Beyond Literal Meaning: Linguistic And Cognitive Features Of Figurative Language Processing And Production

STEPHEN SKALICKY

Follow this and additional works at: https://scholarworks.gsu.edu/alesl_diss

Recommended Citation
https://scholarworks.gsu.edu/alesl_diss/44

This Dissertation is brought to you for free and open access by the Department of Applied Linguistics and English as a Second Language at ScholarWorks @ Georgia State University. It has been accepted for inclusion in Applied Linguistics and English as a Second Language Dissertations by an authorized administrator of ScholarWorks @ Georgia State University. For more information, please contact scholarworks@gsu.edu.
ABSTRACT

Figurative language, such as metaphor and irony, has been studied by researchers in a wide variety of fields for the past several decades. A primary goal of this research has been to determine the processes underlying figurative language comprehension (e.g., how the figurative meaning of a metaphor is processed when compared to its literal meaning). While this research has led to a better understanding of the relation between figurative meaning and surface linguistic form, it has also served to underemphasize other potentially important influences on figurative language use, such as participant individual differences (e.g., cognitive ability, language background), linguistic features (e.g., lexical sophistication), and affective perceptions (e.g., humorous reactions). The purpose of this dissertation was to analyze figurative language...
use from a dynamic perspective taking into account a multitude of potential influences on the production, processing, and perception of figurative language through three separate studies. In the first two studies, 61 human participants were recruited to complete a metaphor and sarcasm production experiment as well as a satire processing and comprehension experiment. Individual differences measures (e.g., language background, working memory capacity) were collected, and the metaphors, sarcastic responses, and satirical texts were analyzed for linguistic features related to lexical sophistication and semantic cohesion. Results revealed that a number of individual differences and linguistic features were significant predictors of figurative language production ability, but also that these features interacted with contextual variables (e.g., the metaphor or sarcasm prompts) in a dynamic manner, making it difficult to identify consistent influences on figurative language production ability. Moreover, satire comprehension ability was significantly influenced by affective perceptions of sincerity, positivity, and humor, as well as by relative English age of onset, but not by processing time, suggesting that comprehension of satirical meaning is heavily dependent upon pragmatic inference. In the third study, 423 online workers completed an online creative production task, the results of which suggested that the presence of figurative language significantly increased perceptions of creativity, highlighting further connections between creativity and figurative language. Overall, these results demonstrate the importance of adopting a dynamic approach towards understanding figurative language use.

INDEX WORDS: Satire, Sarcasm, Metaphor, Creativity, Individual differences, Psycholinguistics, Cognitive linguistics
DEDICATION

This dissertation is dedicated to my very understanding wife, Yeti Skalicky, who has supported me without question during every stage of this process.
ACKNOWLEDGEMENTS

This dissertation is the result of both my own efforts as well as many others who have supported me during my entire graduate career. Among the many people who have helped me, Nancy Bell was the first professor to actually treat me like a graduate student, and whose patience and enthusiasm led me through my master’s degree and towards my PhD. Without her I simply would not be where I am today, and I am extremely thrilled that I am able to continue working with her and that she is a member of my PhD committee.

I thank Eric Friginal for the numerous opportunities he has provided for me, and for the fantastic time I had as part of the global TESOL program in Guangzhou. Eric is masterful when it comes to locating resources, publication outlets, and other opportunities for students, and truly cares about ensuring success for all of the graduate students in the Department of Applied Linguistics, and I feel extremely grateful to have worked with him. I also want to thank YouJin Kim, who has been an ally since my very first visit to Georgia State, even before I entered the program. Working for YouJin as a research assistant taught me a great deal about research methodology, second language acquisition, and even what a stranded preposition is. YouJin may work behind the scenes to look out for mine (and others) well being, but she should know that all of her effort is deeply appreciated.

I am also indebted to several other members of the Department of Applied Linguistics. I thank Ute Römer for helping me learn Sinclarian corpus linguistic methods and applying them to a topic of my own interest. I thank Viviana Cortes for teaching me that it is perfectly fine for me to teach writing the way I want to teach it. I thank Stephanie Lindemann for her good humor and Diane Belcher for reminding me about the qualitative side of research.
And, of course, I must provide big thanks to my advisor, Scott Crossley, for his mentorship and other assistance during my time in the PhD program. Although I did not envision working with Scott when I entered the program, I am extremely glad that I did. This is because Scott pushed me to become a true researcher, one who is not afraid to learn new things and instead welcomes the idea that there is always something new to learn. Scott has provided me with an uncountable amount of opportunities that have helped me develop a wide skill set and created memories that I will carry with me forever.

Finally, to all of my friends at Georgia State, including Dave Chiesa, Cindy Berger, Kris Kyle, Rurik Tywoniw, Sarah Goodwin, and others, I thank you for being there and am truly happy that I was able to get to know you all. Working through a PhD is one of the loneliest processes I have endured, but it was made all the more better when surrounded by good friends.
TABLE OF CONTENTS

ACKNOWLEDGEMENTS ........................................................................................................... V

LIST OF TABLES .................................................................................................................... XV

LIST OF FIGURES ..................................................................................................................... XVII

1 INTRODUCTION ................................................................................................................... 1

   1.1 Current Study .................................................................................................................... 10

2 LITERATURE REVIEW ....................................................................................................... 13

   2.1 Types of Figurative Language .......................................................................................... 14

      2.1.1 Metaphor and Simile ................................................................................................. 14

      2.1.2 Verbal Irony and Sarcasm ......................................................................................... 16

      2.1.3 Satire ........................................................................................................................ 19

   2.2 Figurative Language Research ....................................................................................... 21

      2.2.1 Metaphor processing ................................................................................................. 21

      2.2.2 Verbal irony processing ........................................................................................... 22

      2.2.3 Satire processing ....................................................................................................... 26

   2.3 Variables Beyond Literal and Figurative Meaning ............................................................ 27

   2.4 Individual Differences and Figurative Language ............................................................. 28

   2.5 Cognitive Measures ........................................................................................................ 28

      2.5.1 Intelligence ................................................................................................................ 29

      2.5.2 Working memory capacity (WMC) ............................................................................ 32
2.5.3 Need for cognition (NFC). ................................................................. 35

2.5.4 Abstract thinking (AT). ..................................................................... 36

2.5.5 Summary of cognitive measures. ....................................................... 38

2.6 Demographic Measures ........................................................................ 39

2.6.1 Gender and age. ................................................................................ 39

2.6.2 Language knowledge. ....................................................................... 42

2.7 Summary of Individual Differences....................................................... 48

2.8 Figurative Language and Linguistic Features ....................................... 48

2.8.1 Lexical sophistication. ...................................................................... 52

2.8.2 Semantic cohesion. .......................................................................... 53

2.8.3 Summary of linguistic features. ......................................................... 55

2.9 The Pragmatics of Figurative Language ............................................ 55

2.9.1 Summary of pragmatic influences. .................................................. 61

2.10 Literature Review Summary ............................................................... 61

3 INDIVIDUAL DIFFERENCES AND LINGUISTIC FEATURES OF
METAPHOR AND SARCASM PRODUCTION ........................................... 63

3.1 Introduction .......................................................................................... 63

3.2 Research Questions .............................................................................. 65

3.3 Method .................................................................................................. 65

3.4 Participants ........................................................................................... 66
3.5 Materials .................................................................................................................. 66

3.5.1 Demographic survey. ............................................................................................. 66

3.5.2 Fluid intelligence test. .......................................................................................... 66

3.5.3 Crystallized intelligence test. ............................................................................... 67

3.5.4 Broad retrieval tests. ........................................................................................... 68

3.5.5 Working memory capacity test. ............................................................................. 69

3.5.6 Abstract thinking survey....................................................................................... 70

3.5.7 Need for cognition survey..................................................................................... 70

3.5.8 Language aptitude test......................................................................................... 71

3.5.9 Metaphor production items. ................................................................................ 72

3.5.10 Sarcasm production items. ................................................................................ 72

3.6 Procedure .................................................................................................................. 74

3.6.1 Online survey. ....................................................................................................... 74

3.6.2 Laboratory session................................................................................................. 74

3.6.3 Metaphor production tasks.................................................................................... 76

3.6.4 Sarcasm production tasks...................................................................................... 78

3.6.5 Human ratings........................................................................................................ 79

3.6.6 Linguistic features................................................................................................ 82

3.6.7 Language background.......................................................................................... 85

3.7 Statistical Analysis .................................................................................................... 86
3.8  Results ................................................................................................................................. 90

3.9  Dimension Reduction ........................................................................................................... 90

3.9.1  Broad retrieval .................................................................................................................. 90

3.9.2  Sarcasm ratings. .............................................................................................................. 91

3.9.3  Metaphor ratings. ............................................................................................................ 92

3.10  Descriptive Statistics .......................................................................................................... 93

3.10.1  Individual differences. .................................................................................................. 93

3.10.2  Human ratings. ............................................................................................................. 93

3.11  Correlations ........................................................................................................................ 94

3.11.1  Individual differences..................................................................................................... 94

3.11.2  Individual differences and production measures. ......................................................... 94

3.11.3  Linguistic features and production measures. .............................................................. 95

3.12  Linear Mixed Effects Models ............................................................................................. 96

3.13  Individual Differences in Metaphor Production ............................................................... 97

3.13.1  Results: Metaphor conceptual distance ratings. ......................................................... 97

3.13.2  Results: Metaphor novelty/mirth scores. ..................................................................... 101

3.13.3  Summary of individual differences in metaphor production................................. 108

3.14  Individual Differences in Sarcasm Production ................................................................. 110

3.14.1  Results: Sarcasm incongruity ratings. .......................................................................... 110

3.14.2  Results: Sarcasm novelty/mirth ratings. ..................................................................... 118
3.14.3 Summary of individual differences in sarcasm production. 120

