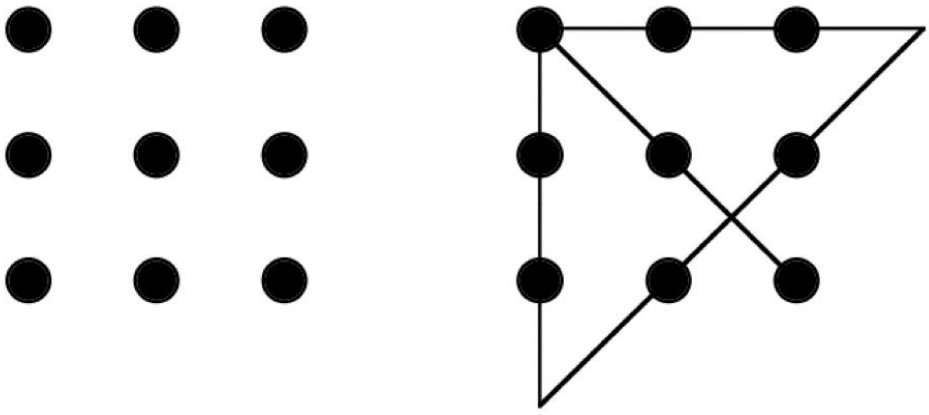

Figure 1: *The nine-dot problem (left) and one of its possible solutions (right).*

However, these problems have several critical limitations. Due to their long residence in psychological literature and classrooms, the typical subject population (i.e., college undergraduates) has likely encountered them before - thus, rendering the need for restructuring moot. Second, many of these problems are one-and-done (such as the nine-dot), limiting the systematic study of the cognitive mechanisms employed in them. Third, these problems take a great deal of time to solve (if they are solved at all), further restricting the amount of data that can be collected in an experimental session. Fourth, there is the popular, though circular assumption that a problem is an "insight problem" because it is used to study insight (Öllinger and Knoblich, 2009). Until recently, insight problems were typically contrasted with "noninsight" problems (e.g., arithmetic; Metcalfe and Wiebe, 1987) as a control without collecting self-reported insight from participants (see Webb et al., 2016). This is further exacerbated by the fact that these presumed noninsight problems produce affective and behavioral patterns indicative of insight (Davidson, 1995; Webb et al., 2016), and that false insights can occur (Danek et al., 2014; Hedne et al., 2016). Finally, it is only recently that there has been a systematic investigation into the strength and reliability of different problem types to elicit insight (Danek et al., 2016; Webb et al., 2018), despite the acknowledged heterogeneity of such problems (Weisberg, 1995).

In the following section, I will expand on these and other issues, in addition to highlighting proposed solutions.

0.1.2 Solving the Problem of Insight

0.1.2.1 Problems of Definition and Measurement

Most studies use a global assessment of insight and give participants an operational definition by which to judge if it has occurred. These definitions vary, being subject to what different researchers find most salient about the experience. For example, the following is provided

by Jung-Beeman et al. (2004) to their participants (emphasis mine):

> A feeling of insight is a kind of "Aha!" characterized by **suddenness** and **obviousnes. You may not be sure how you came up with the answer**, but are relatively **confident** that it is correct without having to mentally check it. It is as though the answer came into mind all at once – when you first thought of the word, you simply knew it was the answer. This feeling does not have to be overwhelming, but should resemble what was just described. (p. 507)

This is typical of what is provided for holistic judgment of insight. It captures the dimensions of suddenness, obviousness, certainty, and lack of conscious strategy. However, it is missing the key dimension of *pleasure* – the positive emotional burst at finding (what appears to be) the correct solution (Danek et al., 2014). Other researchers rely on as few as two dimensions, such as suddenness and surprise (Cushen and Wiley, 2012). Thus, there is no universal criteria by which to demonstrate that insight has occurred. This limitation is well-recognized, yet remains deeply entrenched in the field's methodology.

There is a recent trend of studies which measure the various components implicated in insight (e.g., pleasure, surprise/suddenness, relief, impasse, confidence, strength of aha!) (Danek and Wiley, 2017; Webb et al., 2016, 2018). I believe this is a proper direction to take – especially if we are to disentangle these constituents - namely confidence, which appears to be a heuristic for identifying insight (Danek and Salvi, 2020).

Traditional research and theory posit that feelings of insight only accompany correct solutions, since a proper restructuring of the problem is assumed to have taken place (Wertheimer, 1925). In fact, many studies collect no measures of subjective insight and simply assume that it occurred by merit of an "insight problem" being solved (Topolinski and Reber, 2010). This is especially problematic when using hybrid problems that can be solved through insightful or analytic means (Bowden and Jung-Beeman, 2003). Further, empirical evidence

suggests that erroneous or "false" insights can take place during incorrect solution attempts (albeit at a lower frequencies) (Danek et al., 2014; Hedne et al., 2016; Salvi et al., 2016). An amusing example of this is presented in Figure 2 (adapted from Blackmore, 2011). To assume that this is definitionally impossible ignores a crucial aspect of insight's phenomenology and limits theoretical development. There is also evidence for "negative insights," or "d'oh!/uh-oh" moments, in which people experience insight after being shown the solution or realize it after a failed attempt (Gick and Lockhart, 1995; Hill and Kemp, 2018). Lastly, not only can solvers report certainty in incorrect solutions, but they can offer unanticipated, yet viable, solutions (Olteţeanu, 2016). Such instances of divergent thinking and creativity must be accounted for and explored to attain a comprehensive account of insight.

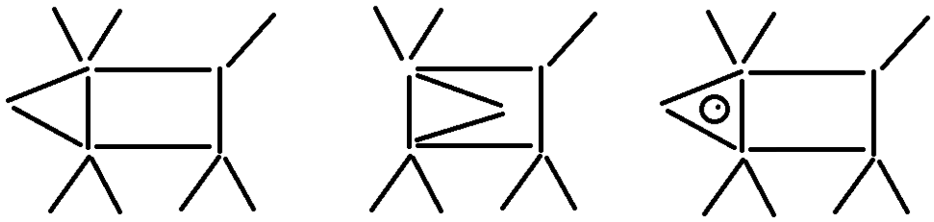

Figure 2: *A problem which asks the solver to move two lines to make the cow face the other way (left).* The predefined solution (center). A novel solution produced by someone who misinterpreted the instruction "move two lines" as "make two moves," leading them to draw a circle and dot to represent a rightward gaze (right). While incorrect, the solver nevertheless reported having an insight.

Finally, how do we know we have captured insight? A popular direction in recent decades has been to conduct brain imaging studies to identify insight's neural correlates (Aziz-Zadeh et al., 2009; Bowden et al., 2005; Jung-Beeman et al., 2004; Kounios and Beeman, 2014). While informative, there are two limitations to this approach. First, it still relies upon correlational results and rarely makes behavioral predictions based on problem features. Second, it is impractical for most empirical research. Thus, in addition to mapping the so-called "insightful brain" (Shen et al., 2013), we must also explore why certain problems elicit feelings of insight, while others do not. To accomplish this, we must develop domain-specific measures, make specific behavioral predictions based on problem features, and develop and

validate datasets for new and existing problems. I will explore the latter issue in the following section.

0.1.2.2 Lack of Cognitively-Defined Behavioral Instruments with Normative Datasets

While used for the better part of a century, classic insight problems have several properties that make them suboptimal for empirical study. Due to their long residence in psychological literature and college classrooms, the typical subject population has likely encountered them before. Additionally, these problems take a great deal of time to solve, limiting the amount of data collected in single experimental sessions. For example, a seminal paper by Schooler et al. (1993) only used seven problems, each of which took several minutes to solve.

To illustrate these challenges, consider the following problem which they used:

> A dealer in antique coins got an offer to buy a beautiful bronze coin. The coin had an emperor's head on one side and the date 544 B.C. stamped on the other. The dealer examined the coin, but instead of buying it, he called the police. Why?
>
> *Solution*: In 544 B.C. Christ had not been born, so a coin from that time would not be marked "B.C." (before Christ). (p. 182)

Though such problems may elicit insight, they have several practical limitations. The first is that classic problems such as this are largely idiosyncratic. While they may share surface features (Gilhooly and Murphy, 2005), they are rarely defined according to the cognitive functions used to solve them. Thus, beyond broad demarcations such as "spatial" or "verbal" problems (the latter of which incorporates our example), they do not belong to clearly defined and internally reliable problem classes. The second issue is such problems' difficulty. In the

case that a participant solves one, it will usually be after a long period, yielding a single data point. Further, while such studies may report main effects, they infrequently provide solution rates, response times, and rates of reported insight for individual problems. Indeed, Schooler et al. (1993) also fall victim to assuming insight has taken place by merit of participants solving such "insight problems" in contrast to "noninsight" problems (i.e., those solved in a logical, incremental fashion, such as through deduction).

To continue with the Schooler et al. (1993) example, reviewing their other six problems reveals little functional overlap between the cognitive mechanisms and strategies required to solve each problem. Thus, even if problem differences regarding success rate, experienced insight, and solution latency are reported, it would be difficult to determine *why* these differences exist. Further, even if these problems were defined in an operationally precise manner, their scant number limits the confidence with which we may draw conclusions.

I would like to emphasize that the point here is not to negate or detract from the impact of Schooler et al. (1993)'s study. It is merely one, salient example of the many issues prevalent in the field.

Because of these limitations, researchers have called for experimental paradigms and problem sets that are procedural, easy to administer, facilitate robust data collection over single experimental sessions, reliably elicit reports of insight, and are classified according to the cognitive functions involved in solving them (Batchelder and Alexander, 2012).

Consequently, insight research has experienced a boon in recent decades thanks to problem sets designed to address these concerns (Bowden et al., 2005). Informally designated "contemporary" insight problems (Webb et al., 2016), these include rebus puzzles (MacGregor and Cunningham, 2008, 2009), anagrams (Novick and Sherman, 2003; Salvi et al., 2016), and magic tricks (Danek et al., 2014). Perhaps the most prominent, however, are compound remote associate (CRA) problems (Bowden and Jung-Beeman, 2003), which were developed

as a modified version of the Remote Associates Test (RAT) pioneered by Mednick (1962). In CRAs, individuals are presented with three cue words and must produce a solution word that is common to all three, forming compound words and/or phrases. For example, the solution to the problem triad "COTTAGE, SWISS, CAKE" is CHEESE (forming "COTTAGE CHEESE," SWISS CHEESE," and "CHEESECAKE," respectively). This task is designed such that a solver must break free of high-frequency associations to access globally satisfactory solutions (Gupta et al., 2012).

CRAs have several desirable properties: 1) Many can be completed in short experimental sessions (i.e., 30 seconds, opposed to the six minutes plus allotted in some classic problems), 2) They can be solved through both insightful and non-insightful means, 3) Participants have reliably demonstrated that they can make subjective judgments of insight regarding them, 4) They can be used in neuroimaging studies to identify neural correlates of insight, 5) It is easy to collect solution accuracy and time latency data regarding them and, importantly, 6) They have a large ($n = 144$), normed database of problems (Bowden and Jung-Beeman, 2003).

While certainly useful, this database is simultaneously overutilized and underexplored. These problems are relied upon to de facto cause and study insight, though behavioral predictions are rarely made based on their features (e.g., lexical and syntactic properties). A few noteworthy exceptions include investigations into semantic convergence (Bowers et al., 1990), priming effects (Wiley, 1998; Topolinski and Strack, 2008), and modeling search behavior, such as through Latent Semantic Analysis (LSA) (Gupta et al., 2012; Smith et al., 2013). While promising, such ventures are sparse in the literature. I extend this effort in Chapter 1.

Further, even when these well-defined problems are used, it is uncommon for researchers to share their own normative data. This is especially true of studies adopting Bowden and Jung-Beeman's (2003) database, where it is rare to find normative data for solution frequency,

occurrence of insight, and/or time courses for individual problems, making it difficult to judge their reliability.

Exceptions to this trend include Threadgold et al.'s (2018) database of 84 English rebus puzzles and Salvi et al.'s (2016) validation of Italian problem sets. However, these cases are exceptions that prove the rule. If we are to judge the validity and reliability of these instruments as measures of insight, we must develop and share such data structures. This is especially critical if researchers are to extend existing problems sets or develop new problem classes entirely.

To address these issues, I introduce a new problem solving task used to capture and explore the insight phenomenon: Joke completion (Chapter 2). Further, I share its normative data in Table B.2 to encourage collaboration, inform future studies, and help guide theoretical development.

While it is unlikely to develop a general theory or comprehensive model of insight problem solving, we can make meaningful, informed progress in specific problem domains (Batchelder and Alexander, 2012). To do so, however, will require these instruments and a transparent, concerted effort to share their results. Through investigating the strength and reliability of different problem types to elicit the affective dimensions and cognitive processes of insight, we may better understand its phenomenology and make more precise predictions about – or even facilitate – its occurrence.

0.1.2.3 Lack of Qualitative Investigation

To gain a comprehensive understanding of this phenomenon, we must explore the reported occurrence, nature, and strategies (or lack thereof) in purported insight. This includes qualitative research (e.g., Jarman, 2014; Ovington et al., 2018), analysis of open responses corresponding to experimental data (e.g., Danek et al., 2014), and exploring the characteristic

differences of problem types (e.g., Webb et al., 2016).

In reviewing the literature, I have come to believe that there is an aversion toward strictly qualitative efforts. Due to the elusive nature of insight, cognitive scientists seem to fear making bedfellows with a "soft" approach, as if such an effort would somehow tarnish the rigor of the field. This proprietary attitude is, in fact, part of what limits our progress. Much can be gleaned from qualitative data and the metacognition of respondents. Understanding the reported mental processes of problem solvers is just as critical for theory development as the raw data which they produce (Batchelder and Alexander, 2012).

Thus, to make meaningful progress we must undertake qualitative endeavors to inform and supplement our quantitative efforts. Otherwise, we will remain ignorant of critical aspects of insight's character and commit similar oversights to those made in the past (e.g., false insights, ignoring valid but unanticipated responses, negative insight; Danek et al., 2014; Hill and Kemp, 2018).

0.1.3 Research Aims

As just demonstrated, the challenges inherent to the field are plentiful. This has led some researchers to doubt the plausibility of a formal theory of insight (Batchelder and Alexander, 2012). While I agree that such an encompassing account is unlikely (at least, at present), I do believe that we can make domain-specific progress in service of this goal by addressing the specific limitations and needs of the field. Accordingly, my work in this realm has three primary aims:

1. Detail the cognitive functions used in specific problem classes and make behavioral predictions based on their features and demands. Specifically, I explore language processing in CRA problems (Chapter 1).

2. Develop new problem classes and protocols which identify, facilitate, and operationally describe the occurrence of insight (Chapter 2).

3. Explore the strategies and affective constituents reported in insight problems to better account for the qualitative nature of insightful versus analytic problem solving (Chapter 2).

Chapter 1

An insight into language: Investigating lexical and morphological effects in compound remote associate problem solving

1.1 Introduction

Insight has sparked some of history's greatest accomplishments – from Einstein's special theory of relativity to Newton's universal law of gravitation. These sudden "aha!" moments also permeate our everyday lives – from practical household problems to puzzles in video games. However, our understanding of the processes underlying insight have remained subject to empirical gaps and theoretical debate (Batchelder and Alexander, 2012). Indeed, a prevailing assumption of the literature has been that insight occurs by merit of one solving an "insight problem" (Topolinski and Reber, 2010). To make meaningful progress toward

understanding insight, we must first explore the cognitive mechanisms involved in problems in which it is reported.

One such class of problems are CRAs (Bowden and Jung-Beeman, 2003). The CRA task was developed as a modified version of the RAT (Mednick, 1962), which has been correlated with performance in insight problems. The difference between the original RAT and CRAs is that the latter only uses structural associates based on syntax (Worthen and Clark, 1971). In CRAs, people are presented three cue words and must produce a solution word which is common to all three, forming compound words and phrases. For example, the solution to the triad "COTTAGE, SWISS, CAKE" is CHEESE (forming "COTTAGE CHEESE," "SWISS CHEESE," and "CHEESECAKE," respectively). The task is designed such that a solver must break free of high-frequency associations to access globally satisfactory solutions.

CRAs have many advantages over classic insight problems:

1. They have large, normed databases.
2. Many can be completed in single, short experimental sessions.
3. They can be solved with and without insight
4. People have reliably demonstrated that they can make subjective judgments of insight regarding them.
5. They can be used in neuroimaging studies to identify the neural correlates of insight.
6. They can be supplemented with time-based measures of solution latencies.

As a result, they have been widely used to explore various cognitive domains, such as intuition (Topolinski and Strack, 2008), sleep (Cai et al., 2009), and computational/deep learning (Olteţeanu, 2015).

Much of the past research on the RAT and CRAs has defaulted to a correlational account that simply assumes insight and ignores the underlying processes that may drive it. This problem was highlighted by Topolinski and Reber (2010), who point out that many researchers neglect to explain the phenomenology of insight, yet rely on it as a sufficient condition.

Recent studies have attempted to mend this by modeling CRA performance. Gupta et al. (2012) were among the first to provide a formalized account of individual differences in CRA search behavior. They employed a norm-based model that defined the best guess at solutions based on the average of cues in the Word Association Space (WAS) (Steyvers et al., 2005). This was contrasted with a frequency-biased model that assumes people's search is biased by word fluency, based on work by Griffiths et al. (2007) with PageRank and associative frequency. As predicted, they found that the probability of a given response is biased toward high-frequency words. Thus, people perform poorly if they are biased in favor of high-frequency incorrect words, precluding access to low-frequency correct responses.

This work was extended by Olteţeanu and Falomir (2015), who developed the comRAT-C; a computational model that solves compound RAT queries, based on a cognitive theoretical framework for creative problem solving (CreaCogs) (Olteţeanu, 2016). The knowledge base (KB) comprising the CRAs themselves used language data (2-grams pruned for relevance) from the Corpus of Contemporary English (COCA). They found that the comRAT-C used a convergence process similar to that of human solvers, and that the frequency of cues in the KB influences responses. The comRAT-C was able to correctly solve 64 of the 144 items in Bowden and Jung-Beeman's (2003) list of normed CRAs, in addition to suggesting unlisted, yet plausible solutions in more than 20 cases – suggesting its own form of creativity. Overall, their study laid a solid computational framework for formalizing the processes in CRA problem solving.

A promising experimental approach was taken by Smith et al. (2013), who used Latent Semantic Analysis (LSA) to evaluate the similarity between people's guesses, word cues, and

answers. They accomplished this by having participants enter every word considered while searching for the answer, regardless of their correctness. By doing this, they focused on the search processes used when generating candidate answers through a probabilistic sampling framework. They found sequential dependencies between responses in a problem, with subjects generating semantically similar chains of responses. Additionally, people seemed to focus primarily on one cue at a time. However, their procedure assumes that guesses accurately reflect the implicit nature of the search, even though the very act of conscious report may alter the search process.

The main body of work on CRAs has focused on associative aspects − not the requisite that responses be syntactic compounds (with the notable exception of Olteţeanu and Falomir, 2015). Indeed, research has largely ignored the morphological properties of the compounds themselves and how they affect performance and the likelihood of reported insight. We have thus failed to adequately address a critical aspect of their character. This approach has potentially restricted us from discovering how people attain insight in these problems. A look at the nature of compounds and their lexical elements is necessary to better understand the underlying cognitive processes involved in these problems.

1.1.1 Compound Word Research

Early research of compound words used a lexical decision paradigm (Taft and Forster, 1975), which measures peoples' response times (RT) in classifying words and nonwords. One such study found that only the lexical status of the first constituent word in a compound affects processing, with longer RT for word-word and word-nonword pairs (e.g., DUST-WORTH, FOOTMILGE) than nonword-word and nonword-nonword pairs (e.g., TROW-BREAK, MOWDFLISK) (Taft and Forster, 1975). Thus, it appears that morphological decomposition takes place when processing compound words, instead of the words being

stored and retrieved as a whole.

There has been considerable work on visual word recognition in recent years, facilitated by databases containing lexical characteristics and behavioral data, such as latencies of word naming and lexical decisions for large sets of words (e.g., Balota et al., 2007) and investigations of word length (New et al., 2006). Though initially focused on monosyllabic and monomorphemic words, this work has been extended to address processing in multisyllabic words (Yap and Balota, 2009) and English compound words.