3.15 Linguistic Features and Metaphor Production 121

3.15.1 Results: Metaphor conceptual distance ratings. 121

3.15.2 Results: Metaphor novelty/mirth scores. 125

3.16 Linguistic Features and Sarcasm Production 127

3.16.1 Results: Sarcasm incongruity ratings. 127

3.16.2 Results: Sarcasm novelty/mirth ratings. 130

3.17 General Discussion 134

3.18 Conclusion 142

4 INDIVIDUAL DIFFERENCES AND LINGUISTIC FEATURES OF SATIRE PROCESSING 143

4.1 Introduction 143

4.2 Research Questions 146

4.3 Method 146

4.4 Participants 147

4.5 Materials and Study Design 147

4.5.1 Demographic survey and individual differences tests. 147

4.5.2 Satirical and non-satirical texts. 147

4.5.3 Linguistic features. 150

4.5.4 Comprehension and affective perceptions questions. 151
4.5.5 Genre background survey

4.6 Procedure

4.6.1 Online survey

4.6.2 Laboratory session

4.6.3 Satire processing experiment

4.6.4 Comprehension question ratings

4.6.5 Genre background information

4.7 Statistical Analysis

4.8 Results

4.8.1 Affective ratings

4.8.2 Reading times

4.8.3 Satirical comprehension

4.9 Conclusion

5 INFLUENCE OF CREATIVE LANGUAGE ON PERCEPTIONS OF CREATIVITY

5.1 Introduction

5.2 Research Question

5.3 Method

5.4 Participants

5.5 Materials
5.5.1 News story prompts ................................................................. 176

5.6 Procedure ................................................................................. 179

5.7 Data Coding ........................................................................... 183

5.8 Statistical Analysis .................................................................. 183

5.9 Participant Creativity Ratings .................................................. 185

5.10 Linguistic Creativity in Participant Responses ......................... 186

5.11 Ordinal Linear Mixed Effects Model ....................................... 187

5.12 Discussion .............................................................................. 189

5.13 Conclusion ............................................................................. 194

6 CONCLUSION ............................................................................. 194

6.1 Future Directions ...................................................................... 201

6.2 Conclusion ............................................................................... 204

REFERENCES ................................................................................. 205

APPENDICES ..................................................................................... 229

Appendix A. Demographic and Language Background Survey ........... 229

*Informed Consent* ......................................................................... 229

*Demographic and Language Information* ..................................... 231

Appendix B: Fluid Intelligence Task .............................................. 234

Appendix C. Crystallized Intelligence Test ....................................... 235

Appendix D. Abstract Thinking Test .............................................. 240
LIST OF TABLES

Table 3.1 Participant demographic information. .......................................................... 66
Table 3.2 Inter-rater reliability scores for human ratings of metaphor and sarcasm. ........ 81
Table 3.3 Descriptive statistics and internal consistency for individual differences measures. 93
Table 3.4 Descriptive statistics for human ratings of metaphors and sarcastic responses. .... 94
Table 3.5 Pearson's correlations among the cognitive individual differences measures. ....... 95
Table 3.6 Correlations among participant individual differences and human ratings of metaphor and sarcasms. .......................................................................................... 95
Table 3.7 Correlations among linguistic features and human ratings of metaphors and sarcasms. .................................................................................................................. 95
Table 3.8 Linear mixed effects model predicting human ratings of conceptual distance scores of participant metaphor production. ................................................................. 98
Table 3.9 Linear mixed effects model predicting human ratings of novelty/mirth scores of participant metaphor production ..................................................................................... 102
Table 3.10 Linear mixed effect model predicting human ratings of incongruity of participant sarcastic responses ............................................................................................... 111
Table 3.11 Linear mixed effect model predicting human ratings of novelty/mirth of participant sarcastic responses ............................................................................................... 118
Table 3.12 Linear mixed effects model predicting conceptual distance scores for metaphors using linguistic features .......................................................................................... 121
Table 3.13 Linear mixed effects model predicting novelty/mirth scores for metaphors using linguistic features ............................................................................................... 125
Table 3.14 Linear mixed effects model predicting incongruity scores for sarcastic responses using linguistic features

Table 3.15 Linear mixed effects model predicting novelty/mirth scores for sarcastic responses using linguistic features

Table 4.1 Move structure of Onion scientific report articles genre

Table 4.2 General topics of satirical and non-satirical texts per sub-experiment

Table 4.3 Mean and standard deviation for linguistic features of satirical and non-satirical texts

Table 4.4 Results from Genre Background Knowledge survey.

Table 4.5 Descriptive statistics and Welch t-test results for affective ratings of sincerity, humor, and positivity

Table 4.6 Linear mixed effects model predicting reading times (in seconds) for satirical and non-satirical texts

Table 4.7 Logistic linear mixed effects model predicting satirical comprehension using individual differences and reading times

Table 5.1 Ordinal linear mixed model predicting participant comparison ratings of creativity.
LIST OF FIGURES

Figure 3.1 Sample conventional metaphor screen from E-Prime ................................................. 77
Figure 3.2 Sample novel metaphor prompt screen from E-Prime .................................................. 78
Figure 3.3 Sample sarcasm prompt using three-panel comic from E-Prime ................................. 79
Figure 3.4 Interaction between metaphor type and WMC (conceptual distance) ...................... 99
Figure 3.5 Interaction between metaphor type and Crystallized Intelligence (conceptual distance) ........................................... 100
Figure 3.6 Interaction between metaphor type and Need for Cognition (metaphor novelty/mirth) .......................................................................................................................... 105
Figure 3.7 Interaction between Broad Retrieval and Production Time (metaphor novelty/mirth) .......................................................................................................................... 106
Figure 3.8 Interaction between English age of onset and Crystallized Intelligence (metaphor novelty/mirth) .................................................................................................................. 107
Figure 3.9 Interaction between English age of onset and Broad Retrieval (sarcasm incongruity) ........................................................................................................................................ 112
Figure 3.10 Interaction between production time and Crystallized Intelligence (sarcasm incongruity) ........................................................................................................................................ 113
Figure 3.11 Interaction between prompt type and Crystallized Intelligence (sarcasm incongruity) ........................................................................................................................................ 114
Figure 3.12 Example of desert island comic used as sarcastic stimuli ..................................... 115
Figure 3.13 Interaction between prompt type and English age of onset (sarcasm incongruity). 116
Figure 3.14 Example black and white comic used as a sarcasm prompt ..................................... 117
Figure 3.15 Sarcastic comic stimuli related to travelling the world in the military ................. 129
Figure 3.16 Interaction between MRC Familiarity and prompt (sarcasm novelty/mirth)........ 132
Figure 3.17 Interaction between Brysbaert Concreteness and prompt (sarcasm novelty/mirth) 132
Figure 4.1 Text stimuli screen for a satirical text shown during satire processing experiment.. 155
Figure 5.1 Procedure flow.............................................................................................. 182
Figure 5.2 Participant creativity self-ratings...................................................................... 186
Figure 5.3 Interaction between self-ratings and incidence of linguistic creativity............... 187
1 INTRODUCTION

Figurative language has been studied by researchers in linguistics, psychology, and rhetoric for the past several decades (Gibbs & Colston, 2007, 2012; Glucksberg, 2001; Ortony, 1993; Roberts & Kreuz, 1994; Simpson, 2003). Examples of figurative language include metaphor, idioms, irony, and satire, which exist in both spoken and written language (Burgers, van Mulken, & Schellens, 2012; Corts & Meyers, 2002; Gibbs, 1994, 2000; McCarthy & Carter, 2004). Despite differences in form and function, one common definition that groups these varied examples of figurative language together is that they are typically described as “speech where speakers mean something other than what they literally say” (Gibbs & Colston, 2012, p. 1). This definition has led to the assumption that figurative language represents a deviation from normal, literal language use (Glucksberg, 2001), and has shaped the scope and purpose of much of the scientific research into figurative language.

This assumption has led to a prevalence of studies that focus on developing and testing models of language processing specifically designed to account for figurative language comprehension (Filik, Leuthold, Wallington, & Page, 2014; Gibbs & Colston, 2012; Giora, 2003). In general, this strand of research has been influenced by one major theory related to figurative language processing, known as the Standard Pragmatic Model (SPM). This theory emerged from early research in language pragmatics and postulates that the interpretation of figurative meaning first requires the recognition, comprehension, and rejection of a figurative utterance’s literal meaning in order to arrive at a secondary, figurative interpretation (Grice, 1975, 1978; Sperber & Wilson, 1981). For example, in order to understand the figurative meaning behind the metaphor all lawyers are sharks, the SPM predicts a hearer must first recognize that a speaker does not literally mean lawyers are the same as predatory fish in the sea,
and then construct a new, figurative meaning based on a reinterpretation of the utterance (e.g., all lawyers are ruthless). Therefore, many psycholinguistic experiments test figurative language comprehension by comparing reading times for literal and figurative language that have the same approximate meaning (e.g., comparing reading times for *all lawyers are sharks* against *all lawyers are ruthless*). In general, the authors of these studies point towards differences in online processing\(^1\) between literal and figurative utterances as evidence that figurative language is processed in a qualitatively different manner when compared to non-figurative language.

Despite revealing important information about figurative language comprehension, the focus on testing theories of figurative language processing has served to stall current scientific understandings of figurative language (Gibbs & Colston, 2012). This is evident in the marginalization of studies investigating other variables that may play a role during figurative language use, as well as the lack of methodological diversity in figurative language research (Gibbs & Colston, 2012). The purpose of this dissertation is to first review how a multitude of features beyond literal and figurative meaning are important to consider when researching figurative language use, such as participant individual differences, linguistic features of the stimuli, and pragmatic and social variables that influence meaning comprehension. Then, connections among these features and figurative language use will be explored through three original experiments that also address calls for methodological diversity and improvement. The results from this dissertation will thus contribute towards a holistic understanding of figurative language use that complements the existing body of research. Before describing the research questions and experimental studies in detail, the following section presents a brief summary of

---

1 Online processing refers to real-time processing, typically measured through reading times or by recording the eye-movements of participants with an eye-tracking camera.
the different research gaps which allow for an expansion of the current scope and purpose of figurative language research, and briefly touches on how this dissertation will address them.

Most notably, relatively few studies have investigated figurative language production (Benedek et al., 2014; Corts & Meyers, 2002; J. Johnson & Rosano, 1993). Examining figurative language processing only provides a partial understanding, yet the focus on models of processing continue to place studies of figurative language production on the periphery. While some research investigating figurative language production does exist, the majority has only focused on one type of figurative language: metaphor (Beaty & Silvia, 2013; Benedek et al., 2014; Chiappe & Chiappe, 2007; J. Johnson & Rosano, 1993; Pierce & Chiappe, 2008; Silvia & Beaty, 2012). Therefore, more work is needed examining the production of different types of figurative language in order to complement the wide number of studies related to figurative language comprehension. One approach to studying figurative language production is to provide participants with stimuli designed to elicit figurative language and then rate the resulting produced language for different qualities related to specific types of figurative language (e.g., conceptual distance in metaphors; Chiappe & Chiappe, 2007; Pierce & Chiappe, 2008). The first study of this dissertation adopts this approach towards the production of metaphor and sarcasm, which simultaneously builds upon the current knowledge of metaphor production while contributing new information regarding sarcasm production.

Secondly, a number of factors are present when hearers process language, whether figurative or literal, and these factors may in turn influence the online processing of figurative language. Because of this, some researchers have argued that the focus on testing models of figurative language processing is misguided because figurative language employs the same fundamental language processing and comprehension mechanisms as non-figurative language
Arguments from this position state that any perceived differences in figurative and non-figurative language processing stem from interactions with linguistic, cognitive, pragmatic, and other factors, and that there is not a qualitative difference between how a hearer processes and interprets literal and figurative meanings. Accordingly, researchers would benefit by adopting a dynamical systems approach towards understanding figurative language; one which recognizes the regularities and instabilities in research findings surrounding figurative language as representative of the dynamic nature that a wide variety of variables exert on figurative language use (Gibbs & Colston, 2012). Currently, there are a number of studies that have considered the linguistic, cognitive, and pragmatic influences on figurative language processing beyond figurative and literal meaning, and their results cohere to provide some evidence in favor of a dynamic approach to understanding figurative language use.

Additionally, research has revealed that the linguistic characteristics of figurative language can influence how figurative language is processed. For example, highly formulaic instances of figurative language (e.g., familiar metaphors and ironies) result in similar processing speeds for figurative and literal utterances with approximately the same meaning, suggesting that the figurative meaning is immediately available for more familiar examples of metaphors and ironies (Cronk & Schweigert, 1992; Giora & Fein, 1999; Keysar, 1989). Additionally, the surrounding linguistic contexts, such as negation, positive or negative quantifiers, and previous uses of figurative language in the same text also influence the speed of figurative language processing when compared to literal counterparts (Filik & Moxey, 2010; Giora, 2016; Giora et al., 2007, 2013; Giora, Fein, & Schwartz, 1998). Other studies have also demonstrated that some examples of metaphor, verbal irony, and satire employ extremely infrequent language patterns
(e.g., word and grammatical combinations) when compared to a generalized corpus of language (Deignan, 1999a, 1999b; Goatly, 2012, 2017; Partington, 2007, 2011; Philip, 2011; Skalicky, in press), which may in turn influence the speed of processing and level of comprehension. Finally, computationally derived linguistic features related to lexical sophistication (i.e., complexity and diversity of words) and cohesion (i.e., degree of semantic overlap) have been used to successfully distinguish figurative from non-figurative language (Gibbs & Colston, 2012; Skalicky & Crossley, 2015).

As a whole, these studies suggest that the linguistic characteristics of figurative language, such as familiarity of use, deviation from expected linguistic patterns, and other features related to lexical sophistication and semantic cohesion may influence the manner in which figurative language is processed and understood. However, many of these variables are not typically controlled for in psycholinguistic experiments testing figurative language processing or production. Furthermore, relatively few studies have considered the role of lexical sophistication and semantic cohesion during figurative language production and processing, despite these features exerting large influences on non-figurative language comprehension. Accordingly, in this dissertation, the role of lexical sophistication and semantic cohesion is considered during both the production and processing of figurative language.

In addition to linguistic features of figurative language, other research has illustrated links between different individual differences (i.e., different cognitive and other traits such as intelligence and age) and figurative language production and comprehension ability. For instance, studies have found associations between different measures of intelligence (e.g., problem solving ability or broad memory retrieval) and the ability to produce creative and original metaphors (Beaty & Silvia, 2013). Working memory capacity (i.e., the ability to solve
problems while holding information in one’s short term memory) has been linked to metaphor production (Chiappe & Chiappe, 2007; Pierce & Chiappe, 2008) and also to metaphor and sarcasm processing ability (Kaakinen, Olkoniemi, Kinnari, & Hyönä, 2014; Olkoniemi, Ranta, & Kaakinen, 2016). Furthermore, exposure to sarcasm has been associated with abstract thought (Huang, Gino, & Galinsky, 2015), and a preference to perform difficult cognitive tasks has been linked to the ability to process metaphors (Olkoniemi et al., 2016). Finally, age and general levels of background knowledge are significant predictors of satire comprehension (Boukes, Boomgaarden, Moorman, & de Vreese, 2015; Pfaff & Gibbs, 1997; Simpson, 2003; Skalicky & Crossley, in press). The wide number of connections among different individual differences and the ability to produce and comprehend different types of figurative language reported in these studies highlights the importance of taking multiple factors related to the individual into account when investigating figurative language use. However, many studies only test one or two of these variables for only one or two types of figurative language. In both the first and second study of this dissertation, a wide range of participant individual differences are measured using surveys and tests, allowing for a holistic consideration of many different individual differences during the production or processing of metaphor, sarcasm, and satire.

Language background (and proficiency) is another important individual difference that affects the ability to produce and process figurative language. Specifically, figurative language is processed differently between one’s first and second languages, and also between bilinguals and monolinguals (Golden, 2010; Heredia & Cieślicka, 2015; J. Johnson & Rosano, 1993; Titone, Columbus, Whitford, Mercier, & Libben, 2015; Vaid, López, & Martínez, 2015). Moreover, one’s aptitude for language learning is strongly related to one’s language background. Language aptitude measures the ability to learn new languages by measuring one’s ability to process and
comprehend different features of language (e.g., vocabulary, grammar) both explicitly (i.e., analytic skills) and implicitly (i.e., inductive learning) and therefore may also be predictive of figurative language production and comprehension ability (Dörnyei & Skehan, 2003; Li, 2015). However, the majority of published figurative language studies present stimuli only in participants’ first languages and without measuring the additional language knowledge of participants. Because bilinguals and monolinguals process non-figurative language differently (Tokowicz, 2014), they also may process and produce figurative language differently, and it is thus crucial to include language knowledge as a variable when examining figurative language processing and production. Therefore, in addition to the individual differences mentioned above, this dissertation collects data from both native and non-native English speaking participants and also measures their language knowledge using surveys in order to develop a detailed language profile (i.e., number of languages known, proficiency with those languages, and contexts where those languages are used).

In addition to the influence of linguistic features and individual differences, pragmatic and social variables are also important to take into consideration when investigating figurative language use. Figurative language is used to purposefully impart specific pragmatic effects, such as saving face, promoting humor, demonstrating creativity, and fostering group solidarity (Dews & Winner, 1995; Gerrig & Gibbs, 1988; Gibbs, 2000; Jorgensen, 1996; Roberts & Kreuz, 1994). Moreover, metaphorical, metonymic, and ironic language take on different meanings when used in different genres and registers (Burgers et al., 2012; Deignan, Littlemore, & Semino, 2013). Other studies have demonstrated that manipulation of the pragmatic context by supplying more or less information regarding the purpose or intent of figurative language results in the figurative meaning being directly processed from examples of sarcasm (Gibbs, 1986; Ivanko & Pexman,
2003) and satire (Pfaff & Gibbs, 1997). In addition, the ability to empathize with others by recognizing and processing emotions aids in figurative meaning comprehension (Akimoto et al., 2014; Akimoto, Miyazawa, & Muramoto, 2012; Nicholson, Whalen, & Pexman, 2013), as do perceptions of group membership and humor (Boukes et al., 2015; LaMarre, Landreville, & Beam, 2009; Lee & Kwak, 2014; Skalicky & Crossley, in press). These studies suggest that pragmatic meaning is an important component of figurative meaning, yet the vast majority of studies investigating figurative language processing do not typically take pragmatic meaning into consideration (Gibbs & Colston, 2012). In order to address this gap, the second study in this dissertation will consider the influence of pragmatic information on satire processing by capturing affective data related to satire (i.e., perceptions of humor, sincerity, and positivity) and genre knowledge (i.e., how often do they read news satire).