Research suggests that English compounds are processed differently from length and frequency-matched monomorphemic words. For instance, both semantically-transparent compounds (e.g., ROSEBUD) and opaque compounds (e.g., HOGWASH) are processed more quickly than their monomorphemic counterparts (e.g., GIRAFFE) (Ji et al., 2011). This sense of morphological complexity has ignited debate in the psycholinguistic literature, with competing perspectives on compound representation and processing (see Fiorentino and Poeppel, 2007).

The current study investigates the roles of three lexical properties involved in compound processing and, by extension, CRAs: Word familiarity, semantic transparency, and lexeme meaning dominance. Thus, we investigate if and how they differentially affect CRA performance and the likelihood of insight. To do this, we used Juhasz et al.'s (2015) database of 629 compound words to construct 21 novel CRA problems. This database, which adapted items from the English Lexicon Project (ELP: Balota et al., 2007), compiled subjective ratings for six properties believed to affect morphological processing. The questionnaires used by these authors are available in their Supplementary Materials. We will now briefly explore each of these selected properties and justify their inclusion in this study.

1.1.1.1 Familiarity

Whole word frequencies may be interpreted as analogous to whole word access and have thus been studied in compound word recognition (Juhasz et al., 2015). However, English compound frequencies tend to be low relative to other languages, resulting in experimental challenges and a consequential gap compared to Dutch (Kuperman et al., 2009) and Finnish (Kuperman et al., 2008) counterparts. Rated familiarity can be regarded as a measure of subjective frequency and has been demonstrated to affect word recognition in English monomorphemic words. In particular, familiarity has been shown to influence eye fixation duration, along with word frequency (Juhasz and Rayner, 2003). This was further demonstrated in an experiment by Juhasz et al. (2008), which found that familiarity affected gaze duration for both long (ten or more letters) and short (seven or fewer letters) English compound words.

We thus contend that ratings of familiarity can be used as a subjective proxy for word frequency and have a role in affecting morphological processing and CRA performance.

1.1.1.2 Semantic Transparency

Semantic transparency also plays an important role in how compounds are processed and represented (Libben, 1998). A fully transparent compound is one in which both constituents contribute to the meaning of the compound word (e.g., SUNLIGHT), while a fully opaque compound is one in which neither constituent contributes to its meaning (e.g., FLAPJACK). There are also partially-opaque compounds, in which only one constituent contributes to the compound's meaning (e.g., JAYWALK, CHEAPSKATE) (Juhasz et al., 2015).

Libben (1998) proposed a model in which semantic transparency is represented in two distinct ways: The semantic relationship between the meaning of a constituent morpheme within a

compound, and the meaning of the morpheme independent of it. For example, the opacity of the compound SHOEHORN results from HORN not being transparently related to the compound as a whole, whereas SHOE is fully transparent. Thus, it is classified as a T-O compound (wherein T = transparent, O = opaque). Compounds require some level of semantic transparency to be tied to semantic representations of their lexemes. Using a lexical decision task, Libben et al. (2003) found that fully opaque and T-O compounds were responded to more slowly than other compound types, though there was a significant priming effect on all four compound types relative to neutral primes.

Research has demonstrated that semantically transparent compounds are especially susceptible to morphological decomposition, and that semantic priming only seems to occur when there is at least one transparent lexeme. Using Dutch compounds, Sandra (1990) used semantic associates of constituents as primes for transparent (e.g., BIRTHDAY primed by DEATH), opaque (e.g., SUNDAY primed by MOON), and pseudo-compounds (e.g., BOYCOTT primed by GIRL). Facilitatory priming effects were only observed for constituents in transparent compounds.

1.1.1.3 Lexeme Meaning Dominance

Compared to other languages, English compound words tend to be right-headed (i.e., the second constituent word – or lexeme - is the semantic head of the compound). This lexemic dominance primarily defines the meaning of the compound. In a study by Inhoff et al. (2008), location and word frequencies of lexemes were manipulated in lexical decision, naming, and sentence reading tasks. They found an effect for larger word frequency for the dominant lexeme in each task. Lexeme dominance also affected first fixations on compound words. These results suggest the headedness of a compound affects how it is recognized and subsequently processed.

Since all the word cues presented in the CRAs in this experiment are the second lexemes, their contribution to the overall meaning of the compound should affect the speed of access when solving each problem.

1.2 The Present Study

In accordance with the evidence above, we predicted that CRA problems beginning with word cues that are the most familiar, are the most semantically transparent, and have right-headed lexeme dominance would result in the highest levels of performance and reporting of insight.

To test this, we staggered the presentation of word cues on-screen, with cues either increasing or decreasing in ratings for the relevant lexical domain. Thus, we actively constrained and manipulated the search processes used by solvers. As CRA triads are commonly presented at once, this presents an experimental departure that, we hypothesize, differentially affects performance and captures some of the latent features of this process. To our knowledge, this is one of the first studies to actively manipulate cue presentation in CRAs in such a way with precise behavioral predictions.

Another departure is in how problems are scored. CRAs are typically scored according to whether a submission conforms to all three cue words and to the suggested response of the researchers, precluding other "incorrect," yet plausible responses. This does not allow for the investigation of partially correct responses, in which fewer than three cues are satisfied by the solution candidate. We address this issue using a lexicon to test whether submitted responses form valid compounds against each individual cue presented, either as a prefix or suffix. This allows for a more comprehensive account of the processes and strategies employed in such problems.

Together, this study contributes to the CRA literature in three major ways. First, it uses a staggered presentation of word cues, facilitating semantic activation and lexical search behavior. Second, it investigates the morphological properties of the compounds themselves. Finally, it uses partial scoring for each word cue. The goal of this study is to determine how the aforementioned lexical properties affect solution retrieval and if they influence performance and the probability of reported insight.

1.3 Method

1.3.1 Participants

Each experimental condition was composed of two counterbalanced groups, comprising six groups total. All participants ($N = 128$) were University of California, Irvine undergraduate students who were awarded course credit through the SONA system for their role in the study. The age distribution was 18-21 ($n = 110$), 22-25 ($n = 11$), 26-30 ($n = 5$), and 31-40 ($n = 2$). Everyone identified as a native English speaker, with 53 participants identifying as multilingual (though additional languages spoken were not specified).

In the Familiarity condition, Group 1 consisted of 23 participants ($n = 21$ females), and Group 2 consisted of 22 participants ($n = 14$ females).

In the Lexeme Meaning Dominance condition, Group 3 consisted of 21 participants ($n = 14$ females), and Group 4 consisted of 21 participants ($n = 19$ females).

Lastly, in the Semantic Transparency condition, Group 5 consisted of 21 participants ($n = 16$ females), and Group 6 consisted of 20 participants ($n = 18$ females).

Fourteen participants were excluded from the final analysis, as they did not meet the criteria

of answering at least two of the three practice problems correctly.

1.3.2 Materials

We constructed 21 novel CRA problems from compounds that had at least three common stems (thus forming three cues with a common solution). For example, there are 10 compound words in Juhasz et al.'s (2015) database with the shared prefixed stem FOOT. The mean ratings for whatever variable was in question (on a 1-to-7 scale for familiarity and transparency and on a 1-to-10 scale for lexeme meaning dominance) were then sorted in descending order and the words with the highest and lowest values were selected. The mean of these two values was then calculated and the compound word with that value or its closest approximate was selected as the middle term. Using the same example of FOOT for the variable of familiarity: FOOTPRINT has the highest value at 7, FOOTPATH has the closest approximate to the mean with a value of 5.85, and FOOTHILL has the lowest value at 4.71. This forms the CRA problem "PRINT, PATH, HILL," with the solution FOOT. All compounds in this database begin with a prefixed solution stem. Thus, unlike other studies, the solution is always the first lexeme in the compound.

In the event of a tie between two compound word values, the compound with the closest letter length to the other two words was selected. If the competing compound had the same length, the tie was broken by identifying which one more closely matched the mean age of acquisition value of the other two compounds.

Due to the limited number of candidate items, some words were repeated in both problem and solution terms. For example, PORT occurs in the problems "PORT, BASE, SICK" and "FOOD, PORT, BOARD." There was also an instance of a having the same phonetic representation (WASTE and WAIST). Participants were told that words may occur more than once both as cues and as solutions.

Each condition was counterbalanced so that problems were presented in both ascending and descending order across two groups. This was done to control for potential order effects.

1.3.3 Procedure

Participants were given instructions and a working definition of "insight" (*Insight occurs when the answer suddenly pops into your head, accompanied by a strong burst of positive emotion ("aha!").*). They were then given an example CRA problem ("CREAM, SKATE, WATER," solution = ICE) and were asked to complete three practice problems with feedback. All four of these problems were pulled from Bowden and Jung-Beeman's (2003) set of CRA norms and were the four easiest problems with uniformly prefixed solution stems.

The experiment was conducted using a MATLAB interface. Word cues were presented sequentially with 5 s delays between each cue. The first word cue appeared in the left-center of the screen, the second appeared in the center, and the third in the right-center. Cues remained on-screen after their presentation for the remainder of the trial. Figure 1.1 demonstrates the display of a typical problem trial.

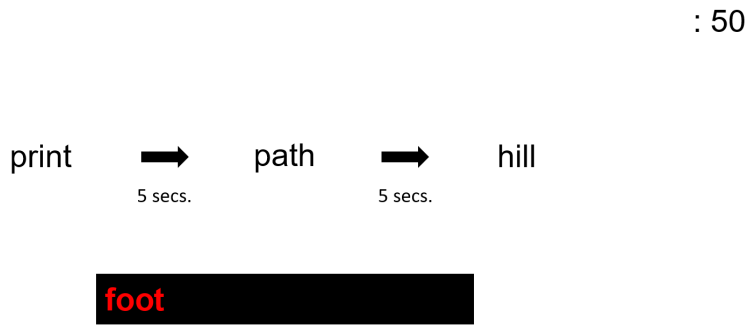

Figure 1.1: *Example of problem trial.*

Each trial lasted for 1 min. A countdown timer appeared on the top right-hand corner of

the screen when 50 s remained and turned red when 10 s remained. Participants typed their responses in a black field below the cues. They were encouraged to answer as quickly and as accurately as possible. They could submit their response at any time following the presentation of the third cue. Participants were forced to proceed after the minute had expired and whatever was typed into the solution box was accepted as the submitted response.

Following each problem trial, participants were asked to report the level of insight they experienced on a scale of 1 ("no insight") to 7 ("complete insight"). They were also reminded of its operational definition on the bottom of the screen.

At the end of the experiment, participants were asked to provide a brief (150-word maximum) description of what strategies they used to solve these problems. We also asked them to describe the difference they felt between solving problems with and without the feeling of insight. This was done to determine individual differences in reporting criteria and as a check for cross-validity with our definition. This data will also be evaluated to inform future, related experiments. Participants were scored based on how quickly and accurately they responded to each problem.

1.4 Results and Discussion

First, we tested the hypothesis that presentation order of cues according to ratings in each lexical condition would affect performance. These results are shown in Figure 1.2. Note: "Direction" denotes whether the cue presentation sequentially increased or decreased for the lexical property in question (that is, "Down" indicates that the first cue had the highest rating for the property, while "Up" started with the lowest rating). Normative data for the familiarity, lexeme meaning dominance, and semantic transparency conditions are presented

in Tables A.1, A.2, and A.3 in Appendix A, respectively. It appears that the only observed difference was in familiarity, with a higher proportion of problems successfully solved when they began with the most familiar word cue ($M = 0.383$, $SD = 0.126$), rather than the least familiar cue ($M = 0.301$, $SD = 0.161$, $t(226) = 4.304$, $p < .001$, $d = 0.570$). The estimated Bayes factor suggested that the data were .001:1 in favor of the alternative hypothesis, suggesting decisive evidence for a presentation order effect (Jeffreys, 1961). While this finding was not shared by the other properties (lexeme meaning dominance and semantic transparency), there are several other important findings – some of which challenge widely-accepted assumptions regarding the "special process" view (Bowden et al., 2005) of the insight phenomenon.

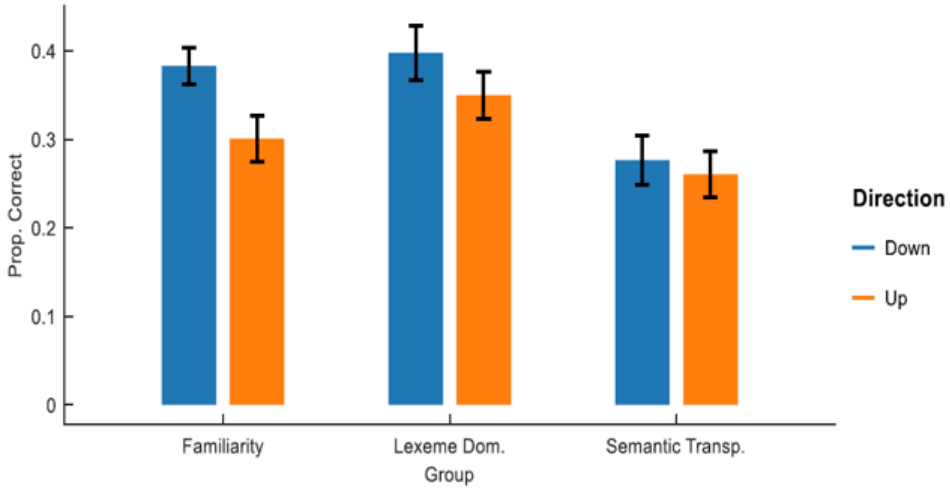

Figure 1.2: *Differences in performance for each lexical property.*

For further analysis, we used English compounds derived from the Touchstone Applied Science Associates (TASA) corpus and derived a lexicon of over 122,000 words, including hyphenated compounds. We used this lexicon to test whether a submitted response forms a valid compound against each individual cue presented, either as a prefix or suffix. The results of individual cue matches are shown in Figure 1.3, which demonstrates that the proportion correct for suffixes is smaller than that of prefixes. Further, submitted responses had a smaller likelihood of being valid prefixes for middle cues ($M = 0.351$, $SD = 0.162$) than

for first cues ($M = 0.411$, $SD = 0.182$) and last cues ($M = 0.402$, $SD = 0.174$, $F(2,228) = 8.049$, $p < .001$). The estimated Bayes factor suggested that the data were .032:1 in favor of the alternative hypothesis, or rather, 31.25 times more likely to occur under the model including an effect for cue position than the model without it, providing strong evidence for its effect. This suggests that participants were alternating between cues when attempting to generate a solution, rather than using parallel processing. One possible explanation is that since cue presentation was staggered – and thus their search was guided – there may be primacy and recency effects, whereby they were able to test and generate more candidate solutions following the first word cue, then worked backwards once all cues were presented using the third cue.

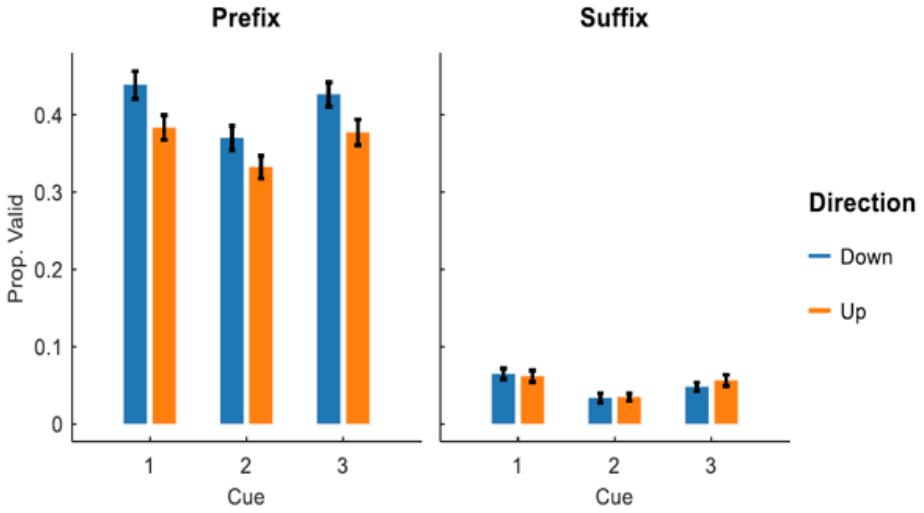

Figure 1.3: *Proportion of valid prefixed (left) and suffixed (right) responses for each word cue position, according to lexicon.*

Another interesting finding was that ratings of insight decreased as time elapsed throughout trials, as demonstrated in Figure 1.4. This finding holds for both correct and incorrect trials. This seemingly challenges the popular assertion that there must be a period of impasse, or mental block, preceding the experience of insight (Ohlsson, 1992). To the contrary, there were higher ratings of insight in the immediate time following the presentation of all three cues (i.e., 10-20 s) than in the time before the end of each trial (50-60 s). It is possible

that participants simply rated solutions that they perceived to be correct as insightful *de facto* (hence, being submitted quickly), and correctly rejected the occurrence of insight for incorrect solutions proffered as a final guess before trials ended.

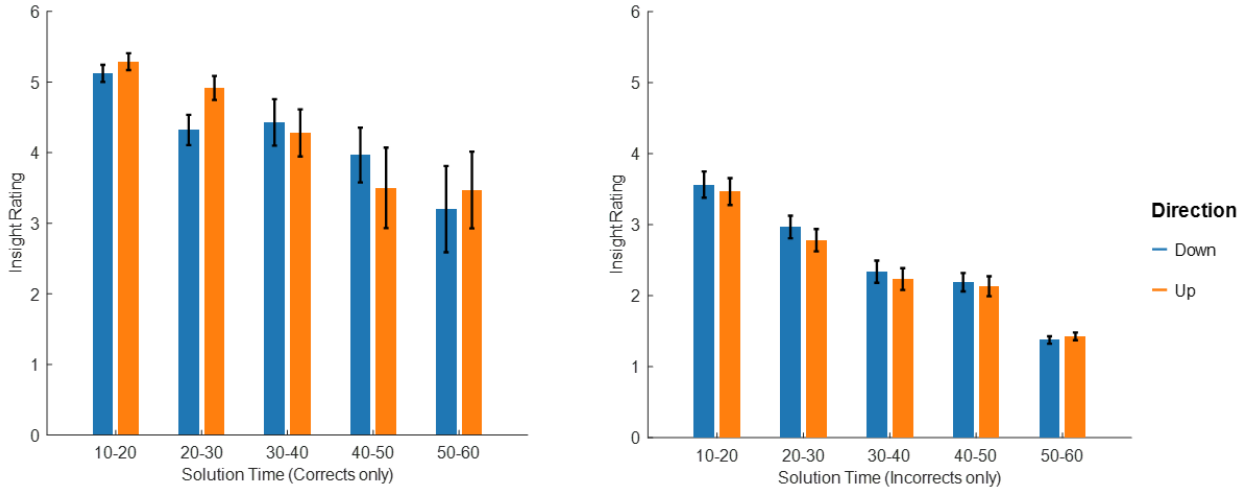

Figure 1.4: *Insight ratings as a function of solution time (s) for correct solutions (left) and incorrect solutions (right).*

Finally, there is the reporting of insight, itself. As demonstrated in Figure 1.5, fewer cues were likely to be solved as trial time elapsed. The magnitude of reported insight also increased along with the number of cues correctly solved. Rather than an all-or-none experience – the "sudden, certain burst" frequently reported and used as a necessary criterion (Chronicle et al., 2004) - it appears that participants used ratings of insight to indicate confidence in their answers. Indeed, these ratings increased as a function of the number of cues their proposed solution fit. There is not the presence of absolute insight for totally correct trials (in which all three cues are satisfied by the proposed solution), nor the absence of insight if this is not achieved. Rather, it exists on a continuum. This suggests more of an analytic approach, in which participants reliably monitor their progress in each problem and the likelihood of success using insight as a proxy for said progress. This contrasts previous research which states that incremental feelings of "warmth" do not precede moments of insight and are instead relegated to analytic or non-insightful problem solving strategies (Metcalfe and Wiebe, 1987). It should be noted again that a property of CRAs is that they

can be solved with or without insight. What we argue here is the usefulness of insight ratings in CRAs to indicate perceived progress.