Finally, the field of figurative language research could further benefit from methodological diversity and improvements related to stimuli and comprehension measures. Despite the large number of published studies investigating figurative language, there still exist types of figurative language that have not been scientifically studied. This is most evident in the research on verbal irony, which has mainly focused on sarcasm. However, many other types of verbal irony exist that differ from sarcasm (Gibbs & Colston, 2007), and a full investigation of verbal irony should consider these other subtypes. One specific example is satire, a more subtle form of verbal irony used by authors to subtly criticize a satirical target, most commonly political or culturally related. Despite satire’s ubiquitous presence in film, newspaper, and television (Simpson, 2003), it has undergone very little linguistic and scientific investigation (cf. Pfaff & Gibbs, 1997; Skalicky, in press; Skalicky & Crossley, 2015, in press). Additionally, much of the stimuli used during processing experiments is comprised of invented examples on the part of the
researchers, calling into question the naturalistic quality of the stimuli (Gibbs & Colston, 2012). In this dissertation, authentic examples of satirical text taken from The Onion are used as stimuli in a processing experiment in order to address these gaps.

In tandem with investigating a wider range of figurative language types, better comprehension measures should also be employed that provide a more accurate understanding of the meaning a hearer obtains from an example of figurative language. In addition to the semantic meaning of the message, this also includes a hearer’s ability to determine authorial intent, pragmatic force, and affective responses (e.g., humor, irritation) produced after processing examples of figurative language (Colston & Katz, 2005; Gibbs & Colston, 2012; Pfaff & Gibbs, 1997). In this dissertation, a wider range of comprehension measures are collected from participants during the processing experiment in order to address these concerns.

This brief review highlights the need to consider factors beyond literal and figurative meaning when testing how hearers understand and process figurative language (i.e., cognitive, linguistic, and pragmatic influences). It further touches upon the need for methodological diversity and improvement (i.e., better comprehension measures), all of which aids in adopting a dynamic approach to the study of figurative language. While attempting to account for all of the features that affect figurative language production and comprehension during a single scientific study is difficult, a combination of studies measuring a large number of variables from a wide variety of participants can provide a more holistic understanding of figurative language use when compared to previous studies focusing primarily on the comprehension of figurative or literal meanings.
1.1 Current Study

Through a series of experimental studies, this dissertation will address many of the topics outlined above in regards to how different types of figurative language (i.e., metaphor, sarcasm, and satire) interact with linguistic, cognitive, and pragmatic features during figurative language production, which is generally understudied, as well as figurative language processing and comprehension. In doing so, the role of individual differences, language background, pragmatic context, and affective perceptions during the processing, comprehension, and production of figurative language will be considered from a wide variety of participants, previously unexplored lexical features (i.e., lexical sophistication and cohesion) will be examined, and a currently understudied form of figurative language (i.e., satire) will be investigated.

Two main approaches will guide the focus of this dissertation. First, individual differences and their relation to figurative language processing, comprehension, and production will be examined in order to examine the roles individual features of participants play during figurative language use. Secondly, the linguistic features of figurative language will be analyzed for lexical sophistication and cohesion in order to investigate whether lexical features beyond the literal or figurative meaning of figurative language play any meaningful roles during the processing and production of figurative language. These two questions will be answered through three separate studies that also take into account calls for methodological innovation, including considering features of meaning comprehension beyond literal and semantic interpretations of figurative language and investigating a previously underexplored type of figurative language (i.e., satire). The following research questions guide this dissertation.

1. What individual differences (including language background), if any, influence the production, processing, and comprehension of figurative language?
In order to fully explore answers to these research questions, this dissertation reports the results from three original experimental studies. Study 1 investigates the roles of individual differences and language background during the production of figurative language (specifically, metaphor and sarcasm). To do so, figurative language production data from 60 participants (30 native English speakers and 30 non-native English speakers) is analyzed for several features by human raters in order to rank the quality of figurative language produced by the participants. At the same time, a large number of participant and linguistic variables are measured and recorded, including demographic features (e.g., age, country of birth), cognitive traits (e.g., intelligence, working memory capacity), and language background (e.g., first language, English proficiency). Additionally, the participant output is measured for linguistic features related to lexical sophistication and semantic cohesion. Statistical analyses are then performed to determine whether any of these variables are significantly associated with human ratings of figurative language production.

The following research questions guide Study 1:

1. What role, if any, do individual differences such as demographic features, cognitive ability, and language background play in the production of metaphor and sarcasm?
2. Can human ratings of the quality of figurative language production be predicted using linguistic features of lexical sophistication and semantic cohesion?

Study 2 examines the online processing and comprehension of an underexplored type of figurative language (i.e., satire). In order to do so, reading times and comprehension responses from the same 60 participants from Study 1 are collected as they read different examples of
satirical and non-satirical texts. Participants are also asked to provide the intended meaning of the texts, and their answers are measured for the presence or absence of satirical message comprehension. Additionally, the satirical texts are measured for linguistic features related to lexical sophistication and semantic cohesion. As in Study 1, statistical analyses are then performed to determine if any significant relations exist among participant individual differences, language background, textual linguistic features, and satirical processing and meaning comprehension.

The following research questions guide Study 2:

1. Are there significant differences in affective perceptions between satirical and non-satirical texts?

2. Do individual differences, affective perceptions, and measures of lexical sophistication and semantic cohesion influence the processing of satirical texts?

3. Do individual differences, affective perceptions, measures of lexical sophistication and semantic cohesion, and processing time interact with satirical meaning comprehension?

Study 3 employs an online task in order to investigate the influence of exposure to figurative language on perceptions of creativity during a creative production task. In order to do so, a large population of online workers are recruited to produce creative and engaging reactions to two different news story headlines. After providing their initial answer, the workers then compare their answers to pre-constructed responses that include differing levels of figurative language (specifically, metaphor and sarcasm) and rate whether they felt their response was more, less, or equally as creative as the pre-designed responses. They then complete the task a second time for the second headline.

The following research question guides Study 3:
1. What effect (if any) does the presence or absence of creative language have on perceptions of creativity?

This dissertation is organized as follows. First, research related to the scientific study of figurative language use is reviewed in more detail in Chapter 2. Then, the methods, analyses, and results for Study 1, Study 2, and Study 3 are presented in Chapters 3, 4, and 5, respectively. Chapter 6 then presents a summary of the findings, answers the main dissertation research questions, and discusses their implications for figurative language research.

2 LITERATURE REVIEW

Figurative language and thought are tightly linked and ever-present in human communication. Scholars from a wide range of academic disciplines including literature, rhetoric and composition, cognitive linguistics, computational linguistics, humor studies, and pragmatics have collectively produced an overwhelming amount of literature regarding this topic. Figurative language’s ability to capture the attention of many different scholars in different disciplines is a testament to its compelling yet enigmatic nature as an object of study. Indeed, even a task seemingly as essential as providing definitions for different types of figurative language differs vastly among disciplines and is still hotly debated today within certain fields (e.g., cognitive linguistics). As such, any review of figurative language research must be clear regarding the theoretical and empirical foundation motivating the review.

The purpose of this review is to first define the three types of figurative language studied in this dissertation before further describing the research gaps introduced in Chapter 1. Because this dissertation is motivated by work in the fields of cognitive linguistics, pragmatics, and natural language processing, and focuses specifically on metaphor, sarcasm, and satire, scholarship from other disciplines (e.g., literature) and other types of figurative language (e.g.,
idioms) are not reviewed. This review first defines metaphor, sarcasm, and satire before covering studies that have explored how individual differences, linguistics features, and pragmatic functions influence metaphor and sarcasm production and satire processing.

2.1 Types of Figurative Language

2.1.1 Metaphor and Simile.

Metaphor is the most widely studied form of figurative language (Gibbs & Colston, 2012), and the one most closely associated with language and thought (Gibbs, 1994; Glucksberg, 2001; Tay, 2015). Conceptual metaphors, or the “understanding and experiencing of one kind of thing in terms of another” (Lakoff & Johnson, 2008, p. 5, emphasis original) are said to be not just an example of linguistic creativity, but are reflective of fundamental processes behind human thought (Gibbs, 1994, 2011; Lakoff & Johnson, 2008). Specifically, based on work in cognitive linguistics, researchers have argued that conceptual metaphor is representative of the system employed by the human mind to categorize and understand different entities in the world (Gibbs, 2011; Kövecses, 2002; Steen, 2015; Tay, 2015). For example, exposure to conceptual metaphors has been shown to affect choices people make when asked to solve problems that require complex reasoning (Thibodeau & Boroditsky, 2011, 2013, 2015) and also prove persuasive when used during political propaganda (Charteris-Black, 2011).

Conceptual metaphors are usually described using an A is B formula, where a target entity A is categorized differently by placing it into a different conceptual domain B (Gibbs, 2011; Kövecses, 2002; Lakoff & Johnson, 2008). This re-categorization is completed when some, but not necessarily all elements of the source domain (B) are transferred to the target (A), which provides new ways of interpreting and understanding the target entity (Gibbs, 2011). For instance, in the conceptual metaphor lawyers are sharks, abstract elements of sharks (e.g.,
predatory, dangerous) are mapped onto concrete understandings of lawyers, so that the basic understanding of a lawyer (i.e., a person trained in the law) is adjusted to now also include new concepts related to sharks (i.e., a person trained in the law who is also ruthless and predatory; Gibbs, 1994; Glucksberg, 2001). Another example of a conceptual metaphor is the argument is a war metaphor, which represents the process of having an argument as a series of battles, where people win, lose, or draw (Kövecses, 2002; Lakoff & Johnson, 2008). For example, one can defend their point of view or concede the argument, showcasing the adversarial nature of arguments. Conceptual metaphor is thus an incredibly efficient way to communicate meaning because it allows for the expression of difficult concepts in a relatively concise yet vivid manner (Gibbs, 1994).