One limitation to the current work is that it used novel CRAs instead of those with established norms for difficulty and magnitude/frequency of reported insight (such as Bowden and Jung-Beeman, 2003). Applying lexical ratings to such a database for the dimensions present in Juhasz et al. (2015) would be informative for future studies. Future research could also have subjects generate their own list of compounds given a set of word stems. Through doing this, researchers could collect latency data for how long people take to produce words, indicating their availability in memory. Researchers could also use LSA to analyze these participant-generated sequences of compounds to describe search behavior. These data could be applied across participants to establish cross-reliability and a more naturalistic set of items with norms.

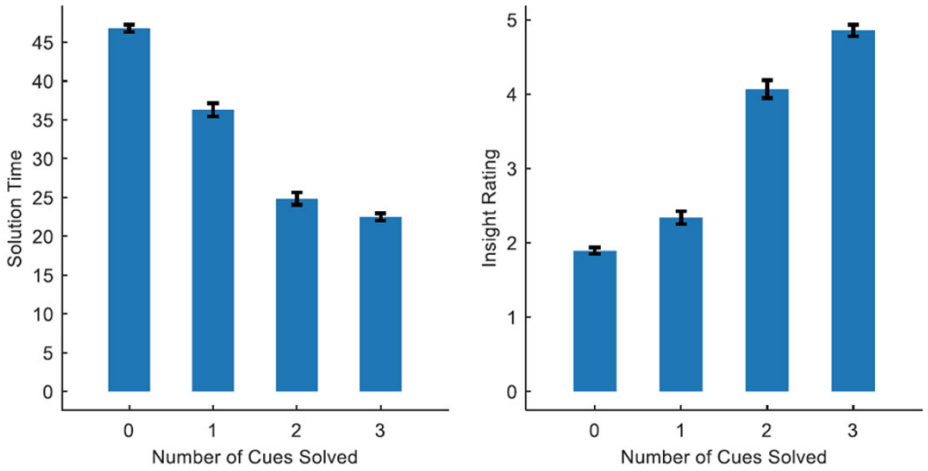

Figure 1.5: *Solution time (s) as a function of cues solved (left); Insight ratings as a function of cues solved (right).*

All problems in the current study have suggested solutions that are the first lexeme in the compound. Since stem placement seems to matter in the processing of compounds (Taft and Forster, 1975), it may be beneficial to compile and use compounds with common prefixes and suffixes in future studies.

The self-identified magnitude of insight in our study was still based on subjective report. While this study focuses on the cognitive processes underlying these problems, rather than attempting to formalize insight in a significant manner, similar studies attempting to so may wish to include neural and/or physiological covariates to identify correlates of insight (e.g., EEG, fMRI, skin conductance, eye-tracking) (see Bowden et al. (2005) for suggested neurocognitive approaches). Future studies should also explore participants' differences in reporting thresholds, as one person may be more willing to identify the occurrence of insight than another. These individual differences could be applied to a signal detection theory model.

This study offers modest progress into understanding the linguistic contributors to CRA processing. There are other factors that should be investigated, such as if compounds with noun-noun links and adjective-noun links differentially affect performance. Other variables to investigate are word length effect (New et al., 2006), imageability, age of acquisition, sensory experience, or a combination of the above.

There may also be a reading direction effect present, as cue presentation always proceeded from left-to-right on the screen. To circumvent potential perceptual biases, future studies using a similar design may benefit from counterbalancing the order of reading direction, as well.

Lastly, it is important to be remain cognizant that not all insight problems are the same, and the phenomenology in CRAs may differ from that of other insight problems. It would be premature to make any sweeping statements about modeling insight from discoveries made in one class of problems.

1.5 Conclusion

If we are to solve the problem of insight, we must better understand the cognitive processes underlying the methods we use to study it. Since we have largely neglected to explore these commonly-used procedures, we have defaulted to assumptions that they are "insight problems" simply because they elicit feelings of insight (based on the many and inconsistent criteria of researchers). While there have been both promising empirical and theoretical attempts to address this problem in recent years, much work remains. Better understanding the driving mechanisms, including lexical properties, within CRA problem solving will further inform us about how creativity is exercised and, perhaps, how insight is attained.

Chapter 2

An *aha!* walks into a bar: Joke completion as a form of insight problem solving

2.1 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 and Davidson, 1995). These unpredictable moments of revelation permeate our daily lives and are believed to have facilitated some of history's greatest achievements (Hill and 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 et al., 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 et al., 2018), these include rebus puzzles (MacGregor and Cunningham, 2009), anagrams (Novick and Sherman, 2003), and CRA problems (Bowden and Jung-Beeman, 2003), which are based on the RAT pioneered by Mednick (1962). 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 data sets, 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.

2.1.1 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 and 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 and Raskin, 1991), and a

sudden representational shift accompanied by a feeling of surprise and pleasure (Canestrari et al., 2018; Kozbelt and Nishioka, 2010). Further, performance on insight problem solving tasks, such as the RAT (Mednick, 1962), has been found to correlate with humor production and comprehension (Sitton and 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 this area has typically used cartoon caption generation tasks (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 et al., 2017). Our joke completion task is the first to our knowledge to have a robust sample, collect original productions, explicitly assess insight, and be operationalized to allow for theory-driven data collection.

There are three main 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 (Attardo et al., 2002).

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.

To demonstrate this process, consider the following joke stem from our task: "A TREE walks into a bar..." As a reminder, the goal is to generate a novel punchline that is a wordplay of the subject (in this case, TREE) and adheres to the script (in this case, a bar). There is no obvious, optimal punchline (i.e., solution), though there may be inappropriate noise in the form of weak or unrelated candidates as you conduct your search through semantic memory. One heuristic for solving this problem is listing candidate words related to the subject and testing them one-by-one against the stem. For instance, you may generate the words BARK, OAK, and LEAF. Consequently, you may produce the following: 1) "and starts BARKING"; 2) "are you OAKAY?"; or 3) "The bartender says, "You gotta LEAF!". Incidentally, these were all solutions offered by participants in the present study and are ordered here according to increasing ratings of funniness. Further examples from the current study are presented in Table B.1 in Appendix B.

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 LSA (Landauer et al., 1998). Having normed databases of user-generated jokes may also inform the development of computational and cognitive models of humor (Kao et al., 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 and 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 and rebus puzzles).

2.2 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 characteristics in an established insight problem task: Rebus puzzles (MacGregor and 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 and 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 and 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 a valid insight task, self-estimated funniness (i.e., confidence) and externally-rated funniness (i.e., performance) would correlate with reported insight (Danek and Salvi, 2020; Salvi et al., 2016). We also tested the assumption of impasse – namely, that insight would increase with trial time, indicating an overcoming of mental barriers (Ohlsson, 1992). We performed corresponding analyses for the rebus puzzle task to evaluate similarities.

Third, we analyzed open response data regarding participants' reported solution strategies in each task. We did this to better understand the cognitive processes described when solving these problems - namely, in how insightful and analytic experiences differ.

Lastly, we tested whether our results replicated MacGregor and Cunningham's (2009), as we adopted their rebus puzzles for our study. This further serves to test the validity and reliability of their problem set.

2.3 Method

2.3.1 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 min. Eight participants were excluded from final analysis; five due to repeated participation, one due to missing data, one due to low-effort responses, and one due to being located outside of the United States.

2.3.2 Materials

For the joke completion task, we constructed three joke stems inspired by 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 total, each participant completed 18 joke completion trials. Normative data for each stem/subject combination is presented in Table B.2 in Appendix B.

For the rebus task, we adopted 24 problems from MacGregor and Cunningham (2009). These

are presented in Figure B.1 in Appendix B. We scored performance based on their suggested solutions. Normative data from these problems is presented in Table B.3 in Appendix B.

2.3.3 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 s. 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 labeled 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 instances of the word "SECRET" stacked vertically, with the top word circled. They were also given the target solution: "Top Secret." Each rebus puzzle trial lasted 60 s. 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.

Lastly, participants gave open response descriptions of how they knew they had experienced insight and reported any specific strategies they had used when solving joke completion and rebus problems.

2.4 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 ($M = 0.474$, $SD = 0.213$), $t(99) = 3.758$, $p < .001$, $BF_{10} = 68.94$. Participants also spent significantly more time in joke completion trials ($M = 37.25$, $SD = 14.177$) than in rebus puzzle trials ($M = 21.40$, $SD = 7.576$), $t(99) = 11.94$, $p < .001$, $BF_{10} > 100$. 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 demarcating 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 2.1. 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.

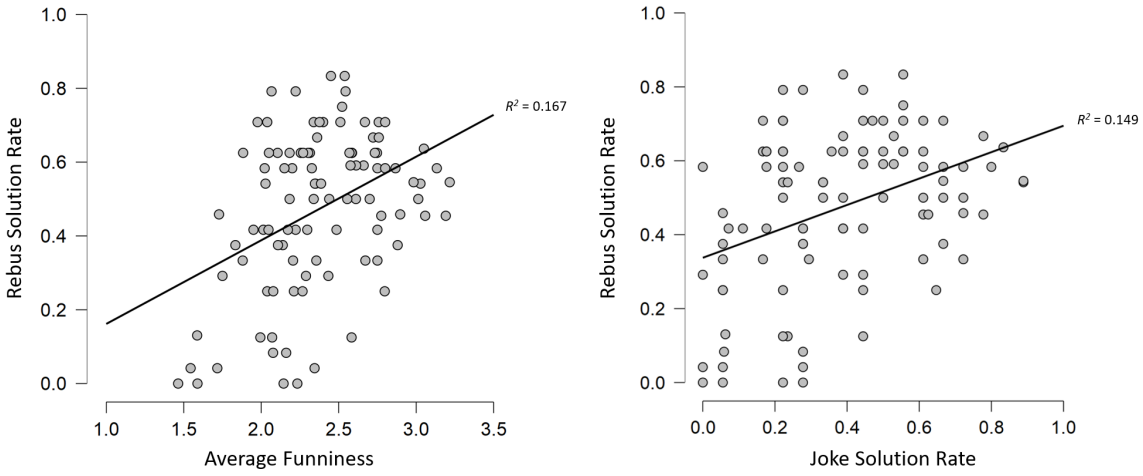

Figure 2.1: *Average funniness (left) and solution rate (right) in joke completion problems by solution rate in rebus puzzles.*

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 (n = 771, or 32.1% of all cases) and 5 (n = 566, or 23.6% of all cases), we saw the opposite trend for joke completion, with the *lowest* frequencies at 1 (n = 314, or 17.4%) and 5 (n = 127, or 7.1%), 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 2.2.

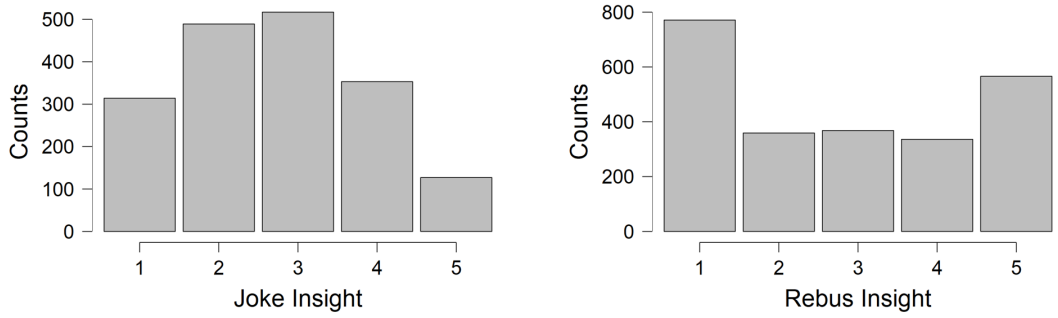

Figure 2.2: *Frequency of reported insight in joke completion problems (left) and rebus puzzles (right).*

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, $r_s = 0.218$, $p < .001$, $BF_{10} > 100$, 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 insight increased with successful performance on rebus puzzles. Similarly, a paired samples t-test revealed that there was a higher degree of reported insight for correct rebus solutions ($n = 1130$, $M = 3.843$, $SD = 1.262$) than for incorrect rebus solutions ($n = 1270$, $M = 1.909$, $SD = 1.232$), $t(1129) = 37.84$, $p < .001$, $BF_{10} > 100$.

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 s 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, $p < .001$, $BF_{10} > 100$. This indicates that both performance and insight in joke completion problems decreased as trial time elapsed. These trends are illustrated in Figure 2.3 and Figure 2.4 (left). Thus, responses submitted earlier in trials were more likely to be funny and to accompany greater feelings of insight.

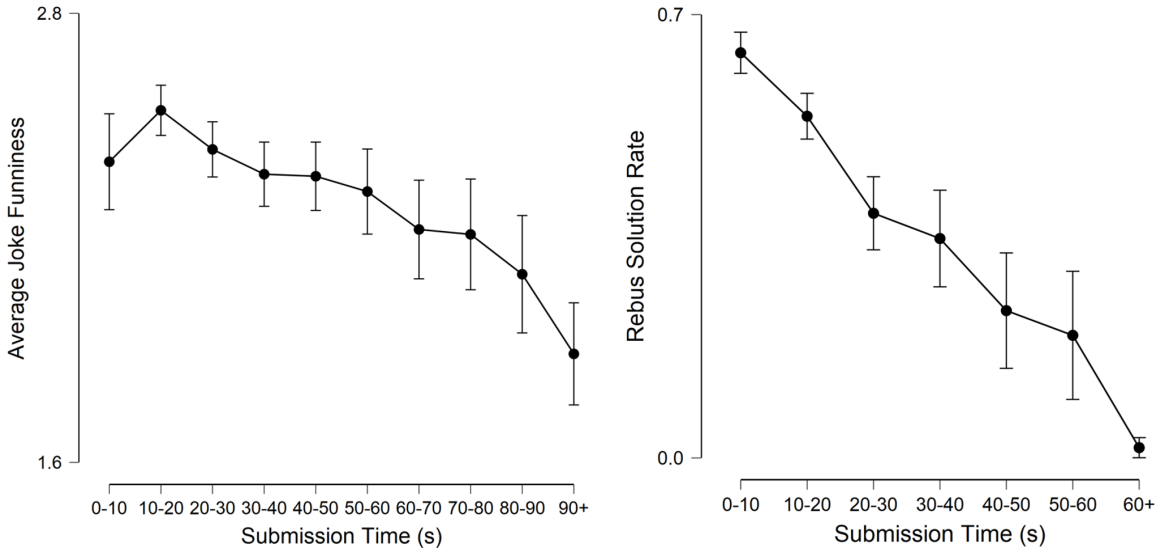

Figure 2.3: *Average joke funniness (left) and rebus solution rate (right) within each 10 s trial time interval (error bars denote 95% confidence intervals).*

This trend is once again reflected in the rebus task. Most responses to rebus puzzles were submitted within the first 10 s 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, $p < .001$, $BF_{10} > 100$, as well as rebus accuracy, r_s = -0.345, $p < .001$, $BF_{10} > 100$. This indicates that performance and reported insight in rebus puzzles decreased as trial time elapsed. This is illustrated in Figure 2.3 and Figure 2.4 (right).

Next, we analyzed participants' reported strategies used in each task. Twenty-six percent

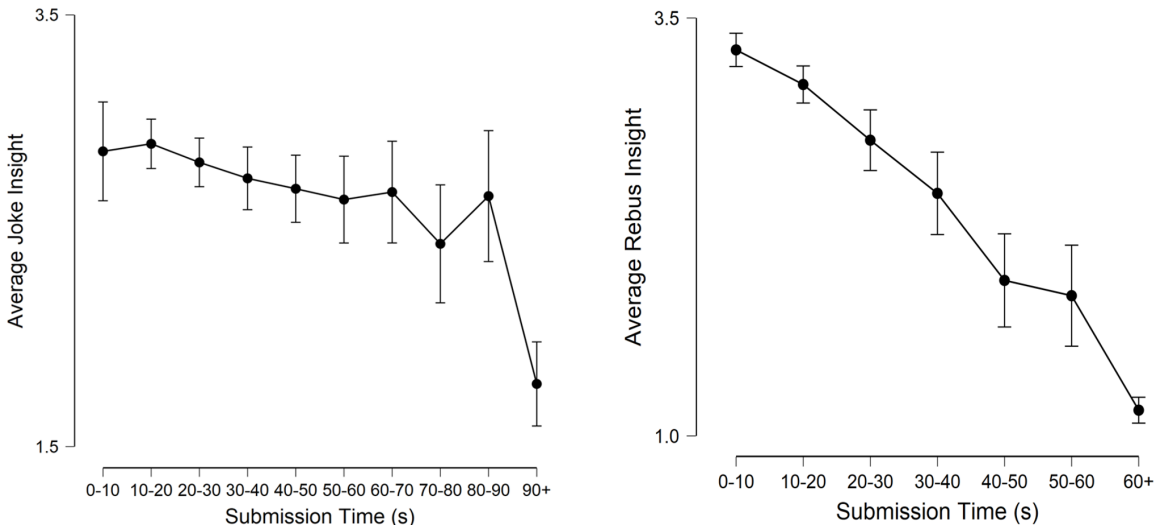

Figure 2.4: *Average insight ratings for joke completion problems (left) and rebus puzzles (right) within each 10 s trial time interval (error bars denote 95% confidence intervals).*

of participants reported using strategies to solve joke completion problems. The following are examples of typical responses: 1) "I was just starting to get the idea to look for words that were comically linked to the sentence"; 2) "I thought of words that were commonly associated with the given word. For example I thought of bark, leaves, trunk etc and then I thought about what double meanings any of those words could have that would also work for the joke scenario"; 3) "Coupling of words with know[n] associations that have symbolic meaning just as much as material meaning"; 4) "I tried to think of words in the semantic field"; 5) "I thought of every word that related to the subject. For instance, "clock - time, hands, tick-tock." Then it was easy to think of a punchline." An independent samples t-test revealed that there was no difference in average funniness between those who reported using strategies in joke completion (M = 2.469, SD = 0.346) and those who did not (M = 2.345, SD = 0.394), $t(98)$ = -1.416, p = 0.160, BF_{10} = 0.560. Similarly, there was no difference in magnitude of insight between self-reported strategists (M = 3.013, SD = 0.682) and non-strategists (M = 2.613, SD = 0.786), $t(98)$ = -2.308, p = 0.023, BF_{10} = 2.304. Lastly, there was no difference in estimated funniness between self-reported strategists (M = 2.357, SD = 0.727) and non-strategists (M = 2.135, SD = 0.737), $t(98)$ = -1.324, p = 0.189, BF_{10} =

0.503.

Twenty-seven percent of participants reported using strategies to solve the rebus puzzles. The following are examples of typical responses: 1) "[I] read the words, looked for visual clues (spacing, positioning, direction)"; 2) "I found that my brain would first be stumped. But then I told myself to RELAX, take a moment to sound out each item out loud and maybe it would come to me that way, and it did!"; 3) "I just tried to talk out the pictures to see if they were a recognizable phrase." There was no significant difference in solution rates between those who reported using strategies in rebus puzzles ($M = 0.582$, $SD = 0.204$) and those who did not ($M = 0.434$, $SD = 0.204$), $t(98) = -3.225$, $p = 0.002$, $BF_{10} = 19.01$. Similarly, there was no difference in magnitude of insight between self-reported strategists ($M = 3.066$, $SD = 0.728$) and non-strategists ($M = 2.728$, $SD = 0.734$), $t(98) = -2.048$, $p = 0.043$, $BF_{10} = 1.417$.

Lastly, we examined whether our results replicated MacGregor and Cunningham's (2009) using the same rebus puzzle set. As in their study, participants were scored based on their proportion of correct answers. We analyzed this data using a repeated-measures analysis of variance (ANOVA), with number of restructurings (1 or 2) and restructuring level (sub-word, word, or supra-word) as independent variables. Results indicate significant main effects for number of restructurings, $F(1, 99) = 132.33$, $p < .001$, $BF_{10} > 100$, and restructuring level, $F(2, 198) = 80.7$, $p < .001$, $BF_{10} > 100$, and a significant interaction, $F(2, 198) = 25.84$, $p < .001$, $BF_{10} > 100$. The nature of these effects is illustrated in Figure 2.5.