Closely related to the notion of conceptual metaphor are similes, which compare two different entities using a fixed formula of A is like B (e.g., lawyers are like sharks). Some theories of metaphor comprehension argue that all metaphors are automatically reinterpreted as similes (i.e., the Comparison Theory; Gibbs, 1994; Glucksberg, 2001), but this theory has failed to account for several key differences between metaphor and simile, namely that while similes prompt a reconsideration of one entity in light of another, they do not complete the re-categorization process attributed to conceptual metaphor (i.e., similes only prompt a comparison between two things at the concrete, but not conceptual level; Haught, 2013). Additionally, it is comparatively easy to find some similarities between any two entities included in a simile, but conceptual metaphors require that the domain mappings from a source to a target are apt and meaningful (Gibbs, 1994; Glucksberg, 2001, 2003; Haught, 2013). That being said, in actual use, metaphor and simile are similar because the reason speakers use either one is to explain one thing in terms of another (Chiappe & Chiappe, 2007; Pierce & Chiappe, 2008). Because this
dissertation is not focused on further investigating the definitions of metaphor or simile, nor theories related to metaphor interpretation, metaphors and similes are treated as approximately similar linguistic phenomena that both serve to describe one thing in terms of another.

2.1.2 Verbal Irony and Sarcasm.

Verbal irony is another widely investigated form of figurative language (Gibbs & Colston, 2007). Theoretical definitions of verbal irony have been in a state of constant flux since the oppositional theory of irony first defined irony as an utterance which means the opposite of its literal form (Colston, 2017; Colston & Gibbs, 2007; Gibbs & Colston, 2007). For instance, if a bleary-eyed office worker arrived unenthusiastically to work on a Monday morning, she might quip “I love Mondays.” The oppositional account of verbal irony defines this utterance as ironic because it is clear that the speaker actually means the opposite of what she says (i.e., that she does not in fact like Mondays). However, this definition of irony is unable to account for ironic utterances that cannot be simply reinterpreted as the opposite of their literal meaning. For instance, a frustrated teacher exclaiming “I love it when you’re on time” when a student comes late to class does not actually mean the teacher enjoys it when the student is late, but the utterance is nonetheless ironic in that situation because of the opposition between what was said and the context of the situation.

As such, the definition of verbal irony has been refined by scholars over the past several decades. The most influential theoretical definition of verbal irony, known as the Standard Pragmatic Model (SPM), was a result of Grice’s work on conversational maxims and pragmatic implicature (Grice, 1975, 1978). According to this model, ironic meaning is the result of a reinterpretation of an ironic utterance based on the pragmatic context within which it is stated. To understand the ironic meaning, a hearer must first interpret the literal meaning of an ironic
utterance, reject the literal meaning based on contextual information, and then compile a new, ironic meaning, based on their understanding of the speaker’s pragmatic intent and the surrounding context (Colston, 2017). A classic example used to describe how this process unfolds is the phrase “what lovely weather” spoken by someone during a pouring thunderstorm. According to the SPM, a hearer must first decode the speaker’s meaning, reject the meaning (because it is stated during a thunderstorm), recognize that the speaker is purposefully violating Gricean maxims, and finally reinterpret the sentence figuratively (i.e., that the speaker is expressing their distaste for the current weather).

The SPM has been criticized and revised over the years, primarily because of its prediction that hearers must always initially comprehend the literal meaning of an ironic utterance before arriving at the ironic meaning (Colston, 2017; Gibbs, 1986, 2002; Giora, 2003). Various refinements of the SPM include use-mention theories, which posit that ironic utterances involve a mention or echoing of some previous belief, value, or experience, that may or may not be attributed to the speaker themselves (Sperber & Wilson, 1981; Wilson & Sperber, 1992). Pretense theories of irony claim that a speaker is pretending to be naïve when producing ironic utterances (Clark & Gerrig, 1984), whereas allusional pretense theories combine elements from use-mention and pretense theories into a single theory (Kumon-Nakamura, Glucksberg, & Brown, 1995). More recently, the relevant inappropriateness theory of verbal irony asserts an utterance is ironic if it is both inappropriate yet relevant in a particular context (Attardo, 2000). The irony as clash theory simplifies matters by stating that all ironic utterances simply contain some form of clash between meanings, and that all the aforementioned theories of verbal irony share this perspective (Garmendia, 2014). Finally, a recent review of verbal irony and its related
theories provided an inclusion definition which states that verbal irony is an utterance which expresses meaning that is somehow contrary to reality (Colston, 2017).

This brief review of the theoretical debate surrounding the definition of verbal irony demonstrates that it is difficult to locate an all-encompassing definition with any measure of specificity. The difficulty associated with locating an inclusive definition of verbal irony may in part be due to its status as an umbrella term, suggesting that verbal irony is best conceptualized as a family of related language phenomena, such as hyperbole, understatement, and rhetorical questions (Colston, 2017; Gibbs, 2000; Gibbs & Colston, 2007). This difficulty is exacerbated because many studies of verbal irony only focus on one specific type of verbal irony: sarcasm.

Sarcasm is a most commonly considered a negative form of verbal irony which expresses a negative sentiment from the perspective of a speaker, and may or may not be used to enhance or mitigate criticism of others (Bowes & Katz, 2011; Colston, 1997, 2017; Dews & Winner, 1995). Indeed, the examples of verbal irony stated above (e.g., “I love Mondays” and “what lovely weather”) are examples of sarcasm, as they express a negative sentiment from the speaker. There have been several attempts to define the required linguistic and social components of a sarcastic utterance, including presence of a victim, facial and paralinguistic cues, tone of voice, and specific linguistic features (Campbell & Katz, 2012; Caucci & Kreuz, 2012; Colston, 2000; Kreuz & Roberts, 1995). Results from these studies suggest that while all of these features are capable of inducing sarcastic meaning, none of them are required. Thus, for the purposes of this dissertation, the definition of sarcasm is retained as a specific type of verbal irony designed to express a typically (but not always) negative point of view towards another entity.
2.1.3 Satire.

In contrast to metaphor and sarcasm, satire is relatively understudied, especially from cognitive and linguistic perspectives (Simpson, 2003). Satire is typically defined as a literary genre, where an author attempts to critique a social or moral wrong they perceive to exist in the world (A. Nilsen & Nilsen, 2008; Phiddian, 2013; Simpson, 2003). However, satire’s treatment as a literary genre has led to a dearth of empirical, scientific investigations of satire, especially when compared to other examples of figurative language (Simpson, 2003). Furthermore, because literary investigations of satire tend to focus on variations in technique among authors, they are of little use to the current investigation and will not be considered further.

Instead, the definition of satire used in this dissertation is based largely on a book-length treatment of the topic provided by Simpson (2003), wherein he defines satire as a discursive practice that uses verbal irony to establish and resolve an incongruity between a satirical target, a satirical author, and a satirical audience in order to critique or mock a satirical target, commonly resulting in humor between the satirical author and audience. According to Simpson (2003), satire comprehension involves three stages: a prime, a dialectic, and an uptake. The prime stage is an echoic reference to a type of discourse, register, or genre event that may or may not be fully real. The dialectic phase is where an internal element of the text conflicts in some way with the prime and thus signals an ironic discrepancy between the expectations formed based on a hearer’s knowledge of the prime (e.g., the genre of political news). During the uptake phase, the hearer must notice and understand that the satirical text is projecting a satirical, rather than literal, meaning. Successful understanding of the satirical meaning typically results in humor on the part of the audience. In sum, satire is a discursive practice that relies on subtle forms verbal irony in order to critique a specific (and most likely political) satirical target (A. Johnson,
Examples of contemporary American satire include satirical television and print news outlets, such as *The Daily Show*, *The Colbert Report*, *Last Week Tonight*, *The Onion*, and *The Borowitz Report*. All of these examples are similar because they masquerade as legitimate news sources but then present satirical spins on the news (or produce fictional news in some cases), although they range in regards to how obvious their satirical purpose is. For example, Stephen Colbert’s character on *The Colbert Report* satirized conservative talk show host Bill O’Reilly by appropriating O’Reilly’s mannerisms and political perspective in order to subtly criticize conservative politics in the United States, whereas hosts of *The Daily Show* are comparatively more overt when delivering their humorous and critical messages. Examples of satire used in this dissertation come from *The Onion*, one of the oldest and most well known satirical newspapers in the United States. *The Onion* regularly publishes news stories and headlines on its website and social media (e.g., Facebook and Twitter). Satirical news stories in *The Onion* range from subtle criticism of politics and others aspects of American culture (e.g., the headlines “Drone Places Fresh Kill on Steps of White House” and “Obama Not Ruling Out U.S. Military Action In Congress” serve to criticize the United States government), wordplay and ambiguity that taps into American stereotypes (e.g., the headline “Black Guy Asks Nation for Change” during the 2008 U.S. presidential election between Barack Obama and John McCain), and reporting of otherwise mundane events as newsworthy (e.g., the headline “Area Man Mystified By Layout Of Adjacent Town’s Kroger” presented as newsworthy).
2.2 Figurative Language Research

As mentioned in Chapter 1, the vast majority of linguistic research investigating figurative language has focused on developing and testing models of figurative language processing. In general, these studies compare reading times for literal and figurative utterances that are thought to represent similar meanings, and were prompted by the claims made by the SPM (i.e., that figurative meaning interpretation is a two-stage process that requires an initial comprehension of an utterance’s literal meaning before arriving at the intended figurative meaning). Thus, the SPM would predict that figurative utterances should take longer to read than matched literal utterances because two meanings (i.e., the literal and figurative) must be identified rather than one (i.e., just the literal).