As in MacGregor and Cunningham (2009), there was an increase in solution rates across rebuses requiring two levels of restructuring (their percentages denoted with asterisks); from 23% (28%*) at the sub-word level, through 36% (53%*) at the word level, to 49% (68%*) at the supra-word level. There were significant differences between each pair of means ($p < .001$). Also like in the original study, there was not a corresponding linear increase in solution rates for rebuses requiring one level of restructuring, with 53% (75%*) at the sub-

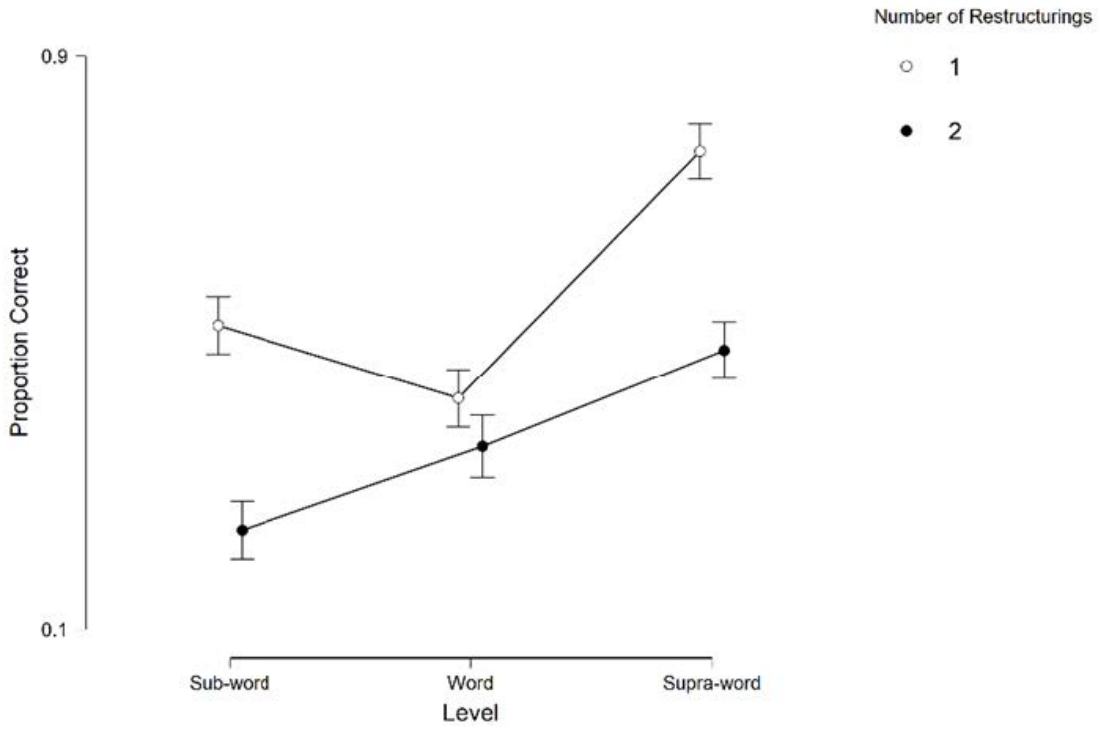

Figure 2.5: *Mean proportion of rebuses correct by restructuring level and number of restructurings (error bars denote 95% confidence intervals).*

word level, 42% (71%*) at the word level, and 77% (86%*) at the supra-word level. Our results replicated their findings that solution rates were higher for rebuses requiring one restructuring over two and higher for restructurings at the supra-word level than at lower levels on the parse tree. However, overall performance was notably worse in the current experiment.

2.5 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 for 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 the most 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 and 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 s for rebuses and 10-20 s 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 s 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 latency, 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 and Murphy, 2005; Paulewicz et al., 2007) and working memory (Chuderski and Jastrzebski, 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 and Salvi, 2020). This accuracy effect seems supported by the fact that confidence correlates strongly with reported insight experiences (Danek and Wiley, 2017; Webb et al., 2016). However, even when confidence is mentioned as a requisite of insight, this effect persists (Hedne et al., 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 and 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 and 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.

2.6 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.

Chapter 3

Perceptions of AI engaging in human expression

3.1 Introduction

As the capabilities of AI systems accelerate, they increasingly surpass us in several domains - even those once thought exclusively human. From making medical diagnoses (Dawes et al., 1989; Grove et al., 2000) to offering jail-or-release decisions (Kleinberg et al., 2018), algorithms can outperform human experts in a number of tasks. Given the capability of AI in such tasks, we as users should defer to their guidance to make optimal decisions. However, a growing body of evidence reveals certain biases in how users seek and weigh advice from algorithms compared to that of humans, resulting in sub-optimal decision making (Burton et al., 2020; Jussupow et al., 2020). Given the prevalence of these systems and their demonstrated ability to inform high-stakes decisions, it is critical to identify and eliminate such biases.

While people have expressed skepticism toward algorithms for decades (Meehl, 1954), em-

pirical work exploring these attitudes is quite recent. Such work has identified two primary forms of bias: *Algorithm aversion* (Castelo et al., 2019; Dietvorst et al., 2015) and *algorithm appreciation* (Logg et al., 2019). Algorithm aversion occurs when users prefer human to algorithmic judgement - even when the latter is proven superior. Conversely, algorithm appreciation occurs when users prefer algorithmic to human judgment. These seemingly incompatible findings raise a host of important questions: Why does aversion occur in some contexts, while appreciation occurs in others? How strong are these biases? Can they be overcome (and, if so, how)?

These differences may be accounted for, in part, by task characteristics (Castelo et al., 2019; Jussupow et al., 2020). For instance, users tend to prefer human judgement on subjective tasks (e.g., book, movie, and joke recommendations: Sinha and Swearingen, 2001; Yeomans et al., 2019), while deferring to algorithmic judgement on objective tasks (e.g., logic problems: Dijkstra et al., 1998; Logg, 2017). However, it should be noted that users *have* demonstrated aversion to decision aids even when they are perfectly accurate in objective judgment (e.g., target detection: Beck et al., 2005). This suggests that users tend to discount advice from AI when a task invokes domains believed to be inextricably human, such as personal taste, intuition, and experience. We thus tested the hypothesis that these conditions elicit aversion through a task invoking a fundamentally subjective and human expression: Humor. Specifically, telling jokes.

Beyond meeting the criteria of subjectivity, this task is relevant for a number of reasons. First, humor is a universal phenomenon with which we all have experience and possess individual taste regarding (Martin and Ford, 2018). Accordingly, people are able to make many short qualitative assessments of jokes in a single experimental session. Second, this topic is of interest to AI researchers, as developing embodied humor represents a critical challenge in naturalistic AI development (Binsted et al., 2006). Third, humorous virtual agents/AI have been shown to facilitate effective human-computer interaction (Nijholt et al.,

2017). One notable example is Morkes et al. (1999), who found that participants report more similarity to and cooperation with humorous agents when completing a task. Lastly, it has been suggested that aversion occurs in part due to the belief that algorithms should perform with near-perfect accuracy (Dzindolet et al., 2002). This is presumably why users disproportionately punish mistakes when they are committed by an algorithm, compared to a human (Dietvorst et al., 2015). However, it is unclear if users will maintain this expectation in tasks lacking ground truth, such as joke-telling.

Indeed, individual and societal beliefs regarding AI's comparative ability to produce humor remain largely unknown. The work conducted in this area has yielded interesting - if incomplete - results. For example, Tay et al. (2016) found that people perceive jokes as funnier when they are told by human actors than when they are told by robots, but only when joke content is non-disparaging. Interestingly, they found that people express less disgust toward disparaging jokes when delivered by robot actors. Further supporting the importance of content and context in robot-delivered humor, Stoll et al. (2018) found that humor delivered by a robot is perceived as less appropriate in conflict mediation than when delivered by a fellow human. It should be noted, however, that the physical embodiment of agents in these two studies almost certainly affected reception of their attempts at humor. Thus, the question largely remains: Do people respond differently to humor when it is believed to have been created by an AI, versus a human?

3.2 Results

Across two experiments, we address the following question: *Will people rate jokes as less funny if they believe an AI created them?* According to evidence showing that task subjectivity facilitates aversion, we hypothesized that they will. We tackled this question in two ways. In Experiment 1, we had participants rate the funniness of jokes and guess their

likeliest source - a human or AI. We left these jokes' actual sources ambiguous, forcing participants to make their own attributions. This allowed us to evaluate systematic differences in attribution between jokes considered low and high-quality. In Experiment 2, a new set of participants rated these same jokes. However, jokes were now explicitly labeled as either human or AI-created. This allowed us to assess potential differences in evaluations when jokes' (purported) sources were transparent. If aversion is present, we should expect a systematic downgrading of jokes believed to be AI-created in each experiment, regardless of their actual source.

3.2.1 Experiment 1

In Experiment 1, participants were given a randomized sequence of jokes. Participants were not informed of these jokes' actual sources. For each joke, they were asked to rate its funniness and to guess whether it was more likely created by a human or AI. Further details regarding the procedure can be found in the "Method" section below. There are two main results, as shown in Table 3.1 and Figure 3.1. First, participants guess that the funniest jokes were created by humans and that the least funny jokes were created by AI - regardless of their actual source. That is, funniness ratings increase as a function of how definitively human in origin jokes are believed to be. Second, actual human-created jokes (N = 2580, M = 2.609, SD = 1.250) are rated funnier than actual AI-created jokes (N = 860, M = 1.416, SD = 0.829) across all guesses ($t(2227.648)$ = -31.806, $p < 0.001$, $BF_{10} > 100$). These findings suggest that, in accordance with extant findings, humans prefer the output of other humans over AI in highly subjective tasks.

How accurate are participants at guessing jokes' actual sources? More actual human-created jokes are correctly attributed to "probably" or "definitely" being human-created (n = 1743) than to being AI-created (n = 837), while more actual AI-created jokes are correctly at-

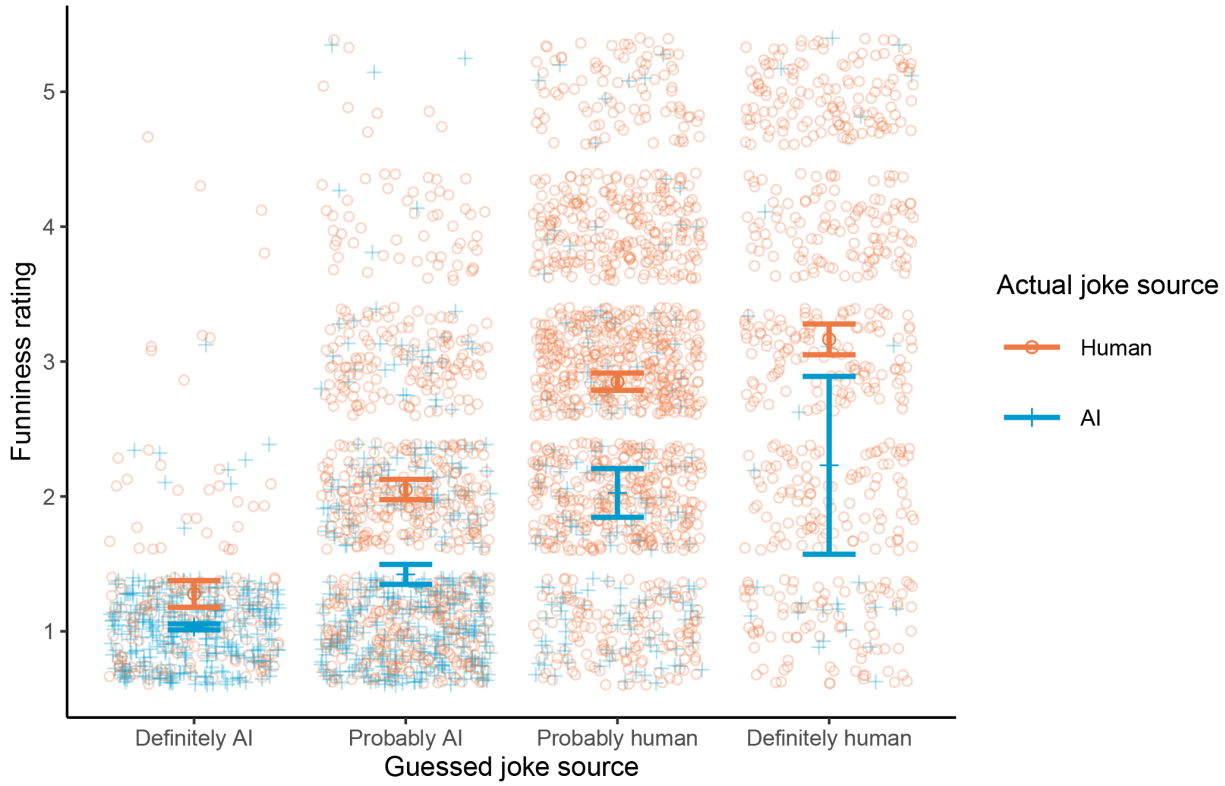

Figure 3.1: *Ratings for human and AI-created jokes across their guessed sources.* Error bars indicate 95% confidence intervals in rated funniness across actual joke sources (human or AI). Points represent an individual joke evaluation (rating and guessed/actual source).

tributed to being "probably" or "definitely" AI-created ($n = 682$) than to being human-created ($n = 178$). This indicates that participants are largely correct in their guesses.

We believe participants adhere to a reasonable heuristic here: When measuring the quality of jokes both by Reddit upvotes and participant ratings, the funniest jokes *do* tend to be human-created. This rating behavior may reflect bias, but it also reveals accurate perceptions of the AI's ability, as demonstrated by the high degree of accuracy in guesses. It is noteworthy, however, that many low-quality, human-created jokes are attributed to the AI (49% of jokes rated less than 3). This suggests that, even though low-quality jokes *are* largely AI-created, there is a tendency to attribute even low-quality human jokes to AI, further supporting the aversion hypothesis.

Table 3.1: *Funniness ratings across actual and guessed joke sources*

Actual Source	Guessed Source	Mean	SD	N
AI	Definitely Human	2.231	1.632	26
	Probably Human	2.026	1.127	152
	Probably AI	1.422	0.732	379
	Definitely AI	1.033	0.197	303
Human	Definitely Human	3.164	1.390	572
	Probably Human	2.851	1.111	1171
	Probably AI	2.051	0.992	664
	Definitely AI	1.277	0.659	173

In sum, the findings from Experiment 1 demonstrate that when joke source is not provided, participants tend to attribute high-quality jokes to humans and low-quality jokes to AI. Further, while these attributions are quite accurate, evidence suggests a bias to attribute low-quality, human-created jokes to AI. These results provide a baseline for attitudes regarding AI's ability to create jokes compared to humans and suggest the presence of algorithm aversion. However, will these attitudes persist when joke sources *are* provided? We tackled this question in Experiment 2.

3.2.2 Experiment 2

In Experiment 2, a new set of participants rated the funniness of the same jokes from Experiment 1, again presented in randomized order. However, each joke was now explicitly framed (i.e., labeled) as being either human or AI-created. The framing for each human-created joke was counterbalanced such that they were all equally-represented as human and AI-created. All AI-created jokes were accurately framed to improve the feasibility of the purported AI's output.

The main finding is that there is no difference in funniness ratings for human-created jokes alternatively framed as human ($N = 2490$, $M = 2.828$, $SD = 1.279$) or AI-created ($N =$

2490, $M = 2.765$, $SD = 1.274$) ($t(4978) = -1.732$, $p = 0.083$, $BF_{10} = 0.142$), as shown in the left two data columns of Figure 3.2. This indicates that, regardless of framing, participants rate human-created jokes equally. (Note that the Bayes factor shows not only that there is no significant difference, but that there is evidence that the funniness ratings are essentially the *same* between framings.) Critically, these results suggest that aversion is *not* present when participants are told the purported source of each joke. This result is inconsistent with the aversion hypothesis, as participants do not systematically downgrade jokes strictly because they are told that an AI created them.

In addition to this key finding, we observe other interesting results. First, human-created and framed jokes are rated significantly higher than their correctly AI-framed counterparts ($N = 1660$, $M = 1.421$, $SD = 0.829$) ($t(4145.024) = -42.987$, $p < 0.001$, $BF_{10} > 100$). Furthermore, human-created but AI-framed jokes are also rated higher than their correctly AI-framed counterparts ($t(4145.756) = -41.164$, $p < 0.001$, $BF_{10} > 100$). These results reinforce the fact that jokes are rated consistently based on quality, not purported origin.

Was our deception effective? Most participants ($n = 68$; 81.928%) report that they did *not* spot our deception. When asked to describe when and how they detected it, those who did cite the experimental design (e.g., "I don't think there's any reason to indicate "AI generated" or "human generated" unless you are looking for bias in responses to the jokes."), previous exposure to jokes (e.g., "I had heard some of the jokes that the AI supposedly wrote, years ago."), the disparity between misrepresented and accurately-framed AI joke quality (e.g., "When I noticed that some of the AI jokes were obviously more coherent than others."), and/or intuition about AI joke quality (e.g., "I just figured that AIs aren't making up jokes especially really bad ones."). Our main result remains when these individuals are excluded from analysis. That is, for individuals who reported being deceived, there is still no difference in ratings between human-labeled ($N = 2040$, $M = 2.756$, $SD = 1.269$) and AI-labeled ($N = 2040$, $M = 2.691$, $SD = 1.261$) human-created jokes ($t(4078) = -1.646$, $p = 0.100$, BF_{10}

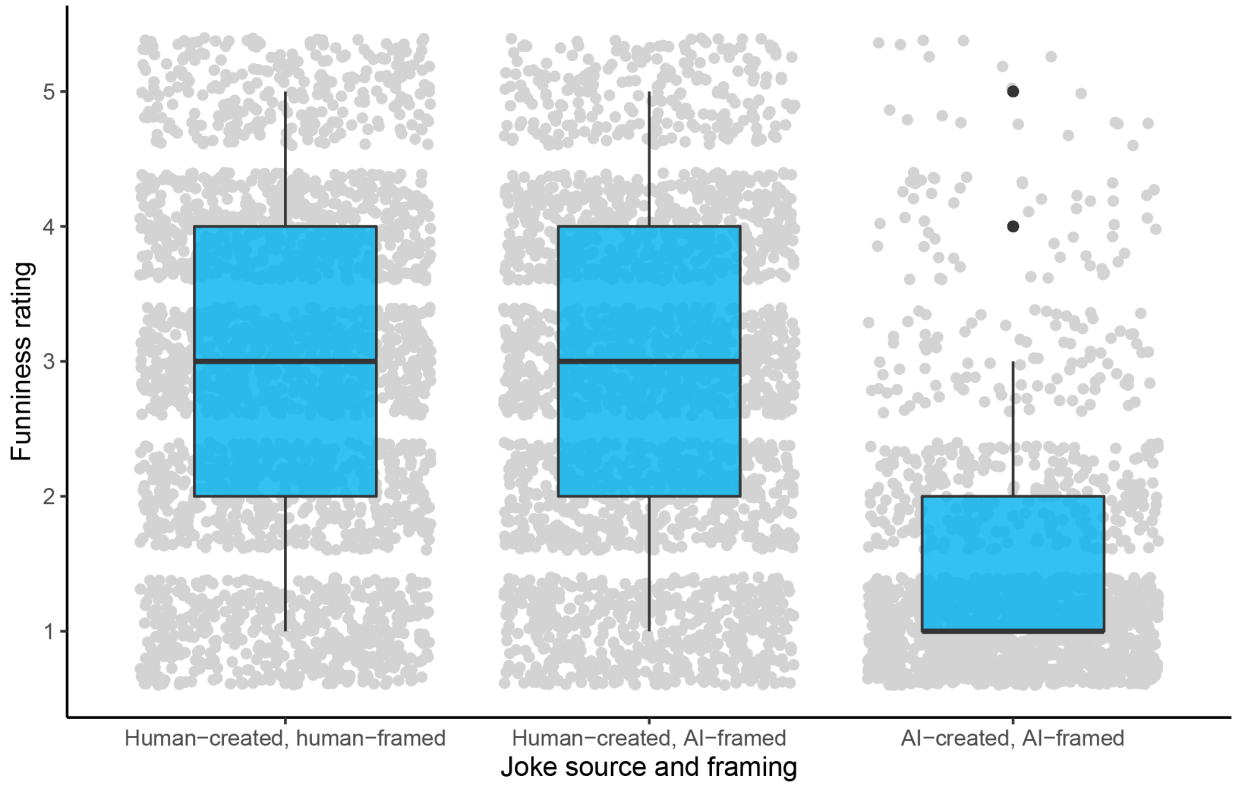

Figure 3.2: *Funniness ratings for jokes across their actual source and framing.* Dots indicate an individual joke rating under each framing.

= 0.136). Thus, the effectiveness of our cover story suggests a willingness to believe in AI's ability to produce human-quality jokes.

Lastly, we compared results between experiments to verify consistency in joke ratings and to identify potential differences in rating trends. Mean joke funniness is higher in Experiment 2 than in Experiment 1 ($t(10078) = 5.175$, $p < .001$, $BF_{10} > 100$). However, only four jokes vary in pairwise magnitude across experiments, suggesting overall reliability in joke ratings.

Overall, the findings in Experiment 2 challenge the hypothesis that there is bias against human-created but AI-framed jokes. Thus, unlike in Experiment 1, there is no indication that algorithm aversion is present.

3.3 Discussion

Across two experiments, we tested the hypothesis that task subjectivity invites algorithm aversion by examining whether people systematically downgrade ratings for jokes believed to be AI-created. The results from Experiment 1 show that people rate jokes guessed to be AI-created more harshly when no source is provided. Furthermore, they frequently attribute low-quality, human-created jokes to AI. This suggests that people may hold latent biases regarding AI versus human-created jokes when there is uncertainty concerning their source – adhering to a heuristic that better jokes are typically a human's and worse jokes are typically an AI's. This supports findings that users favor the work of other humans over AI in subjective domains (Sinha and Swearingen, 2001; Yeomans et al., 2019). Further, this echoes previous findings wherein AI is only attributed responsibility or considered agentic when there is an unfavorable outcome (Hohenstein and Jung, 2020). However, Experiment 2 shows that such biases are absent when joke sources *are* provided, as reflected in similar ratings between human-created jokes framed as either human or AI-created. This suggests that if there are aversions toward AI in this domain, they are weak and easily overcome when presented with counterevidence (i.e., good jokes framed as AI-created). These results contribute to findings that user attitudes toward AI are malleable, even when supposed AI attempt feats believed the province of humans (Castelo et al., 2019). It is once again worth noting that the vast majority of participants in Experiment 2 believed our deception, implying a willingness to accept the ability of current AI systems to produce compelling, human-level jokes (when framed appropriately).

It would be interesting to further explore societal attitudes toward AI that perform other creative work, such as art or poetry. While AI have demonstrated degrees of proficiency in these realms, it remains unclear how they will be broadly received. For instance, findings suggest that the way we talk about AI - as a tool or agent - affects how users allocate credit for machine-generated artwork (Epstein et al., 2020). To understand human beliefs (and

biases) regarding AI capability, it will be necessary to explore these domains.

There are a few limitations to our work. The first is that our findings are domain-specific by design. There are many factors believed to promote algorithm bias (Jussupow et al., 2020). For example, there is evidence that experts are more likely to discount the quality of algorithmic advice than non-experts (Arkes et al., 1986; Gaube et al., 2021; Highhouse, 2008; Logg, 2017). Future work may consider recruiting domain experts to assess potential differences in expressed bias (such as AI researchers and/or professional comedians, in the case of joke appraisal).

An influential finding is that aversion arises due to increased user sensitivity to AI mistakes - especially at the outset of tasks (Dietvorst et al., 2015). Users are more likely to overlook mistakes committed by themselves or another human compared to an algorithm. Since our jokes were randomly presented, we cannot assess if such order effects are present in our study (a "mistake" in this case being a poor-quality joke or non-joke). Subsequent work should evaluate potential order effects to see if this aversion is replicated.

Subsequent work may also categorize these jokes according to their content to see if it affects perceived funniness and source ascription. Indeed, people have shown differential ratings for computer-delivered jokes based on the appropriateness of their content (Tay et al., 2016). Further, the admittedly impoverished nature of the present task (rating jokes on a computer screen) does not capture the many nuances and complexities of humorous communication. To further assess human receptivity toward and discriminability regarding AI humor, various other modes of humor and experimental designs may be used in future work (e.g., employing joking conversational agents) (Dybala et al., 2009; Sjöbergh and Araki, 2009). Such approaches can assess the external validity of this and related work.

Lastly, it should also be noted that the GPT-2 language model from which the actual AI-created jokes were taken has been succeeded by the GPT-3 (Brown et al., 2020), which was

introduced after the completion of our study. The GPT-3 is reported to outperform its predecessor in naturalistic language, so using jokes from this newer model would be useful in assessing both the technical capability of and perceptions toward more advanced creative language systems.

3.4 Conclusion

As AI systems advance, so do our attitudes toward them (Bhattacherjee and Premkumar, 2004). Indeed, our so-called "theory of machine" - our beliefs regarding what algorithms are capable of and how - seems to evolve with the technology itself (Logg, 2017). More and more, AI challenge the delineation of human and machine - often beating us at our own game. This can be in critical tasks, such as making medical diagnoses, or in more casual pursuits, such as making jokes. Our results demonstrate that our beliefs may not be as fixed as once suggested and that we are open to AI occupying spaces once thought our own. Nonetheless, much work remains if we are to fully understand how and why biases toward AI occur.

3.5 Method

Participants

Participants were recruited through Amazon Mechanical Turk. Forty-three participants (n = 20 female) were recruited for Experiment 1 and 83 participants (n = 36 female) were divided between two counterbalanced conditions in Experiment 2. To be eligible for the our study, participants were required to meet the following criteria: 1) Have greater than or equal to 1,000 Human Intelligence Tasks (HITs) approved; 2) Have greater than or equal to

98% HIT approval rate for all requesters' HITs; 3) Be located in the United States; 4) Be fluent in English and; 5) Be 18-years-old or older. One participant declined to have their data included in the final analysis.

All participants provided informed consent before taking part in our study. Furthermore, they were all debriefed about its true nature following its conclusion, including all deception. Following this, they once again consented to allow their data to be used in the final analysis. The University of California, Irvine Institutional Review Board approved this research, which was conducted according to its guidelines.

3.5.1 Stimuli

Sixty items were adopted from a large database of jokes (N = 194,554) scraped from Reddit (Pungas, 2017). This set contains all jokes submitted to the subreddit r/jokes as of February 13, 2017. Jokes were curated by the first author to exclude those which were deemed potentially sexist, racist, or otherwise offensive. We also excluded jokes specific to a given time period or event (e.g., the 2016 US election) to avoid contextual dependencies.

To improve the feasibility of our cover story and provide a baseline for comparison, we inserted 20 actual, machine-produced jokes adopted from the subreddit r/SubSimulatorGPT2. These jokes were trained using a GPT-2 language model (Radford et al., 2019) based on submissions in r/jokes.

A small collection of sample jokes from our study is presented in Table C.1 in Appendix C.

3.5.2 Procedure

At the start of the experiment, participants were directed to read the following text: "The purpose of this study is to test the quality of a new artificial intelligence (AI) joke engine, JOSH ("Joke Ontology and System of Humor"). JOSH is being developed by researchers and uses state-of-the-art deep learning algorithms to construct jokes. However, JOSH is still in the early stages of development and we need your feedback to gauge its effectiveness and to help identify ways to improve it. To do this, we would like you to rate jokes created by JOSH according to how funny you find them. Lastly, we are interested in evaluating how well JOSH's jokes compare to jokes made by actual humans." Here, the script diverged between experiments. For Experiment 1, participants were provided the following: "Following each joke, you will be asked to guess if it was created by JOSH ("AI") or by a person ("HUMAN")." For Experiment 2, participants were provided the following: "Before each joke, you will be told if it was created by JOSH ("AI-GENERATED") or by a person ("HUMAN-GENERATED")." This designation was in orange or blue boldface text above each joke. The jokes actually created by humans were alternatively framed as human or AI-created across two counterbalanced conditions, such that both framings were equally represented for each joke. The jokes actually created by AI were always accurately framed as AI-created.

Participants rated each randomly-presented joke based on its perceived funniness (0 = "not funny at all"; 5 = "very funny"). In Experiment 1, participants guessed each joke's most likely source ("Definitely" or "Probably AI"; "Definitely" or "Probably Human"). Participants were debriefed regarding the true nature of our study, including the deception. They were then asked to indicate if they spotted our deception and, if so, to describe how and when they did to the best of their ability. Finally, they were asked whether they consented to their data being used in the final analysis.

3.5.3 Analyses

Principal data analysis was conducted using JASP (JASP Team, 2020). In addition to standard frequentist statistics, we also report Bayes factors (BFs). The advantages of using Bayesian inference, as well as suggested interpretations of results, are well-outlined in van Doorn et al. (2021).

3.5.4 Data Availability

Stimuli used in and datasets generated during the current study are available through Open Science Framework: `https://osf.io/bpt2d/?view_only=ec1fbeed317748d68ac3b4f170f1c7d9`

Chapter 4

Words of wisdom: Brief textual descriptions accurately convey domain knowledge

4.1 Introduction

Can just a few words reveal how much someone knows about a subject? Many situations in daily life require us to make quick assessments of others' knowledge. Perhaps we ask a friend for a wine recommendation at a new restaurant, or a stranger for directions in an unfamiliar town. How can we infer the extent and validity of their knowledge based solely on the language that they use - especially if we lack their lexicon? Using our wine example: Say your friend recommends an "attractive Zinfandel with a beefy character and a complex chewiness, complimented by its dusty finish"? The specificity of these terms may bolster their authority, expose their ignorance, or simply baffle you.

To further illustrate this, consider Figure 4.1, which features example stimuli and responses

Figure 4.1: *Illustration of the informant discrimination task where the goal is to identify the most knowledgeable informant from the image descriptions from a pair of informants.* Panels (a)-(f) show different example image stimuli and pairs of informant descriptions. The most knowledgeable people of each pair are A, D, E, G, I, and L.

from our current study. In it are images depicting various domains. Below each image are two people's descriptions of the content depicted. Can you tell, based on those descriptions, who the more knowledgeable person is? Perhaps you know a lot about one of these domains, making this distinction easy. But what if you lack this knowledge? How can you tell who knows more? Do you find yourself using particular language cues to infer knowledgeability? If so, how accurate are these inferences? These are the questions that motivated this study.

How we make such inferences can be described broadly as *theory of mind*, or how we reason about the mental states of others based on their behavior (Premack and Woodruff, 1978). There is a wealth of research in this area, exploring issues from how we predict others' intentions (e.g., Luchkina et al., 2018; Meltzoff, 1995), to how we understand false beliefs

(e.g., Baron-Cohen et al., 1985; Gopnik and Astington, 1988; Wimmer and Perner, 1983). However, most of this research concerns how we reason about goal-directed behavior, not the extent of well-learned, crystallized knowledge. Further, scarce work in this area explores how people make judgements about other's domain-knowledge based on language alone. While some developmental studies have explored this topic (e.g., Harris et al., 2018; Landrum and Mills, 2015; Lutz and Keil, 2002), there are relatively few incorporating adults.

The current work tackles the challenge of understanding how we infer knowledge based on others' language by addressing several key questions. First, is there a relationship between a person's own domain knowledge and their ability to assess the same domain knowledge in others? In other words: Does it take an expert to accurately gauge expertise (or lack thereof)? Relatedly, are less knowledgeable people less adept at identifying incorrect information than experts? People's estimates of others' knowledge are often anchored to their own actual or perceived knowledge (Nickerson, 1999). This suggests that, when there is a large discrepancy in knowledge between two individuals, a miscalibration will result. Further, since people seem to exhibit poor metacognition when they lack competence in a domain (the so-called Dunning-Kruger effect; Kruger and Dunning, 1999), we ask if this miscalibration extends to inferences regarding others.

Second, does our ability to infer another person's knowledge increase with the textual information they provide? Previous research has offered support for both elaboration and precision in knowledge communication. For example, experts have been shown to use more relational terms, more nouns, and a greater words-per-sentence ratio than novices (Kim et al., 2011). However, it is unclear to what extent this is helpful when forming an accurate model of others' knowledge. It may be that we reach an "inferential plateau," in which more information is no longer beneficial - perhaps even detrimental. Alternatively, in collaborative formats experts have been shown to effectively adjust their language to compensate for discrepancies between their own knowledge and that of novices, favoring precision (Isaacs

and Clark, 1987).

We explore these questions across two experiments. In Experiment 1, participants complete a series of multiple-choice trivia questions spanning several categories - some of which they identify as being proficient in, others being randomly assigned. They also provide written descriptions detailing the contents of various images representing these categories. In Experiment 2, a new set of participants (henceforth, "evaluators") assess the relative knowledge (based on trivia score) exhibited by Experiment 1's participants (henceforth, "informants"). We did this by having evaluators make pairwise comparisons between two informants' image descriptions, selecting which informant they believe to be more knowledgeable. Furthermore, we varied the number of image descriptions displayed, assessing whether more information is helpful in estimating knowledgeably. We also tested whether linguistic properties, such as description length and number of proper nouns used, affect evaluators' ability to predict informants' knowledge. Lastly, we tested whether the factualness of written descriptions influences evaluators' decisions when making pairwise comparisons regarding informants' knowledge.

4.2 Method

4.2.1 Participants

One hundred participants (i.e., informants) were recruited for Experiment 1 and 160 participants (i.e., evaluators) were recruited for Experiment 2, all through Amazon Mechanical Turk (MTurk). To be eligible for our study, they were required to hold Masters-level status on MTurk, be United States residents, and be at least 18-years-old.

4.2.2 Materials

Nine hundred multiple-choice questions were adopted from a large corpus of trivia questions provided by a trivia question publisher (The Question Co.). These questions span 30 categories (e.g., Video Games, Biology, World Cup). Each question has four alternative, fixed-choice responses.

Three hundred sixty images were compiled by the first author for image description and informant discrimination trials. There are 12 images for each of the above-mentioned categories. Figure 4.1 demonstrates some of these images and informant descriptions.

4.2.3 Procedure

Informants were presented the list of categories and instructed to select two in which they felt the most knowledgeable. Two additional categories were then randomly assigned. This was done to increase individual differences in knowledgeability scores. Throughout Experiment 1, informants completed two different tasks: Describing images and answering multiple-choice trivia questions.

For image description trials, six images were simultaneously presented on-screen with open response fields below each one. Informants were instructed to provide multiple-word descriptions for each of the images, revealing whatever specific knowledge they had for the content depicted therein. If they lacked specific knowledge, they were asked to describe the image in general terms. Informants were explicitly discouraged from using external sources to look up answers (e.g., Wikipedia). If they deviated from the page (e.g., by opening another tab), they were given a warning. After three warnings, they were removed from the study. There were eight such trials of this, with two for each category. In total, each informant provided 48 image descriptions.

For the multiple-choice trivia portion of our study, informants completed two blocks of 15 questions per category. They were given 20 seconds to complete each question. A countdown timer on the upper-right corner of the screen indicated this, turning red when 5 seconds remained. If no response was selected by this time, the choice buttons were locked and the informant was forced to proceed to the next question. Additionally, informants provided a confidence rating for each response (25% ("Guessing"), 40%, 55%, 70%, 85%, 100% ("Absolutely Certain")). No feedback was provided. At the end of each block, informants were asked to estimate how many of these 15 questions they answered correctly. In total, each informant completed 120 multiple-choice questions.

In sum, for each category, each informant produced 12 image descriptions, 30 multiple-choice answers, and two performance estimations.

Half of Experiment 2's procedure followed that of Experiment 1, with evaluators selecting and being assigned categories to complete multiple-choice questions from. The key difference is in the image description portion of the experiment. Instead of providing descriptions of images themselves, evaluators assessed informants' knowledge based on their image descriptions from Experiment 1.

Evaluators were presented with images and corresponding descriptions provided by pairs of informants. For a particular pair of informants, evaluators predicted which of the two informants was likely to perform better on the multiple-choice portion of the category depicted, with three confidence ratings for each informant ("Very Confident," "Somewhat Confident," and "Guessing"). The pairs of informants were randomly selected from Experiment 1 with the constraint that the informants had different accuracy scores on the multiple-choice questions of the category. There were eight prediction trials per category, with each trial involving a different pair of informants. The informants were labeled with distinct letter pairs (A-B, C-D, etc.). The number of images and corresponding descriptions presented varied (1, 3, 6, or 12) across trials. Overall, each evaluator answered 32 prediction trials.

4.2.4 Scoring of Image Descriptions

Image descriptions were manually scored for veracity of facts by the first author. Facts were scored as *correct* when they specifically and accurately represented the content of the image. For example, Informant D in Figure 4.1 correctly identifies the author as Stephen King and his primary genre as horror. Conversely, facts were scored as *incorrect* when, while specific, they inaccurately represented the content of the image. For example, Informant C for the same image incorrectly attributes *Jurassic Park* to King, while it was actually written by Michael Crichton. Lastly, facts received no score when they were generic descriptions of the content depicted or reiterated information already stated in the prompt or image, itself. Note that descriptions may contain several correct or incorrect facts, or no facts at all.

4.2.5 Data Analysis

For all analyses, we utilize Bayes factors (BFs) to determine the extent to which the observed data adjust our belief in the alternative and null hypotheses. Values of $3 < BF < 10$ and $BF > 10$ indicate moderate and strong evidence against the null hypothesis, respectively. Similarly, values of $1/10 < BF < 1/3$ and $BF < 1/10$ indicate moderate and strong evidence in favor of the null hypothesis, respectively (Jeffreys, 1961; Rouder et al., 2009, 2012). In order to improve readability, BFs larger than 100 are reported as $BF > 100$. We use CI to refer to the 95% credible interval using Bayesian estimation methods.

For the Bayesian Pearson correlation and t-tests, we computed the BFs with the software package JASP (JASP Team, 2022) using the default priors that came with the software. For the logistic regression model, we applied Bayesian inference using a Markov chain Monte Carlo approach based on slice sampling. We ran the slice sampler with 8 chains with a burn-in of 1000 iterations and took 100 samples from each chain after each 20th iteration. Convergence of the sampler was tested using standard methods. The Bayes factors for the

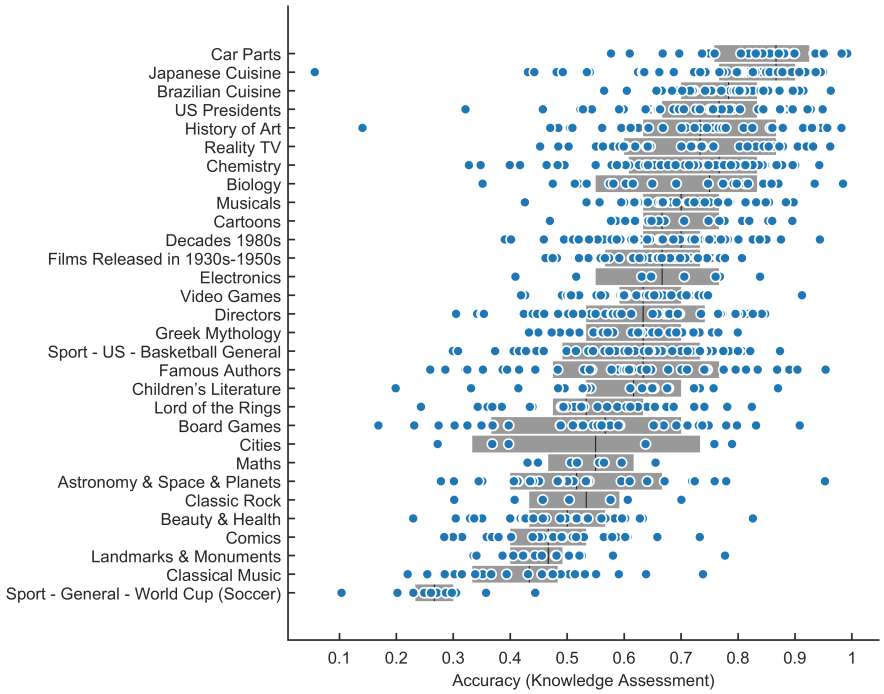

Figure 4.2: *Individual participant accuracy across knowledge assessment categories.* Gray bars show the 25%-75% quartiles. Results are combined across participants in Experiment 1 and 2.

logistic regression model was performed using the Savage Dickey method (Wagenmakers et al., 2010).