2.2.1 Metaphor processing.

With regards to metaphor, psycholinguistic research has demonstrated conventional (i.e., familiar and commonly used) and novel (i.e., uniquely coined and infrequently encountered) metaphors take no longer to process when compared to literal counterparts (Blasko, 1999; Gibbs & Colston, 2012; Glucksberg, 2003; McElree & Nordlie, 1999). Additionally, metaphorical meaning is activated even when metaphors are placed in literal biasing contexts (Keysar, 1989), demonstrating that metaphorical meaning is primary and interpreted first when hearers encounter metaphorical utterances. Keysar (1989) interpreted this “metaphor interference effect” (p. 376) as evidence that the metaphorical meaning of statements is readily available regardless of context, providing additional evidence that the literal meaning of a metaphorical utterance does not need to be interpreted before the figurative meaning of a metaphorical utterance is obtained. Based on these findings, Keysar (1989) argued the processing of metaphors was fundamentally
no different from the processing of literal language, because the meaning congruent to a particular context was the meaning participants arrived at the quickest.

The unique features of metaphor may account for the apparent ease in processing of metaphorical meaning. As mentioned above, researchers in cognitive linguistics posit that metaphorical comparisons are representative of how the human mind conceptualizes the world (Gibbs, 2011; Glucksberg, 2001, 2003; Lakoff & Johnson, 2008), and therefore the default state of thought may in fact be metaphorical, allowing for effortless interpretation of figurative meaning from metaphorical utterances. Additionally, many metaphors are so entrenched in language they have been described as “dead” (Veale, 2012b, p. 7), suggesting that whatever literal meaning used to exist in a particular metaphorical phrase is no longer the primary meaning (e.g., describing the body of a wine; Veale, 2012b). Regardless of the reasons why, the majority of research suggests that hearers are able to understand metaphorical meaning directly from metaphors without needing to first understand the literal meaning (Glucksberg, 2003).

2.2.2 Verbal irony processing.

With regards to verbal irony processing, the research is not as clear. Based on psycholinguistic experiments of ironic language processing, two main alternative theories have been produced: a direct access view and a salience first view. The direct access view is based on Gibbs (1986) and other studies demonstrating no significant differences in reading times for ironic utterances and their literal counterparts, and also that manipulations made to the context of an utterance significantly predicted comprehension and reading times. Specifically, Gibbs (1986) found that as the surrounding linguistic contexts were changed to bias ironic interpretations of sarcastic statements, participant readings times became quicker, as well as their ability to describe and remember the utterances as ironic. Gibbs (1986) argued the contextual information
present in his examples allowed for the direct processing of ironic meaning, and that the probabilistic incorporation of contextual information during reading is how any meaning of any utterance, literal or figurative, is obtained. In other words, the Direct Access view does not assume any single utterance contains an inherent, decontextualized meaning, and that the interpretation of figurative meaning requires processes no different than the interpretation of literal meaning.

The assumptions put forth by the SPM were also called into question by Giora (1997), who believed the divide between literal and figurative meaning overlooked the actual linguistic mechanism responsible for differences in processing time: the role of salience. Initially proposed to explain the processing of metaphor, Giora (1997) argued the salient meanings of a phrase or utterance are always processed first, regardless of contextual influence. This is because salient meanings are frequent, familiar, conventional, and thus represent prototypical or standardized meanings (Giora, 2003). Importantly, these features are graded, in that their relative strengths determine how salient a particular meaning is to a particular person. Salient meanings are “coded in the mental lexicon” (p. 15), allowing for quicker access and retrieval of salient meanings when compared to novel meanings that must be computed online.

Salient meanings are not specific to literal or figurative interpretations, and are so strongly activated they can mitigate the effects of prior context biasing a non-salient meaning. This feature, according to Giora (1997), is what explains the results of research showing no significant differences between processing times for metaphor (e.g., Keysar, 1989), because metaphorical meanings are encoded into the lexicon. When hearers encounter figurative language that represents a non-salient meaning, they must first interpret the literal meaning of the utterance in order to determine the figurative meaning. In this manner, Giora’s Graded Salience
Hypothesis (GSH) subsumes both the Direct Access and Standard Pragmatic Model of figurative language processing. When meanings are salient, they can be directly processed without needing to consider alternative meanings (i.e., Direct Access view) because they are readily available in the lexicon. When meanings are not salient, the literal meaning is first interpreted, rejected, and ultimately replaced with the intended figurative meaning, derived from contextual clues (i.e., the Standard Pragmatic Model).

Giora and colleagues have published a number of studies finding continued support for the GSH while also responding to criticisms of methodology and stimuli (e.g., Gibbs, 2002). For instance, Giora and Fein (1999) tested reading times for familiar and less familiar ironies and found that reading times for familiar ironies were significantly faster than less familiar ironies and also not significantly different from their literal counterparts, suggesting that salience of familiar ironies allowed the figurative meaning to be directly accessed when compared to less familiar ironies. In another study, Giora, et al. (2007) tested whether encountering verbal irony in a particular stimuli would facilitate subsequent processing of verbal irony, but found that examples of irony were always read slower than their literal counterparts, regardless of the level of preceding irony. These results were replicated by Fein, Yeari, and Giora (2015).

These researchers have also published findings that do not fully support the predictions of the GSH. Specifically, Giora, Drucker, Fein, & Mendelson (2015) reported the results of a study that demonstrated negative constructions, such as *punctuality is not his forte*, were interpreted as sarcastic verbal irony by default, regardless of surrounding contexts, and argued that neither the GSH nor Direct Access view could account for these findings, because sarcastic verbal irony should not be the default, salient meaning associated with these utterances. Giora et al. (2015) questioned whether this provided evidence of constructional knowledge being highly associated
with figurative meaning for negative constructions, explaining why the figurative meaning was able to be accessed directly.

Despite a large number of studies demonstrating slower processing times for ironic utterances compared to their literal counterparts from Giora and other researchers, which align with the predictions of the GSH (Akimoto et al., 2012; Filik et al., 2014; Filik & Moxey, 2010; Giora et al., 1998; Kaakinen et al., 2014; Pexman, Ferretti, & Katz, 2000; Schwoebel, Dews, Winner, & Srinivas, 2000), proponents of the Direct Access view maintain that figurative meaning is not processed in a qualitatively different manner than literal meaning, and thus reject the GSH because it includes a component that states figurative processing resorts to the SPM for unfamiliar and non-salient examples of verbal irony. In a clarification of the Direct Access view, Gibbs (2002) made it clear that the Direct Access view rejects the SPM because the full literal meaning of an utterance does not need to be interpreted before arriving at figurative meaning. Instead, as utterances are read, multiple possible meanings compete for selection, and this process is heavily influenced by the linguistic and pragmatic context of an utterance. Furthermore, differences in reaction times from psycholinguistic experiments are not sufficient evidence of qualitatively different processes of comprehension between literal and figurative meaning, but rather evidence that arriving at a figurative meaning may just take longer. In other words, the same reading comprehension processes occur when computing literal and figurative meaning, except that one instance of language requires more processing time than the other (but is still processed in the same manner). Accordingly, the Direct Access view questions the very nature of the distinction between literal and figurative meaning, arguing that all meaning is computed from language using a similar process, and that the brain does not treat figurative or

In sum, the linguistic research associated with verbal irony processing has dominated current research approaches in figurative language research. Despite many studies providing evidence that verbal irony takes longer to process than literal counterparts, there also exists compelling evidence that much of this research is misguided due to an overemphasis on the distinction between figurative and literal meanings, which may not even exist (see Gibbs & Colston, 2012).

2.2.3 Satire processing.

Unlike metaphor and sarcasm, there are very few psycholinguistic studies testing satire processing (cf. Skalicky & Crossley, in press). Moreover, while there does exist a theoretical model of satire comprehension (Simpson, 2003), it has not been empirically validated. Because Simpson’s (2003) model of satire comprehension posits that satirical meaning is partially caused through ironic opposition, the formalized models of verbal irony comprehension may apply to satire processing as well, but this is also yet to be tested. Aside from Simpson’s model and its potential connections to verbal irony processing, other research has demonstrated that satire comprehension is dependent upon specific satirical strategies, a hearer’s background knowledge, agreement or disagreement with the satirical message, demographic features such as age, and the ability to recognize that an author is attempting to be satirical (Becker, 2014; Boukes et al., 2015; A. Johnson et al., 2010; LaMarre et al., 2009, 2014; Pfaff & Gibbs, 1997; Skalicky & Crossley, in press). Accordingly, investigations of satire processing would benefit from considering these variables in their experimental design.
2.3 Variables Beyond Literal and Figurative Meaning

As can be gleaned from this review so far, it is unclear whether separate models of text processing are needed to specifically account for figurative language processing. Yet, contemporary studies of figurative language processing still aim to test figurative versus literal utterances, despite claims that they may be misguided in assuming slower processing time is indicative of qualitatively different processing mechanisms (Blasko, 1999; Colston, 2017; Gibbs, 1994, 2002; Gibbs & Colston, 2012). This is just one of several problems identified with contemporary research into figurative language outlined in a recent book-length review of the state of the field (Gibbs & Colston, 2012). Other potential problems with the manner in which figurative language is typically investigated include the use of invented examples as stimuli, a lack of comprehensive assessment measures testing figurative comprehension, and the isolation of processing mechanisms from other contextual factors that influence figurative language comprehension (Colston, 2015; Gibbs & Colston, 2012). Additionally, the focus on examining figurative language processing has sidelined studies investigating figurative language production (Benedek et al., 2014; Corts & Meyers, 2002; J. Johnson & Rosano, 1993).

As mentioned in Chapter 1, the purpose of this dissertation is to investigate the influence of features beyond literal and figurative meaning that may shape figurative language production and comprehension ability by considering individual differences, linguistic features, and pragmatic and social-contextual variables. A number of studies have highlighted how these different contextual factors influence figurative language use and are reviewed in the following section. Although these sections are presented as separate influences, there is also overlap among these features. For example, levels of background knowledge can be related to both a sub-type of intelligence (i.e., an individual difference) as well as social identity (i.e., a pragmatic and social-
contextual variable). Therefore, it is important to keep in mind these features are presented in isolation as a matter of organization, and the purpose is not to assume they are separate constructs that do not interact with one another.