4.3 Results

4.3.1 Individual differences in knowledgeability

Informants in Experiment 1 and evaluators in Experiment 2 showed substantial individual differences in knowledgeability within each category (Figure 4.2). The mean accuracy difference between the worst and best participant in each category was .51 (average IQR is 0.19). In addition, participants were more accurate for categories that were self-selected than for categories that were randomly assigned ($M = .730$ vs. $M = .573$, $BF > 100$).

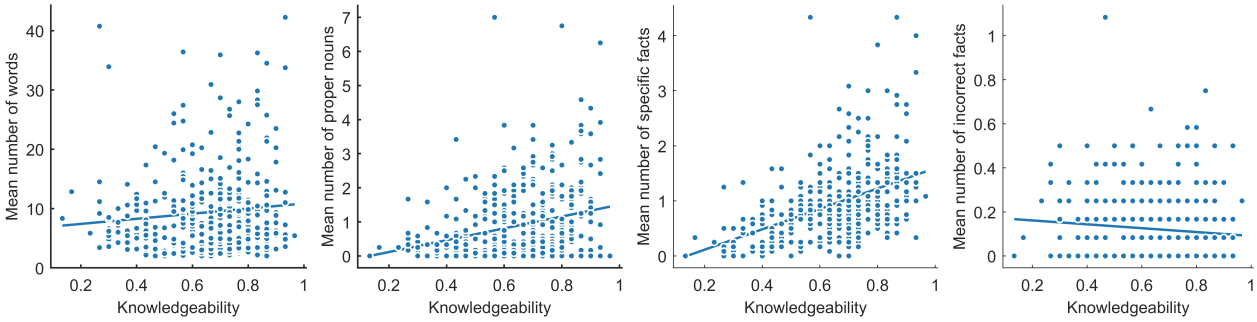

Figure 4.3: *Mean number of words, proper nouns, specific facts and incorrect (specific) facts in informant descriptions as function of informant knowledgeability.*

4.3.2 Knowledgeable informants mention more specific facts

For each informant, we analyzed the descriptions generated in the context of each category. Informants used a median of 7 words (IQR = 4 - 12) in their descriptions. While there was no evidence that more knowledgeable informants used more words (Pearson correlation $r =$ 0.103, $BF = .525$, CI = [0.005, .199]), they did use more proper nouns ($r = 0.271$, $BF > 100$, CI = [0.177, .358]), suggesting that knowledgeable informants used more specific language. To better assess the level of specificity in the descriptions, we scored each description for the number of specific (non-trivial) facts mentioned. Knowledgeable informants produced more specific facts ($r = 0.463$, $BF > 100$, CI = [0.381, .535]). For example, informants who scored above 90% accuracy produced more than three times the number of specific facts as informants below 50% accuracy ($M = 1.54$ vs. $M = .45$). The likelihood that any specific statement contained an incorrect fact was very low ($p = .11$). Therefore, the majority of specific statements were true facts. There was no evidence that knowledgeable informants produce fewer incorrect facts ($r =$ -.099, $BF = .432$, CI = [-0.194, 0.000]). Figure 4.3 shows these relationships.

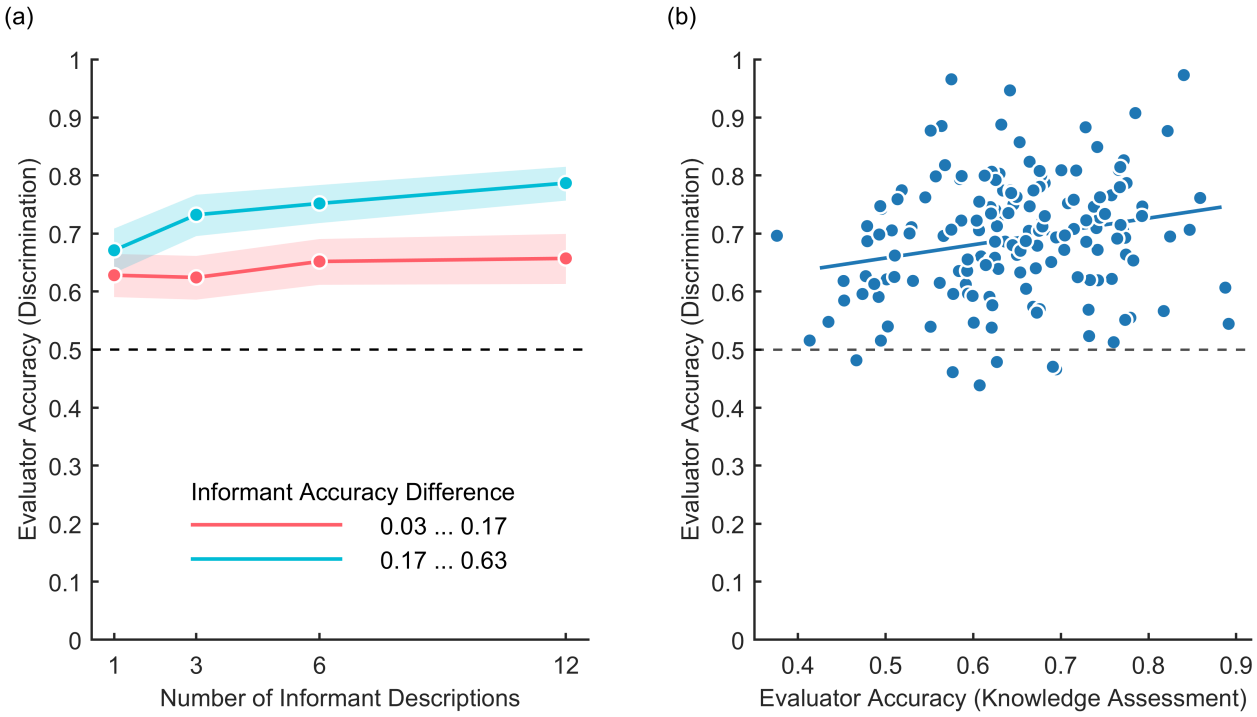

Figure 4.4: *Evaluator accuracy in the discrimination task.* (a) Mean evaluator accuracy for different number of informant descriptions and differences in informant knowledge scores, below (red) and above (blue) the median of differences. (b) Individual evaluator accuracy as a function of their knowledge score. Dashed lines represent chance performance

4.3.3 Factors influencing evaluator accuracy

When presented with image descriptions from pairs of informants, evaluators were generally above chance in determining the most knowledgeable informant. With a single image description, evaluation accuracy was 65%. Performance increased to 74% with 12 descriptions ($BF > 100$, paired sampled t-test). In addition, evaluators were more accurate when they had to discriminate between informants with larger differences in their knowledge scores. Figure 4.4a shows these relationships. There is moderate evidence that the more knowledgeable evaluators performed better on the discrimination task (Pearson $r = .226$, $BF = 5.9$, CI $= [.073, .365]$), but as illustrated in Figure 4.4b, there are substantial individual performance differences across the two tasks.

4.3.4 Level of specificity and factualness in descriptions influences evaluator choice

To investigate how evaluators determine their choice on the basis of the content of the image descriptions from a pair of informants A and B, we estimate a logistic regression model. The model includes as factors the difference in the total number of specific statements from A and B (regardless of whether they are correct or not), as well as the difference in the total number of incorrect statements from A and B. These totals are calculated across all the image descriptions provided by informants A and B on a particular trial. In addition, we also add an interaction effect with the knowledgeability of the evaluator to test if more knowledgeable evaluators are more sensitive to incorrect statements. The simplest model we estimate is the regression model:

$$p(\text{Choose } A) = f\left(w_0 + w_1(n_A - n_B) + w_2\theta(m_A - m_B)\right) \tag{4.1}$$

where f is the logistic function, $n_A - n_B$ is the differential in the total number of specific statements made by A and B, $m_A - m_B$ is the differential in the number of incorrect statements, θ is the knowledgeability of the evaluator, and w are model weights. We applied Bayesian inference to estimate model parameters and also performed a Bayesian model comparison with simpler models that excluded each individual term. The model comparison results show that there is evidence for the first factor involving w_1 as well as the interaction factor involving w_2, ($BF > 100$). The posterior mean of w_1 was positive (M = .36,CI = [.33, .39]) while the posterior mean of w_2 was negative (M = -.44, CI = [-0.51, -0.36]). To facilitate the interpretation of the estimated model, Figure 4.5 shows the model predictions as the number of correct and incorrect statements are varied *independently*. These results show that evaluators of all levels of knowledgeability tend to select informants with a higher number of correct statements. The less knowledgeable evaluators tend to treat incorrect statements

more similar to correct statements – increasing the number of incorrect statements for informant A makes these evaluators *more* likely to choose informant A. However, for the more knowledgeable evaluators, this effect is reversed but not completely. Overall, these results demonstrate that evaluators are sensitive to the number of specific statements and only the more knowledgeable evaluators tend to discount the effect of incorrect statements.

4.3.5 Comparing informativeness of verbal descriptions to prior performance

The results show that evaluators can select, with above chance accuracy, the more knowledgeable informant from a pair of informants on the basis of brief verbal descriptions. A natural question that arises is how informative this verbal information is relative to other sources of information about informant knowledgeability. One such source is the prior performance of the informant on other questions from the same general knowledge category. For example, if we are told that informants A and B previously have scored three and five questions correctly in a particular category, it is reasonable to predict that informant B will score better on future questions from the same category. We computed predictive accuracy based on knowing the outcomes from a given number prior questions. For one, two, three, and four prior questions, average predictive accuracy on future questions is at 59%, 65%, 70%, and 74% (averaged across random partitions of the questions). We compare these results to the predictive accuracy of verbal descriptions in Figure 4.4a. Evaluator accuracy on a single image description (65%) is about as good as making predictions based on knowing the score for two prior questions (on average). Similarly, the evaluator accuracy for 12 image descriptions (74%) corresponds to the level of performance that can be expected when knowing the score for four questions. Therefore, these results show that short verbal descriptions can be diagnostic and can be as informative as knowing how well an informant has scored previously on a few questions.

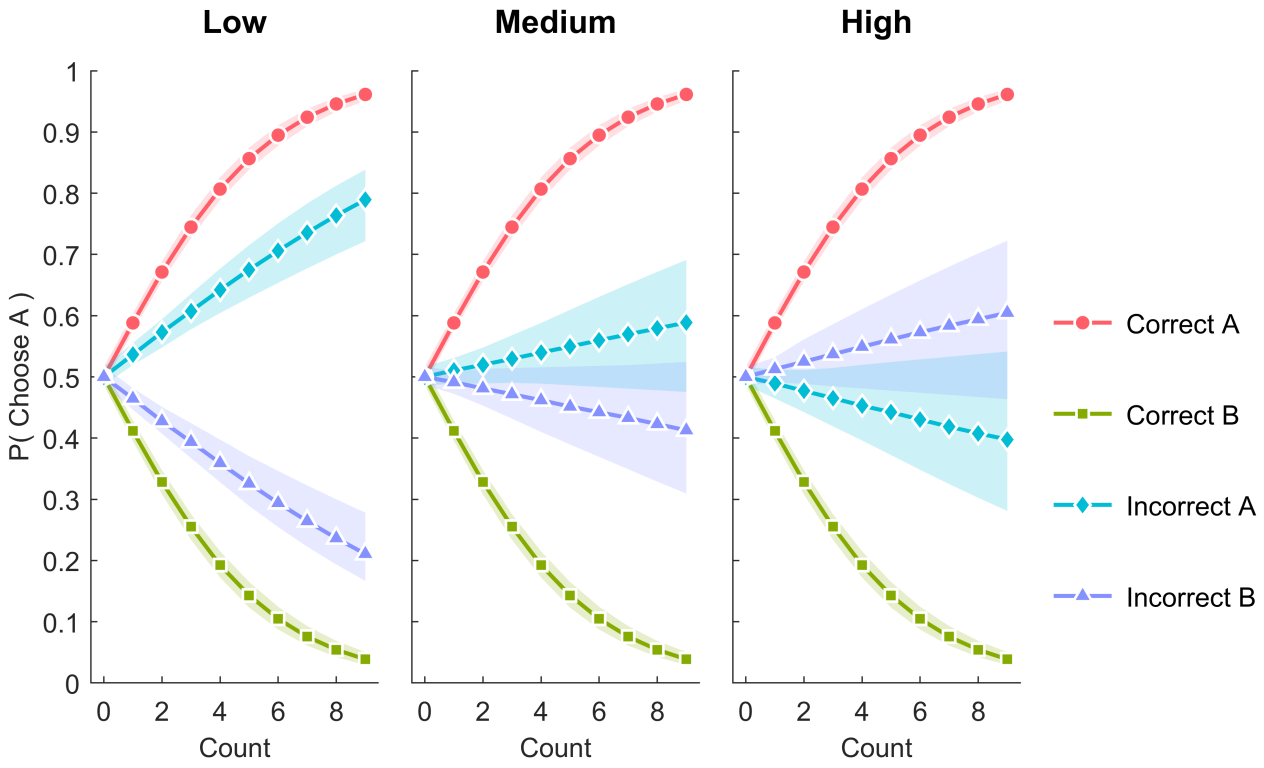

Figure 4.5: *Predicted effects of the independent effects of number of correct and incorrect statements from A and B on the probability of choosing A as the more knowledgeable informant.* Results are separated by the knowledgeability of the evaluator (Low = 0.5, Medium = 0.75, and High = 0.9 accuracy).

4.4 Discussion

Across two experiments, we addressed the fundamental question of whether a few words may provide accurate estimates of domain knowledge in others. There are several key results.

First, the most knowledgeable informants produce the most specific facts, with those performing above 90% accuracy in the trivia task producing more than three times as many specific facts as those who performed below 50% accuracy. Second, evaluators performed above chance (65%) at discerning which informants were most knowledgeable, even with only one image description available. This is striking, given that relatively so little information (around seven words, on average) facilitates such strong inferential accuracy. Moderate evidence suggests that more knowledgeable evaluators performed better at discriminating informant ability, though this is mediated by individual differences across tasks. Third, evaluators at all knowledge levels select informants who produce more specific information. However, less knowledgeable evaluators are less likely to discriminate correct from incorrect facts, while more knowledgeable evaluators are more sensitive to this difference. This suggests that people with less domain knowledge are more readily swayed by specific information, regardless of its ground truth. Lastly, verbal descriptions provide similar predictive accuracy regarding informant performance as knowing their scores on a few trivia questions.

One limitation of our study is that the base rate of incorrect facts was quite low. Future work should investigate if our findings persist when the presence of incorrect statements is matched with − or *exceeds* − their correct counterparts. This may be done through experimental manipulation or through naturalistic means. This approach should further reveal if specificity of statements is a useful heuristic in determining others' knowledge.

Our study offers an important methodological contribution in that it provides an alternative for measuring confidence/expertise, compared to oft-used ratings scales. Measuring the degree of specificity and elaboration provided in statements may better capture competence

than through self-report.

One future direction which we are currently exploring is using natural language processing techniques (e.g., Word2vec; Mikolov et al., 2013) to manipulate the semantic profile of statements. We may then administer statements to a new set of evaluators - some of which are from actual informants, and some of which are altered (machine-generated or manually-so) to replace proper nouns with semantically similar terms (e.g., replace "**Stephen King** is an **American** horror author who wrote *The Dark Tower*" with "**Dan Brown** is a **British** horror author who wrote *The Dark Tower*"). Thus, we can control statement length and vary the factualness and distance of terms, further parsing who can determine expertise in a more controlled environment.

Finally, a natural extension of this work is to examine how domain experts and novices differentially treat misinformation. This is an obvious and growing concern on social media platforms, such as on Twitter (e.g., Kouzy et al., 2020). As our results suggest that low-knowledge evaluators have a comparatively difficult time discriminating true and false statements when they are specific, it is worth exploring if and how this is observed in such real-life contexts. Researchers may also employ effortful deception, such as by having high-knowledge informants construct false statements and examining how effective others are at fact-checking them across the knowledge spectrum. Such approaches can help us better understand how false statements are appraised and are a potentially crucial step in combating the misinformation epidemic.

Bibliography

Amir, O., Biederman, I., Wang, Z., and Xu, X. (2015). 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.

Arkes, H. R., Dawes, R. M., and Christensen, C. (1986). Factors influencing the use of a decision rule in a probabilistic task. *Organizational Behavior and Human Decision Processes*, 37(1):93–110.

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

Ash, I. K. and Wiley, J. (2006). The nature of restructuring in insight: An individual-differences approach. *Psychonomic Bulletin & Review*, 13(1):66–73.

Attardo, S., Hempelmann, C. F., and Maio, S. D. (2002). Script oppositions and logical mechanisms: Modeling incongruities and their resolutions. *Humor: International Journal of Humor Research*, 15:3–46.

Attardo, S. and Raskin, V. (1991). Script theory revis(it)ed: Joke similarity and joke representation model. *Humor: International Journal of Humor Research*, 4:293–347.

Aziz-Zadeh, L., Kaplan, J. T., and Iacoboni, M. (2009). "Aha!": The neural correlates of verbal insight solutions. *Human Brain Mapping*, 30(3):908–916.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., and Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, 39(3):445–459.

Baron-Cohen, S., Leslie, A. M., and Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, 21(1):37–46.

Batchelder, W. H. and Alexander, G. E. (2012). Insight problem solving: A critical examination of the possibility of formal theory. *The Journal of Problem Solving*, 5(1):56–100.

Beck, H. P., Dzindolet, M. T., and Pierce, L. G. (2005). Take the advice of a decision aid: I'd rather be wrong! In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 49, pages 558–562. SAGE Publications Sage CA: Los Angeles, CA.

Bhattacherjee, A. and Premkumar, G. (2004). Understanding changes in belief and attitude toward information technology usage: A theoretical model and longitudinal test. *MIS Quarterly*, 28(2):229–254.

Binsted, K., Nijholt, A., Stock, O., Strapparava, C., Ritchie, G., Manurung, R., Pain, H., Waller, A., and O'Mara, D. (2006). Computational humor. *IEEE Intelligent Systems*, 21(2):59–69.

Blackmore, S. (2011). *Consciousness: An introduction*. Oxford University Press.

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

Bowden, E. M., Jung-Beeman, M., Fleck, J., and Kounios, J. (2005). New approaches to demystifying insight. *Trends in Cognitive Sciences*, 9(7):322–328.

Bower, A. H., Burton, A., Steyvers, M., and Batchelder, W. H. (2019). An insight into language: Investigating lexical and morphological effects in compound remote associate problem solving. In Goel, A. K., Seifert, C. M., and Freksa, C., editors, *Proceedings of the 41st Annual Conference of the Cognitive Science Society*, pages 166–173, Montreal, QB. Cognitive Science Society.

Bowers, K. S., Regehr, G., Balthazard, C., and Parker, K. (1990). Intuition in the context of discovery. *Cognitive Psychology*, 22(1):72–110.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

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

Burnham, C. A. and Davis, K. G. (1969). The nine-dot problem: Beyond perceptual organization. *Psychonomic Science*, 17(6):321–323.

Burton, J. W., Stein, M.-K., and Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, 33(2):220–239.

Cai, D. J., Mednick, S. A., Harrison, E. M., Kanady, J. C., and Mednick, S. C. (2009). REM, not incubation, improves creativity by priming associative networks. *Proceedings of the National Academy of Sciences*, 106(25):10130–10134.

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

Castelo, N., Bos, M. W., and Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, 56(5):809–825.

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

Chronicle, E. P., MacGregor, J. N., and Ormerod, T. C. (2004). What makes an insight problem? The roles of heuristics, goal conception, and solution recoding in knowledge-lean problems. *Journal of Experimental Psychology: Learning, memory, and cognition*, 30(1):14.