2.4 Individual Differences and Figurative Language

Individual differences refer to the many different features that comprise the unique nature of every person. For example, people differ based on intelligence, age, language background, and a seemingly limitless number of other features (Chamorro-Premuzic, Von Stumm, & Furnham, 2011). While it is difficult if not impossible to control and measure every conceivable individual difference and how they may or may not affect figurative language use (Gibbs & Colston, 2012), a number of different studies have directly or indirectly investigated how a wide range of individual differences influence the production and processing of figurative language. This section focuses on three types of individual differences: cognitive measures, demographic measures, and language background.

2.5 Cognitive Measures

Cognitive measures are those that relate specifically to cognitive functions such as attention, memory, and intelligence, and are thus closely related to one’s ability to process, store, and retrieve information (Gruszka, Matthews, & Szymura, 2010). Because these features are related to language processing in general (Tokowicz, 2014), it follows that they may also influence the ability to produce and process examples of figurative language. Although there is still a need for more studies exploring the role of cognitive differences in figurative language use, there has been some work investigating intelligence, working memory capacity, need for cognition, and abstract thinking in the context of figurative language use.
2.5.1 Intelligence.

Intelligence is one of the most common cognitive differences that researchers in any field concerned with individual differences are interested in testing. Specifically, the Cattell-Horn-Carroll model of intelligence divides the construct of intelligence into several different sub-types of intelligence. For instance, fluid intelligence measures problem-solving ability, crystallized intelligence measures generalized background and vocabulary knowledge, and broad retrieval refers to the ability to retrieve information from long term memory (Flanagan, 2013). These different sub-types of intelligence can be measured through the use of tests or surveys that ask participants to locate patterns and solve problems, generate words based on prompts (e.g., list as many words as you can that start with m), and complete standardized tests of learned knowledge.

All three of these intelligence sub-types were demonstrated to influence the ability to produce both conventionalized (i.e., commonly used) and novel (i.e., original) metaphors in two studies (Beaty & Silvia, 2013; Silvia & Beaty, 2012). In a study designed to locate potential connections between intelligence and creativity, Silvia and Beaty (2012) measured participants’ fluid intelligence before asking them to create two novel metaphors related to describing the most boring class they had even been in and the most disgusting food item they had ever eaten. After rating the participant metaphors for creativity using an analytic rubric, Silvia and Beaty (2012) found that participants’ fluid intelligence scores explained 24% of the variance in creativity scores of the metaphors, suggesting that higher fluid intelligence allowed for more creative metaphors. The authors also reported that participants who spent longer thinking about their metaphors produced more creative metaphors, whereas those who spent less time appeared to be relying on past memories of conventional metaphors. Their results thus suggest that metaphor production is strongly associated with inductive problem-solving and reasoning ability.
In a follow-up study, Beaty and Silvia (2013) measured participants’ fluid intelligence, crystallized intelligence, and broad retrieval ability before asking participants to create both novel and conventional metaphors. The conventional metaphors were prompted by providing participants with the framework to construct clichéd metaphors, such as describing life as a series of stages to prompt the metaphor \textit{life is a journey}, whereas the novel metaphors used the same prompts from Silvia and Beaty (2012). The metaphors were again rated for creativity in a similar fashion to the previous study. The authors then used structural equation modeling to identify which of the three intelligence measures (if any) associated with creativity scores. Their results reported strong relations between fluid intelligence and broad retrieval for the creative metaphors and a moderate relation between broad retrieval ability and conventional metaphors, suggesting that different sub-types of intelligence associate with production ability for different types of metaphors. Specifically, creative metaphor production relied on participants’ inductive problem-solving ability as well as access to long term memory, whereas conventional metaphors relied more on participants’ generalized background knowledge.

Other than these two studies, few if any other studies have considered these different sub-types of intelligence during the production of figurative language. Although both Silvia and Beaty (2012) and Beaty and Silvia (2013) used metaphor production as a proxy for investigating links and differences between intelligence and creative ability, their research strongly suggests that the ability to produce creative metaphors is significantly predicted by these three sub-types of intelligence (i.e., crystallized intelligence, fluid intelligence, and broad retrieval). It is therefore important to continue exploring these features as they relate to metaphor production, but also to the production of other figurative language. Specifically, producing examples of
sarcasm may tap into similar or different constructs of intelligence, as there exist both novel and clichéd uses of sarcasm.

In terms of processing, there are also very few studies that examine the role of intelligence during figurative language processing. In regards to metaphor, Kazmerski, Blakso, and Dessalegn (2003) measured participants’ working memory and IQ before testing their ability to read and determine if metaphorical and non-metaphorical utterances were literally true or false. Their results found that IQ and working memory were positively correlated ($r = .69$), that IQ was a positive predictor of participants’ processing time, and that metaphors were more difficult to judge as non-literally true when compared to scrambled nonsense control sentences. Moreover, participants with higher IQ demonstrated greater difficulty in judging metaphors as literally false when compared to lower IQ participants. Kazmerski et al. (2003) interpreted these findings as evidence suggesting higher IQ participants were automatically interpreting the metaphorical meaning of metaphors at a very quick rate, and thus the metaphorical meaning (which would be literally false) was immediately competing with a literal interpretation of each metaphor when participants were asked to judge the literal meaning of metaphors.

In terms of satire, Skalicky and Crossley (in press) found that participant’s scores on generalized tests of world knowledge (i.e., a measure of crystallized intelligence) were predictive of humor ratings for satirical headlines, which in turn serves as a proxy for satire comprehension as recognition of a satirical message commonly results in humor (Simpson, 2003). Therefore, there is some initial evidence that these intelligence sub-types may also affect satire comprehension ability when measured via perceptions of humor, but to date these features have not been shown to influence actual processing time of satire.
Although scarce, the existing research exploring connections between different sub-types of intelligence and figurative language production and processing indicate that intelligence is an important individual difference to consider, as individuals with higher intelligence may produce higher quality metaphors, understand metaphorical meaning quicker, and better recognize satirical meaning.

2.5.2 **Working memory capacity (WMC).**

WMC is defined as the limited amount of information that can be held in the mind as one performs some form of mental work, such as solving problems or learning (Cowan, 2005), and is highly correlated with measures of fluid intelligence (Chamorro-Premuzic et al., 2011; Heitz, Unsworth, & Engle, 2005). Therefore, WMC is different from concepts like short-term memory, as it represents an individual’s executive control over cognitive information processing at any given time. WMC can be measured by through various tests that ask participants to perform complex operations in a set amount of time, such as memorizing a series of letters while also solving math problems or recalling the last \( n \) number of items from a sequence that is a random number of \( m \) items long.

The role of WMC has been investigated during both the production and processing of figurative language. For instance, Blasko (1999) reported that participants with higher WMC were able to describe metaphors with more detail when compared to participants with lower WMC, suggesting that WMC positively correlates with metaphor comprehension ability. In a series of three experiments, Chiappe and Chiappe (2007) tested whether WMC was predictive of both metaphor comprehension and production ability. In Experiment 1, participants were tested for WMC by asking them to provide word endings for a series of sentences, and then to recall all the words they provided after a random number of sentences. Participants also completed a
stroop task, where they were shown words spelling the names of colors presented in a font color that either matched or did not match the spelling of the word itself (e.g., BLUE displayed with blue font or green font) and asked to judge each word’s font color. Participants then completed a task where they read metaphors and provided oral explanations of the metaphors, and the researchers recorded the latency between reading the metaphor and the participants’ explanation of the metaphor as evidence of processing time. The participant descriptions of the metaphors were also coded by the researchers for quality of metaphor explanation. Their results found that WMC positively correlated with metaphor processing and quality of metaphor explanation. In Experiment 2, Chiappe and Chiappe (2007) tested whether WMC was predictive of metaphor production. They asked participants to produce metaphors using prompts based on descriptions of nouns (e.g., billboards are something noticeable and unattractive for the frame some billboards ____). The metaphors were then rated for metaphor quality using a six-point scale. Much like Experiment 1, their results found that participants with higher WMC produced higher quality metaphors. In Experiment 3, the authors included more specific WMC tasks to test metaphor production, finding that WMC tests related more strongly to inhibitory control (i.e., the ability to block out irrelevant information while performing mental tasks) were most predictive of metaphor quality. Taken as a whole, the results from Chiappe and Chiappe (2007) suggest that WMC is related to both metaphor production and processing ability, and may specifically be a result of the inhibitory function of WMC.

In a similar study, Pierce & Chiappe (2008) measured the WMC of participants before they completed a metaphor production activity similar to the one used in Chiappe and Chiappe (2007). However, in this study, the authors recruited additional participants to rate the produced metaphors for both aptness (i.e., how well does the metaphor impart the specific meaning) and
conventionality (i.e., how common the answer provided was). The results were similar to those reported by Chiappe and Chiappe (2007), suggesting that participants with higher WMC produced higher quality metaphors.

WMC has also been found to influence the online processing of sarcasm. Kaakinen et al. (2014) investigated sarcasm processing using eye-tracking techniques. While their main finding was that sarcastic utterances had significantly more re-readings than non-sarcastic utterances, they also found that WMC influenced the number of re-readings. Specifically, participants with higher WMC were more likely to re-read ironic utterances during the first-pass of the utterance (i.e., they chose to re-read the sentence while still completing their first full reading of the utterance, rather than returning to the sentence after moving on to later portions of the text). Kaakinen et al. (2014) interpreted these findings as evidence suggesting that higher WMC participants were quicker to identify the irony in the sarcastic utterances, because the first-pass re-reading was indicative of a participants’ purposeful reconsideration of an utterance’s meaning.