Chuderski, A. and Jastrzebski, 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.

Cunningham, J. B., MacGregor, J. N., Gibb, J., and 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.

Cushen, P. J. and Wiley, J. (2012). Cues to solution, restructuring patterns, and reports of insight in creative problem solving. *Consciousness and Cognition*, 21(3):1166–1175.

Danek, A. H., Fraps, T., von Müller, A., Grothe, B., and Öllinger, M. (2014). It's a kind of magic—what self-reports can reveal about the phenomenology of insight problem solving. *Frontiers in Psychology*, 5:1408.

Danek, A. H. and Salvi, C. (2020). Moment of truth: Why Aha! experiences are correct. *The Journal of Creative Behavior*, 54(2):484–486.

Danek, A. H. and 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.

Danek, A. H., Wiley, J., and Öllinger, M. (2016). Solving classical insight problems without aha! experience: 9 dot, 8 coin, and matchstick arithmetic problems. *The Journal of Problem Solving*, 9(1):4.

Davidson, J. E. (1995). The suddenness of insight. In Sternberg, R. J. and Davidson, J. E., editors, *The nature of insight*, pages 125–155. The MIT Press, Cambridge, MA.

Dawes, R. M., Faust, D., and Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, 243(4899):1668–1674.

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

Dietvorst, B. J., Simmons, J. P., and Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114–126.

Dijkstra, J. J., Liebrand, W. B., and Timminga, E. (1998). Persuasiveness of expert systems. *Behaviour & Information Technology*, 17(3):155–163.

Duncker, K. and Lees, L. S. (1945). On problem-solving. *Psychological Monographs*, 58(5):i–113.

Dybala, P., Ptaszynski, M., Rzepka, R., and Araki, K. (2009). Humoroids: Conversational agents that induce positive emotions with humor. In *AAMAS'09 Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems*, volume 2, pages 1171–1172. ACM.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., and Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors*, 44(1):79–94.

Epstein, Z., Levine, S., Rand, D. G., and Rahwan, I. (2020). Who gets credit for AI-generated art? *iScience*, 23(9):101515.

Fiorentino, R. and Poeppel, D. (2007). Compound words and structure in the lexicon. *Language and Cognitive processes*, 22(7):953–1000.

Fleck, J. I. and Weisberg, R. W. (2013). Insight versus analysis: Evidence for diverse methods in problem solving. *Journal of Cognitive Psychology*, 25(4):436–463.

Gaube, S., Suresh, H., Raue, M., Merritt, A., Berkowitz, S. J., Lermer, E., Coughlin, J. F., Guttag, J. V., Colak, E., and Ghassemi, M. (2021). Do as AI say: Susceptibility in deployment of clinical decision-aids. *NPJ Digital Medicine*, 4(1):1–8.

Gick, M. L. and Lockhart, R. S. (1995). Cognitive and affective components of insight. In Sternberg, R. J. and Davidson, J. E., editors, *The nature of insight*, pages 197–228. The MIT Press.

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

Gopnik, A. and Astington, J. W. (1988). Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. *Child Development*, 5(1):26–37.

Griffiths, T. L., Steyvers, M., and Firl, A. (2007). Google and the mind: Predicting fluency with pagerank. *Psychological Science*, 18(12):1069–1076.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., and Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, 12(1):19–30.

Gupta, N., Jang, Y., Mednick, S. C., and Huber, D. E. (2012). The road not taken: Creative solutions require avoidance of high-frequency responses. *Psychological Science*, 23(3):288–294.

Harris, P. L., Koenig, M. A., Corriveau, K. H., and Jaswal, V. K. (2018). Cognitive foundations of learning from testimony. *Annual Review of Psychology*, 69:251–273.

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

Highhouse, S. (2008). Stubborn reliance on intuition and subjectivity in employee selection. *Industrial and Organizational Psychology*, 1(3):333–342.

Hill, G. and Kemp, S. M. (2018). Uh-oh! what have we missed? A qualitative investigation into everyday insight experience. *The Journal of Creative Behavior*, 52(3):201–211.

Hohenstein, J. and Jung, M. (2020). Ai as a moral crumple zone: The effects of AI-mediated communication on attribution and trust. *Computers in Human Behavior*, 106:106190.

Inhoff, A. W., Starr, M. S., Solomon, M., and Placke, L. (2008). Eye movements during the reading of compound words and the influence of lexeme meaning. *Memory & Cognition*, 36(3):675–687.

Isaacs, E. A. and Clark, H. H. (1987). References in conversation between experts and novices. *Journal of Experimental Psychology: General*, 116(1):26–37.

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

JASP Team (2020). JASP (Version 0.14.1)[Computer software].

JASP Team (2022). JASP (Version 0.16.2)[Computer software].

Jeffreys, H. (1961). *The theory of probability* (3rd ed.). Oxford University Press.

Ji, H., Gagné, C. L., and Spalding, T. L. (2011). Benefits and costs of lexical decomposition and semantic integration during the processing of transparent and opaque english compounds. *Journal of Memory and Language*, 65(4):406–430.

Juhasz, B. J., Lai, Y.-H., and Woodcock, M. L. (2015). A database of 629 english compound words: ratings of familiarity, lexeme meaning dominance, semantic transparency, age of acquisition, imageability, and sensory experience. *Behavior Research Methods*, 47(4):1004–1019.

Juhasz, B. J. and Rayner, K. (2003). Investigating the effects of a set of intercorrelated variables on eye fixation durations in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(6):1312–1318.

Juhasz, B. J., White, S. J., Liversedge, S. P., and Rayner, K. (2008). Eye movements and the use of parafoveal word length information in reading. *Journal of Experimental Psychology: Human Perception and Performance*, 34(6):1560–1579.

Jung-Beeman, M., Bowden, E. M., Haberman, J., Frymiare, J. L., Arambel-Liu, S., Greenblatt, R., Reber, P. J., Kounios, J., and Dehaene, S. (2004). Neural activity when people solve verbal problems with insight. *PLoS Biology*, 2(4):e97.

Jussupow, E., Benbasat, I., and Heinzl, A. (2020). Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion. In *Proceedings of the 28th European Conference on Information Systems*.

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

Kim, K., Bae, J., Nho, M.-W., and Lee, C. H. (2011). How do experts and novices differ? Relation versus attribute and thinking versus feeling in language use. *Psychology of Aesthetics, Creativity, and the Arts*, 5(4):379.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2018). Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133(1):237–293.

Köhler, W. (1925). *The mentality of apes* (E. Winter, Trans.). Harcourt Brace.

Kotovsky, K., Hayes, J. R., and Simon, H. A. (1985). Why are some problems hard? evidence from tower of hanoi. *Cognitive Psychology*, 17(2):248–294.

Kounios, J. and Beeman, M. (2014). The cognitive neuroscience of insight. *Annual Review of Psychology*, 65:71–93.

Kouzy, R., Abi Jaoude, J., Kraitem, A., El Alam, M. B., Karam, B., Adib, E., Zarka, J., Traboulsi, C., Akl, E. W., and Baddour, K. (2020). Coronavirus goes viral: Quantifying the covid-19 misinformation epidemic on twitter. *Cureus*, 12(3).

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

Kruger, J. and Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6):1121.

Kudrowitz, B. M. (2010). *Haha and aha!: Creativity, idea generation, improvisational humor, and product design*. PhD thesis, Massachusetts Institute of Technology.

Kuperman, V., Bertram, R., and Baayen, R. H. (2008). Morphological dynamics in compound processing. *Language and Cognitive Processes*, 23(7-8):1089–1132.

Kuperman, V., Schreuder, R., Bertram, R., and Baayen, R. H. (2009). Reading polymorphemic dutch compounds: toward a multiple route model of lexical processing. *Journal of Experimental Psychology: Human Perception and Performance*, 35(3):876.

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

Landrum, A. R. and Mills, C. M. (2015). Developing expectations regarding the boundaries of expertise. *Cognition*, 134:215–231.

Libben, G. (1998). Semantic transparency in the processing of compounds: Consequences for representation, processing, and impairment. *Brain and Language*, 61(1):30–44.

Libben, G., Gibson, M., Yoon, Y. B., and Sandra, D. (2003). Compound fracture: The role of semantic transparency and morphological headedness. *Brain and Language*, 84(1):50–64.

Logg, J. M. (2017). Theory of machine: When do people rely on algorithms? Working Paper No. 17-086. Harvard Business School, Harvard University.

Logg, J. M., Minson, J. A., and Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151:90–103.

Luchkina, E., Sommerville, J. A., and Sobel, D. M. (2018). More than just making it go: Toddlers effectively integrate causal efficacy and intentionality in selecting an appropriate causal intervention. *Cognitive Development*, 45:48–56.

Lutz, D. J. and Keil, F. C. (2002). Early understanding of the division of cognitive labor. *Child Development*, 73(4):1073–1084.

MacGregor, J. N. and Cunningham, J. B. (2008). Rebus puzzles as insight problems. *Behavior Research Methods*, 40(1):263–268.

MacGregor, J. N. and Cunningham, J. B. (2009). The effects of number and level of restructuring in insight problem solving. *The Journal of Problem Solving*, 2(2):130–141.

Martin, R. A. and Ford, T. (2018). *The psychology of humor: An integrative approach* (2nd ed.). Academic Press.

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

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.* University of Minnesota Press.

Meltzoff, A. N. (1995). Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental Psychology*, 31(5):838–850.

Metcalfe, J. and Wiebe, D. (1987). Intuition in insight and noninsight problem solving. *Memory & Cognition*, 15(3):238–246.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Morkes, J., Kernal, H. K., and Nass, C. (1999). Effects of humor in task-oriented human-computer interaction and computer-mediated communication: A direct test of SRCT theory. *Human-Computer Interaction*, 14(4):395–435.

New, B. et al. (2006). Reexamining the word length effect in visual word recognition: New evidence from the english lexicon project. *Psychonomic Bulletin & Review*, 13(1):45–52.

Newell, A., Simon, H. A., et al. (1972). *Human problem solving*. Prentice-Hall.

Nickerson, R. S. (1999). How we know—and sometimes misjudge—what others know: Imputing one's own knowledge to others. *Psychological Bulletin*, 125(6):737–759.

Nijholt, A., Niculescu, A. I., Alessandro, V., and Banchs, R. E. (2017). Humor in human-computer interaction: A short survey. In *Adjunct Proceedings of INTERACT*, pages 192–214, Bombay, India. Indian Institute of Technology.

Novick, L. R. and 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., and Beaty, R. E. (2017). Ha ha? Assessing individual differences in humor production ability. *Psychology of Aesthetics, Creativity, and the Arts*, 11(2):231–241.

Ohlsson, S. (1984). Restructuring revisited: I. Summary and critique of the gestalt theory of problem solving. *Scandinavian Journal of Psychology*, 25(1):65–78.

Ohlsson, S. (1992). Information-processing explanations of insight and related phenomena. *Advances in the Psychology of Thinking*, 1:1–44.

Öllinger, M. and Knoblich, G. (2009). Psychological research on insight problem solving. In Atmanspacher, H. and Primas, H., editors, *Recasting reality*, pages 275–300. Springer.

Olteţeanu, A.-M. (2015). Towards a visual remote associates test and its computational solver. In *Proceedings of the Third International Workshop on Artificial Intelligence and Cognition 2015*, volume 1510, pages 19–28.

Olteţeanu, A.-M. (2016). From simple machines to eureka in four not-so-easy steps: Towards creative visuospatial intelligence. In *Fundamental Issues of Artificial Intelligence*, pages 161–182. Springer.

Olteţeanu, A.-M. and Falomir, Z. (2015). comRAT-C: A computational compound remote associates test solver based on language data and its comparison to human performance. *Pattern Recognition Letters*, 67:81–90.

Ovington, L. A., Saliba, A. J., Moran, C. C., Goldring, J., and MacDonald, J. B. (2018). Do people really have insights in the shower? The when, where and who of the Aha! Moment. *The Journal of Creative Behavior*, 52(1):21–34.

Paulewicz, B., Chuderski, A., and Necka, E. (2007). Insight problem solving, fluid intelligence, and executive control: A structural equation modeling approach. In *Proceedings of the 2nd European Cognitive Science Conference*, pages 586–591. Laurence Erlbaum Hove.

Premack, D. and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526.

Pungas, T. (2017). A dataset of english plaintext jokes. `https://github.com/taivop/joke-dataset`.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Rouder, J., Morey, R., Speckman, P., and Province, J. (2012). Default bayes factors for anova designs. *Journal of Mathematical Psychology*, 56(5):356–374.

Rouder, J., Speckman, P., Sun, D., Morey, R., and Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, 16(2):225–237.

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

Sandra, D. (1990). On the representation and processing of compound words: Automatic access to constituent morphemes does not occur. *The Quarterly Journal of Experimental Psychology Section A*, 42(3):529–567.

Schooler, J. W., Ohlsson, S., and Brooks, K. (1993). Thoughts beyond words: When language overshadows insight. *Journal of Experimental Psychology: General*, 122(2):166–183.

Shen, W., Luo, J., Liu, C., and Yuan, Y. (2013). New advances in the neural correlates of insight: A decade in review of the insightful brain. *Chinese Science Bulletin*, 58(13):1497–1511.

Shen, W., Yuan, Y., Liu, C., and Luo, J. (2016). In search of the 'Aha!'experience: Elucidating the emotionality of insight problem-solving. *British Journal of Psychology*, 107(2):281–298.

Sinha, R. R. and Swearingen, K. (2001). Comparing recommendations made by online systems and friends. In *DELOS Workshop: Personalisation and Recommender Systems in Digital Libraries*, volume 106.

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

Sjöbergh, J. and Araki, K. (2009). A very modular humor enabled chat-bot for japanese. In *Proceedings of PACLING*, pages 135–140.

Smith, K. A., Huber, D. E., and Vul, E. (2013). Multiply-constrained semantic search in the remote associates test. *Cognition*, 128(1):64–75.

Smith, R. W. and 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. and Davidson, J. E. (1995). *The nature of insight.* The MIT Press.

Steyvers, M., Shiffrin, R. M., and Nelson, D. L. (2005). Word association spaces for predicting semantic similarity effects in episodic memory. In Healy, A. F., editor, *Experimental cognitive psychology and its applications*, pages 237–249.

Stoll, B., Jung, M. F., and Fussell, S. R. (2018). Keeping it light: Perceptions of humor styles in robot-mediated conflict. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 247–248.

Suls, J. M. (1972). A two-stage model for the appreciation of jokes and cartoons: An information-processing analysis. In Goldstein, J. H. and McGhee, P. E., editors, *The psychology of humor: Theoretical perspectives and empirical issues*, pages 81–100. Academic Press.

Taft, M. and Forster, K. I. (1975). Lexical storage and retrieval of prefixed words. *Journal of Verbal Learning and Verbal Behavior*, 14(6):638–647.

Tay, B. T., Low, S. C., Ko, K. H., and Park, T. (2016). Types of humor that robots can play. *Computers in Human Behavior*, 60:19–28.

Threadgold, E., Marsh, J. E., and Ball, L. J. (2018). Normative data for 84 UK English rebus puzzles. *Frontiers in Psychology*, 9:2513.

Tian, F., Hou, Y., Zhu, W., Dietrich, A., Zhang, Q., Yang, W., Chen, Q., Sun, J., Jiang, Q., and Cao, G. (2017). Getting the joke: Insight during humor comprehension–evidence from an fmri study. *Frontiers in Psychology*, 8:1835.

Topolinski, S. and Reber, R. (2010). Gaining insight into the "Aha" experience. *Current Directions in Psychological Science*, 19(6):402–405.

Topolinski, S. and Strack, F. (2008). Where there'sa will—there's no intuition. the unintentional basis of semantic coherence judgments. *Journal of Memory and Language*, 58(4):1032–1048.

van Doorn, J., van den Bergh, D., Böhm, U., Dablander, F., Derks, K., Draws, T., Etz, A., Evans, N. J., Gronau, Q. F., Haaf, J. M., et al. (2021). The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychonomic Bulletin & Review*, 28:813–826.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., and Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, 60(3):158–189.

Webb, M. E., Little, D. R., and 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., and 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.

Weisberg, R. W. (1995). Prolegomena to theories of insight in problem solving: A taxonomy of problems. In Sternberg, R. J. and Davidson, J. E., editors, *The nature of insight*, pages 157–196. The MIT press.

Wertheimer, M. (1925). Über gestalttheorie. *Symposion*.

Wiley, J. (1998). Expertise as mental set: The effects of domain knowledge in creative problem solving. *Memory & Cognition*, 26(4):716–730.

Wimmer, H. and Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128.

Worthen, B. R. and Clark, P. M. (1971). Toward an improved measure of remote associational ability. *Journal of Educational Measurement*, 8(2):113–123.

Yap, M. J. and Balota, D. A. (2009). Visual word recognition of multisyllabic words. *Journal of Memory and Language*, 60(4):502–529.

Yeomans, M., Shah, A., Mullainathan, S., and Kleinberg, J. (2019). Making sense of recommendations. *Journal of Behavioral Decision Making*, 32(4):403–414.

Appendix A

Supplementary Tables for Chapter 1

Table A.1: *Normative Data for Familiarity Conditions*

Problem	Solution	Up/Down[a]	Solution rate	Mean insight rating	Mean solution time (s)[b]
lamp/beam/burn	sun	Up	52%	2.96 ± 1.92	12.19 ± 13.90
burn/beam/lamp	sun	Down	77%	4.23 ± 2.00	5.85 ± 8.78
sonic/vision/star	super	Up	48%	2.78 ± 1.86	18.45 ± 14.77
star/vision/sonic	super	Down	55%	3.59 ± 1.99	17.39 ± 16.29
bay/room/bed	sick	Up	0%	1.52 ± 0.99	36.61 ± 14.17
bed/room/bay	sick	Down	0%	2.00 ± 1.23	31.25 ± 15.24
flank/patient/fit	out	Up	4%	1.26 ± 0.62	39.08 ± 14.42
fit/patient/flank	out	Down	9%	2.00 ± 1.66	33.13 ± 18.47
long/board/ache	head	Up	74%	2.78 ± 1.81	25.41 ± 13.74
ache/board/long	head	Down	73%	2.73 ± 1.70	19.55 ± 15.92
hill/path/print	foot	Up	14%	2.55 ± 1.47	25.52 ± 17.37
print/path/hill	foot	Down	13%	1.65 ± 1.07	35.37 ± 15.05
wise/fire/word	cross	Up	0%	1.86 ± 1.46	35.53 ± 16.12
word/fire/wise	cross	Down	17%	1.52 ± 1.08	37.39 ± 16.37
clasp/brake/shake	hand	Up	55%	3.77 ± 2.09	13.28 ± 15.33
shake/brake/clasp	hand	Down	52%	3.09 ± 2.11	22.05 ± 17.60
strain/sore/sight	eye	Up	52%	3.30 ± 2.14	16.25 ± 16.36
sight/sore/strain	eye	Down	68%	4.09 ± 2.00	15.96 ± 15.09
coat/band/line	waist	Up	0%	2.55 ± 1.82	35.31 ± 15.32
line/band/coat	waist	Down	4%	2.30 ± 1.94	33.35 ± 15.04
shot/keep/set	up	Up	39%	2.48 ± 2.00	27.54 ± 17.48
set/keep/shot	up	Down	18%	2.86 ± 2.08	23.12 ± 18.06
log/less/pack	back	Up	18%	2.91 ± 2.07	27.34 ± 18.23
pack/less/log	back	Down	70%	3.09 ± 2.00	16.06 ± 13.98
water/land/basket	waste	Up	26%	1.61 ± 1.41	36.28 ± 16.65
basket/land/water	waste	Down	18%	2.18 ± 1.71	34.94 ± 15.65
ground/site/fire	camp	Up	13%	3.04 ± 2.14	23.23 ± 16.25
fire/site/ground	camp	Down	50%	4.64 ± 2.30	14.74 ± 13.46
back/bridge/string	draw	Up	9%	3.05 ± 2.34	28.98 ± 18.74
string/bridge/back	draw	Down	13%	2.13 ± 1.84	32.35 ± 19.33
light/side/wood	fire	Up	23%	3.09 ± 1.95	23.40 ± 16.41
wood/side/light	fire	Down	30%	3.00 ± 2.17	23.44 ± 18.01
well/way/case	stair	Up	18%	2.41 ± 1.68	31.33 ± 16.49
case/way/well	stair	Down	9%	2.39 ± 2.08	32.84 ± 19.17
sick/base/port	air	Up	57%	3.78 ± 2.21	19.52 ± 17.70
port/base/sick	air	Down	59%	3.64 ± 2.13	15.14 ± 12.84
light/post/shade	lamp	Up	55%	4.09 ± 2.33	13.37 ± 10.71
shade/post/light	lamp	Down	43%	3.96 ± 2.23	18.19 ± 18.39
board/port/food	sea	Up	32%	2.55 ± 2.06	24.15 ± 16.67
food/port/board	sea	Down	48%	3.30 ± 2.18	23.96 ± 15.78
tight/spout/melon	water	Up	65%	3.48 ± 2.27	22.21 ± 17.72
melon/spout/tight	water	Down	82%	3.55 ± 1.95	12.56 ± 11.46

[a] "Up" indicates cues began with least familiar word cue; "Down" indicates that cues began with the most familiar cue word.
[b] Denotes submission time following presentation of all three cues (i.e., 10 s after start of trial).

Table A.2: *Normative Data for Lexeme Meaning Dominance Conditions*

Problem	Solution	Up/Down[a]	Solution rate	Mean insight rating	Mean solution time (s)[b]
land/basket/water	waste	Up	24%	1.29 ± 0.72	35.43 ± 14.62
water/basket/land	waste	Down	52%	2.14 ± 1.68	26.95 ± 16.58
place/light/fly	fire	Up	48%	3.19 ± 2.34	21.28 ± 16.12
fly/light/place	fire	Down	62%	3.62 ± 2.27	18.16 ± 15.60
front/spout/melon	water	Up	90%	5.29 ± 2.12	10.05 ± 9.04
melon/spout/front	water	Down	95%	5.24 ± 1.81	4.78 ± 6.99
case/well/way	stair	Up	19%	2.57 ± 2.13	27.29 ± 17.76
way/well/case	stair	Down	24%	3.10 ± 2.32	28.56 ± 18.28
out/cuff/gun	hand	Up	57%	3.95 ± 2.56	16.86 ± 18.89
gun/cuff/out	hand	Down	52%	4.14 ± 2.69	16.03 ± 15.76
ground/site/fire	camp	Up	19%	2.81 ± 1.94	25.35 ± 16.80
fire/site/ground	camp	Down	29%	3.43 ± 2.20	26.65 ± 18.74
borne/lift/mail	air	Up	67%	4.24 ± 2.05	15.47 ± 16.35
mail/lift/borne	air	Down	67%	4.00 ± 2.32	14.04 ± 14.41
human/impose/market	super	Up	43%	3.48 ± 2.42	21.48 ± 18.53
market/impose/human	super	Down	71%	3.90 ± 2.39	18.62 ± 17.31
spot/roof/day	sun	Up	48%	4.14 ± 2.26	16.71 ± 14.34
day/roof/spot	sun	Down	52%	4.10 ± 2.30	16.63 ± 16.94
line/band/coat	waist	Up	0%	1.86 ± 1.62	34.05 ± 15.00
coat/band/line	waist	Down	5%	2.95 ± 2.01	29.48 ± 15.84
work/race/note	foot	Up	5%	2.33 ± 1.83	35.73 ± 16.87
note/race/work	foot	Down	14%	1.67 ± 1.56	33.58 ± 20.13
back/string/bridge	draw	Up	14%	2.19 ± 1.83	32.92 ± 18.15
bridge/string/back	draw	Down	5%	2.33 ± 1.74	32.30 ± 16.05
side/break/cry	out	Up	29%	2.62 ± 2.18	29.30 ± 17.34
cry/break/side	out	Down	33%	2.86 ± 2.52	33.44 ± 15.66
grade/stage/tight	up	Up	29%	2.81 ± 2.32	32.85 ± 18.43
tight/stage/grade	up	Down	38%	3.05 ± 2.48	26.62 ± 17.94
front/food/sick	sea	Up	14%	2.71 ± 2.05	30.34 ± 18.59
sick/food/front	sea	Down	19%	2.38 ± 2.11	29.98 ± 20.18
wise/walk/word	cross	Up	5%	1.76 ± 1.55	42.04 ± 13.45
word/walk/wise	cross	Down	14%	2.00 ± 1.55	34.30 ± 18.18
bay/room/bed	sick	Up	0%	3.38 ± 2.11	27.41 ± 16.76
bed/room/bay	sick	Down	0%	1.90 ± 1.14	34.98 ± 18.21
post/light/shade	lamp	Up	43%	3.90 ± 2.05	22.56 ± 19.22
shade/light/post	lamp	Down	29%	3.86 ± 1.82	13.01 ± 10.98
less/bone/seat	back	Up	43%	3.71 ± 2.39	25.08 ± 15.21
seat/bone/less	back	Down	52%	3.81 ± 2.34	20.77 ± 18.36
set/ache/stone	head	Up	81%	5.71 ± 1.85	11.43 ± 15.04
stone/ache/set	head	Down	71%	5.52 ± 2.18	11.54 ± 14.22
ball/wash/witness	eye	Up	62%	4.76 ± 2.28	23.49 ± 17.07
witness/wash/ball	eye	Down	62%	4.62 ± 2.33	15.84 ± 15.50

[a] "Up" indicates cues began with least dominant word cue; "Down" indicates that cues began with the most dominant cue word.
[b] Denotes submission time following presentation of all three cues (i.e., 10 s after start of trial).

Table A.3: *Normative Data for Semantic Transparency Conditions*

Problem	Solution	Up/Down[a]	Solution rate	Mean insight rating	Mean solution time (s)[b]
proof/way/spout	water	Up	50%	3.15 ± 2.01	16.72 ± 11.65
spout/way/proof	water	Down	67%	2.95 ± 1.83	18.87 ± 20.06
fly/ball/wood	fire	Up	35%	3.20 ± 2.21	23.14 ± 18.01
wood/ball/fly	fire	Down	38%	3.52 ± 2.16	28.96 ± 17.40
land/water/basket	waste	Up	25%	2.15 ± 1.57	31.85 ± 16.14
basket/water/land	waste	Down	29%	2.10 ± 1.55	34.22 ± 16.07
wise/fire/road	cross	Up	0%	1.95 ± 1.24	36.46 ± 15.69
road/fire/wise	cross	Down	0%	1.90 ± 1.29	33.79 ± 15.06
maiden/cart/gun	hand	Up	30%	2.10 ± 1.33	28.74 ± 17.96
gun/cart/maiden	hand	Down	29%	3.19 ± 2.14	26.37 ± 18.25
board/front/water	sea	Up	5%	2.81 ± 1.97	33.06 ± 15.79
water/front/board	sea	Down	0%	2.35 ± 1.50	29.38 ± 16.62
well/case/way	stair	Up	14%	2.43 ± 1.83	35.10 ± 15.36
way/case/well	stair	Down	5%	2.15 ± 1.87	32.53 ± 15.03
shade/post/light	lamp	Up	43%	3.95 ± 2.20	20.54 ± 18.64
light/post/shade	lamp	Down	35%	2.85 ± 1.76	24.60 ± 15.76
note/hold/print	foot	Up	24%	3.24 ± 2.07	27.45 ± 17.89
print/hold/note	foot	Down	0%	2.40 ± 1.60	29.13 ± 17.49
coat/line/band	waist	Up	0%	2.45 ± 1.79	28.60 ± 14.55
band/line/coat	waist	Down	10%	2.71 ± 1.87	31.93 ± 18.32
less/strain/sight	eye	Up	52%	2.62 ± 1.56	34.26 ± 16.77
sight/strain/less	eye	Down	60%	2.95 ± 1.79	27.14 ± 17.03
fit/post/grow	out	Up	29%	2.86 ± 2.17	34.92 ± 16.08
grow/post/fit	out	Down	10%	2.10 ± 1.97	38.27 ± 16.59
line/craft/float	air	Up	5%	2.95 ± 1.88	25.24 ± 18.29
flow/craft/line	air	Down	33%	3.81 ± 2.25	26.14 ± 18.64
back/string/bridge	draw	Up	19%	2.86 ± 2.37	31.38 ± 18.23
bridge/string/back	draw	Down	5%	2.10 ± 1.71	32.58 ± 17.79
bay/bed/room	sick	Up	0%	2.80 ± 2.04	27.14 ± 17.60
room/bed/bay	sick	Down	0%	2.14 ± 1.28	36.34 ± 14.84
day/lamp/light	sun	Up	38%	3.14 ± 2.43	29.10 ± 18.76
light/lamp/day	sun	Down	15%	3.40 ± 2.21	20.88 ± 17.63
impose/sonic/human	super	Up	30%	2.30 ± 1.81	32.48 ± 20.36
human/sonic/impose	super	Down	43%	3.05 ± 2.58	28.41 ± 19.68
set/side/hill	up	Up	25%	2.95 ± 2.19	26.97 ± 19.64
hill/side/set	up	Down	38%	3.48 ± 2.50	26.99 ± 18.19
log/hand/ache	back	Up	35%	2.70 ± 1.87	31.55 ± 17.96
ache/hand/log	back	Down	52%	33.3 ± 2.27	31.62 ± 18.98
long/strong/ache	head	Up	57%	2.90 ± 2.21	30.78 ± 17.61
ache/strong/long	head	Down	50%	2.70 ± 1.69	25.92 ± 17.89
ground/site/fire	camp	Up	29%	4.10 ± 2.21	20.97 ± 15.83
fire/site/ground	camp	Down	40%	4.55 ± 2.54	20.14 ± 18.21

[a] "Up" indicates cues began with least transparent word cue; "Down" indicates that cues began with the most transparent cue word.
[b] Denotes submission time following presentation of all three cues (i.e., 10 s after start of trial).

Appendix B

Supplementary Figure and Tables for Chapter 2

Table B.1: *Sample Responses from Joke Completion Task with Mean Funniness Ratings*

Prompt	Mean funniness (1-2)	Mean funniness (2-3)	Mean funniness (3-4)
A TREE walks into a bar…	Are you doing OAKAY? (1.9)	saying "I seem to have lost my roots." (2.3)	And asks for a tall one (3)
A DOCTOR walks into a bar…	I'll take a shot of whiskey, please. (1.9)	and says "I don't have the patience to wait all night for a drink" (2.9)	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)	It asks for a code brew. (2.5)	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)	NASA fast, what'd you say? (2.8)	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)	well, don't get defensive about it (2.7)	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)	Then it should taste just purr-fect. (2.7)	You've gotta be kitten me! (3.1)
GUITAR, I'm braking up with you…	That is off key! (1.9)	There's too great a riff't between us. (2.9)	You're just stringing me along. (3.7)
OCEAN, I'm breaking up with you…	I'll wave bye! (1.9)	I just don't sea a future here. (2.8)	I feel too tide down (3.4)
BOOK, I'm breaking up with you…	Way to write me off (1.9)	I decided to turn the page on my life (2.9)	You judge people by their cover (3.9)

Table B.2: *Normative Data for Joke Completion Task*

Item	Joke Stem	Mean funniness	Solution rate	Mean insight rating	Modal insight rating	Mean solution time (s)
1	A TREE walks into a bar...	2.46	35%	2.77 ± 1.32	4	42.71 ± 22.05
2	A DOCTOR walks into a bar...	2.14	30%	2.54 ± 2.11	3	55.64 ± 23.86
3	A CAR walks into a bar...	2.23	24%	2.61 ± 1.10	2	45.92 ± 22.65
4	A PIRATE walks into a bar...	1.99	13%	2.3 ± 1.19	1	54.96 ± 25.37
5	A COMPUTER walks into a bar...	2.23	29%	2.59 ± 1.19	3	49.67 ± 26.08
6	An ARTIST walks into a bar...	2.10	22%	2.47 ± 1.18	3	49.71 ± 25.19
7	Waiter, there's a BIRD in my soup!	2.34	37%	2.75 ± 1.18	2	34.33 ± 21.09
8	Waiter, there's a SANDWICH in my soup!	1.95	13%	2.29 ± 1.08	1	43.66 ± 26.64
9	Waiter, there's a LAWYER in my soup!	2.32	37%	2.78 ± 1.22	3	33.85 ± 16.75
10	Waiter, there's a CAT in my soup!	2.31	37%	2.63 ± 1.17	3	34.1 ± 20.86
11	Waiter, there's an ASTRONAUT in my soup!	2.44	41%	2.69 ± 1.23	2	35.72 ± 21.99
12	Waiter, there's a PENCIL in my soup!	2.32	33%	2.67 ± 1.09	2	33.69 ± 22.17
13	OCEAN, I'm breaking up with you...	2.74	61%	2.97 ± 1.12	3	24.48 ± 17.57
14	GHOST, I'm breaking up with you...	2.48	43%	3.01 ± 1.09	3	23.83 ± 15.45
15	CLOCK, I'm breaking up with you...	2.62	56%	3.04 ± 1.12	3	23.06 ± 18.95
16	BOOK, I'm breaking up with you...	2.65	56%	2.83 ± 1.14	3	28.09 ± 16.53
17	BANK, I'm breaking up with you...	2.55	48%	2.89 ± 1.12	3	31.85 ± 19.46
18	GUITAR, I'm breaking up with you...	2.88	70%	3.07 ± 1.08	3	25.25 ± 16.77

Note. Joke submissions were designated as being "correct" if their average funniness rating was greater than or equal to 2.6 (surpassing the Likert median).

Table B.3: *Normative Data for the Rebus Puzzle Task*

Item	Solution	Solution rate	Mean insight rating	Modal insight rating	Mean solution time (s)
1	"Dark Ages"	8%	2.65 ± 1.43	1	19.92 ± 15.12
2	"Big Bad Wolf"	78%	3.21 ± 1.48	5	13.60 ± 12.22
3	"Capital Punishment"	49%	2.86 ± 1.50	1	18.28 ± 14.95
4	"Little League"	75%	3.56 ± 1.38	5	9.22 ± 6.67
5	"A Grave Error"	1%	1.26 ± 0.68	1	39.59 ± 20.80
6	"Ambiguous"	31%	2.22 ± 1.40	1	31.37 ± 19.30
7	"Excuse Me"	61%	2.77 ± 1.59	1	23.35 ± 17.26
8	"Tea for Two"	13%	1.95 ± 1.45	1	34.99 ±19.45
9	"Go Stand in the Corner"	10%	2.17 ± 1.25	1	22.48 ± 18.09
10	"A Round of Applause"	54%	3.03 ± 1.65	5	22.33 ± 17.59
11	"Split Personality"	54%	3.23 ± 1.53	5	16.08 ± 12.98
12	"Waterfall"	51%	3.12 ± 1.30	3	12.74 ± 10.82
13	"Paralegal"	23%	2.07 ± 1.34	1	28.02 ± 19.68
14	"Too Big to Ignore"	52%	2.65 ± 1.56	1	24.99 ± 17.76
15	"Sit Down and Shut Up"	43%	2.78 ± 1.76	1	28.71 ± 18.56
16	"Searching High and Low"	24%	2.52 ± 1.56	1	24.08 ± 17.70
17	"Somewhere Over the Rainbow"	89%	3.68 ± 1.48	5	10.92 ± 7.30
18	"A Home Away from Home"	57%	3.16 ± 1.38	3	16.65 ± 14.07
19	"Beating Around the Bush"	83%	3.72 ± 1.43	5	11.38 ± 6.80
20	"Diamond in the Rough"	78%	3.37 ± 1.57	5	18.09 ± 13.74
21	"A Little on the Large Side"	7%	1.62 ± 1.18	1	35.27 ± 19.80
22	"Just Between Friends"	51%	3.18 ± 1.51	5	15.51 ± 13.22
23	"Lying on the Job"	68%	3.19 ± 1.49	5	17.02 ± 13.87
24	"Rock Around the Clock"	70%	3.7 ± 1.42	5	10.44 ± 8.44

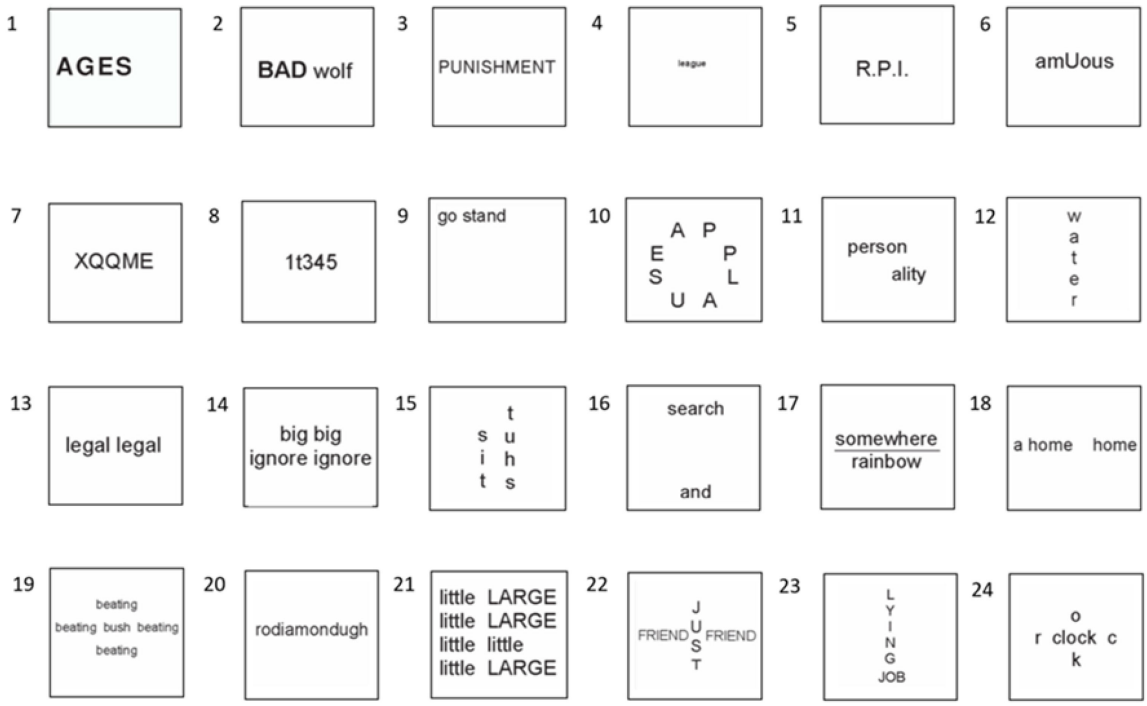

Figure B.1: *Rebus puzzles (adopted from MacGregor and Cunningham, 2009).*

Appendix C

Supplementary Table for Chapter 3

Table C.1: *Sample joke stimuli with their actual and most frequently guessed sources from Experiment 1.*

Joke	Actual Source	Mode Guessed Source
My girlfriend told me to take the spider out instead of killing it. We went and had some drinks. Cool guy. Wants to be a web developer.	Human	Probably Human
As I suspected, someone has been adding soil to my garden. The plot thickens.	Human	Probably Human
I've recently developed a severe phobia of elevators. I'm taking steps to avoid them.	Human	Probably Human
The word 'nothing' is a palindrome. 'Nothing' reversed is 'Gnihton' which is also nothing.	Human	Probably AI
The bartender says "We don't serve time travelers in here." A time traveler walks into a bar.	Human	Probably AI
What's the stupidest animal in the jungle? The Polar bear.	Human	Probably AI
What do you get when you cross a hippie, a hipster, and a vegan? A... vegan? (I'll go with hippie to avoid the hipster joke.)	AI	Probably Human
How do you get a woman pregnant? Just say: "You too, baby."	AI	Probably AI
I was gonna tell the guy who invented the telephone how it works... But he's dead.	AI	Probably AI
Why won't the chicken fly over the balcony? Actually, you'll get it tomorrow.	AI	Probably AI
I tried to ask this guy out. It didn't go well.	AI	Definitely AI
Is it safe to eat in a restaurant while smoking? ... you'll get a bite on your bill.	AI	Definitely AI
When my dad came home he was surprised by the smell of bacon.	AI	Definitely AI