The connection between WMC and metaphor and sarcasm was again highlighted in a recent study by Olkoniemi et al. (2016), who also used eye-tracking to analyze the online processing of metaphor and sarcasm. Based on results from the WMC test, participants were grouped into either a high or low WMC group. In terms of metaphor, their study found that participants were less likely to look back at metaphorical utterances after they had first read them during the second half of the experiment compared to the first, but this did not hold true for the participants grouped into the low WMC group, who exhibited no significant changes in the number of look backs during the second half of the experiment, suggesting that lower WMC participants always needed more processing time to interpret metaphors, regardless of how familiar they were with the reading task. Regarding the sarcastic utterances in the experiment,
the results were similar to Kaakinen et al. (2014), in that participants in the high WMC group initiated more re-readings of sarcastic vs. literal sentences near the second half of the experiment, whereas participants with lower WMC had significantly more look backs for sarcastic utterances when compared to literal during the entire experiment.

In general, the results from these studies suggest that WMC is an important individual difference to consider when studying the production and processing of metaphor and sarcasm. Moreover, the results converge to indicate that higher WMC results in better ability to produce metaphor, and also results in better ability to comprehend metaphor and sarcasm. Many of these studies interpret the positive relation between WMC and figurative language comprehension to be evidence in favor of dual-stage theories that posit an utterance’s literal meaning must first be interpreted before arriving at the figurative meaning, as a higher WMC would allow for more information (i.e., two competing meanings) to be more easily considered at the same time.

2.5.3 Need for cognition (NFC).

NFC measures an individual’s desire to perform cognitive procedures that require effort (Cacioppo, Petty, & Kao, 1984), and it is related to certain personality traits (e.g., openness, goal orientation) as well fluid intelligence, but nonetheless has been validated as its own separate construct (Fleischhauer et al., 2010; Furnham & Thorne, 2013). NFC is measured by asking participants to answer how strongly they agree or disagree with statements such as I like tasks that require little thought once I’ve learned them (Cacioppo et al., 1984). Because of the hypothesized need to consider multiple meanings in order to fully understand figurative meaning, two studies have investigated the potential role that NFC may play during figurative language comprehension. Specifically, along with WMC, Kaakinen et al. (2014) also investigated whether NFC would affect sarcasm comprehension when measured via eye tracking,
and found that NFC did not emerge as a significant predictor of sarcasm processing time.

However, Olkoniemi et al. (2016) also included NFC as a predictor when examining the processing of metaphor and sarcasm, and replicated the finding that NFC did not significantly predict sarcasm reading time, but did find that high NFC affected reading times for metaphors, with higher NFC participants spending more time reading metaphors. Olkoniemi et al. (2016) interpreted this finding to suggest that interpreting metaphors satisfied the desire among higher NFC participants to complete tasks that require more difficult reasoning and inference.

There studies suggest that NFC may be an important consideration for metaphor processing, but not for sarcasm processing. Because NFC is correlated with some measures of intelligence (Fleischhauer et al., 2010; Furnham & Thorne, 2013), which have also been shown to predict metaphor production and processing ability, it may be that Olkoniemi et al. (2016) tapped into these same constructs when measuring NFC in their study. Therefore, it is important to consider NFC in a study that also measures intelligence in order to see if the two measures overlap to a significant degree. Moreover, testing whether NFC is predictive during the production and processing of other forms of figurative language will also help establish whether it is an important feature to considering during figurative language use.

2.5.4  **Abstract thinking (AT).**

Measures of AT identify how different people perceive and identify different events. If someone possesses higher levels of AT, then tend to identify larger, more abstract meanings of an action, which include motivations and future goals of the action. Conversely, someone with lower levels of AT focuses on the concrete and immediate details of an action (Vallacher & Wegner, 1989). As a construct, AT is commonly measured using the Behavior Identification Form (BIF), which is a list of 25 different actions with two different descriptions for each action.
For example, the action of reading is described either as following lines of print (concrete) or gaining knowledge (abstract), and participants are asked to choose which description best matches their interpretation of the action (Vallacher & Wegner, 1989).

To date, only one study has considered the role that AT may have during figurative language production and processing. Huang, Gino, and Galinksy (2015) investigated connections between sarcasm use and creative ability through a series of four experiments. Huang et al. (2015) hypothesized that both sarcasm use and creative ability relied on a similar cognitive construct (i.e., abstract thought). In their first two experiments, Huang et al. (2015) asked participants recruited online to produce, comprehend, or recall past instances of sarcasm (or were assigned to a non-sarcastic control condition) before completing tasks designed to measure creative ability. Regardless of how sarcasm was being used, exposure to sarcasm significantly boosted participants’ performance on subsequent tests of creativity when compared to the control group. In Experiment 3, Huang et al. (2015) replicated the first two experiments using human subjects in a laboratory setting, which allowed them to also include more complex tests of creativity and also measure participants’ AT using questions from the BIF. Crucially, participants completed the BIF questions after being exposed to sarcastic or non-sarcastic stimuli and before completing tests of creative ability. Results reported that AT was significantly higher for participants who produced and listened to sarcastic utterances when compared to the control conditions, and also that participants exposed to sarcasm outperformed participants in the control conditions on tests of creativity. Huang et al. (2015) took these results as evidence suggesting that sarcasm boosted AT, which in turn increased creative ability.

In order to explain the connections between AT and sarcasm use, Huang et al. (2015) cited theoretical descriptions of sarcasm which define sarcasm as an utterance which “signifies
the opposite” (p. 163) and thus results in contradictions between meanings. Huang et al. (2015) postulated that the ability to develop and consider multiple meanings is inherently a creative ability, and thus sarcasm is a creative thought process that benefits from (and induces) higher levels of abstract thinking. While this definition of sarcasm relies on simplistic views of verbal irony and sarcasm that may not accurately describe the nature of figurative meaning (see section above), the overall interpretation that sarcasm use is a creative process may nevertheless provide a beneficial theoretical lens through which to examine sarcasm and other types of figurative language.

2.5.5 **Summary of cognitive measures.**

In sum, a variety of different cognitive individual differences have been shown to interact with both figurative language production and processing. In general, the results collectively suggest that increased cognitive ability, whether measured as intelligence, working memory capacity, need for cognition, or abstract thought are beneficial in both understanding and producing metaphor, sarcasm, and satire. However, many of these studies have only investigated one or two cognitive features for one type of figurative language, and almost no studies have investigated satire. Therefore, it is important to investigate these features in the context of different types of figurative language to see if measures related to one type also affect another (i.e., does intelligence affect sarcasm?), and it is also important to determine if any of these constructs overlap significantly with one another and are thus capturing similar cognitive mechanisms. Finally, the impetus for inclusion of many of these features has been based on the notion that figurative language use requires the consideration of multiple meanings at once, but it is also important to consider other possible alternative explanations for the role that individual differences may play in figurative language use.
2.6 Demographic Measures

2.6.1 Gender and age.

In addition to cognitive individual differences, people also differ based on different demographic features, such as gender and age. Gender has been studied most commonly in the context of sarcasm use, perhaps because of intuitive assumptions about differences between men and women’s language use in general (e.g., assumptions that men are more direct than women). In terms of gender, studies have reported that women tend to use sarcasm less than men and are more likely to interpret an indirect meaning from a speaker’s utterance when compared to men (Gibbs & Colston, 2012). For example, Gibbs (2000) analyzed a corpus of transcribed recordings of spontaneous speech among college students and their friends for instances of different forms of verbal irony, including sarcasm. Among his findings, Gibbs (2000) reported a relatively equal balance in some verbal irony use among males and females, such as hyperbole (41% male, 59% women) and rhetorical questions (46% men, 54% women). However, in terms of sarcasm use, men produced nearly double the amount of sarcasm when compared to women (64% men, 36% women). Therefore, in at least studies investigating sarcasm, it is important to consider the gender of participants as well as the gender composition of groups, as men may use sarcasm more than women.

In terms of age, while studies suggest that even young children have developed an initial ability to understand metaphorical expressions, the ability to fully understand verbal irony does not develop until adolescence, which may be due to young children’s inability to fully interpret pragmatic functions of sarcasm (Gibbs & Colston, 2012). For instance, Pexman, Glenwright, Krol, and James (2005) studied young children’s (7-10 years old) reactions to puppet shows which ended in either a compliment or criticism that was delivered either sarcastically or
directly. Additionally, the authors also described the puppets as either not liking each other, not knowing each other, or liking each other a lot. After watching the puppet show, the authors asked the children to rate how nice or mean the speaker was, whether the speaker was funny, whether the speaker was teasing, and whether the speaker actually believed what he or she said. Their results found that while the children incorporated the relationship information into their ratings during the direct criticism or praise, they did not do so for their ratings in the sarcastic conditions. For instance, the children would recognize that if two close friends were criticizing each other, the criticism was not as harsh as if two strangers were to do so, and that humor or teasing may be the main goal of the criticism. However, when viewing sarcastic criticism or praise, the children did not incorporate the relationship information or indicate that teasing or humor may have been a function of the sarcasm. Pexman et al. (2005) concluded that children did not possess knowledge of stereotypical uses of sarcasm in different relationship settings that adults possess, and therefore while children may be able to understand a sarcastic utterance means something different than its literal form, they lacked the social knowledge to properly interpret the pragmatic functions of the sarcastic utterances.

In general, studies investigating children’s ability to understand verbal irony continue to report similar results, in that children demonstrate signs of developing irony comprehension ability which matures as they reach adolescence (Climie & Pexman, 2008; E. S. Nilsen, Glenwright, & Huyder, 2011; Recchia, Howe, Ross, & Alexander, 2010), although some research indicates that children’s ability to understand sarcasm can be enhanced when performed in contexts that are familiar to the children (Angeleri & Arenti, 2014). Aside from testing the developmental aspects of figurative language comprehension and use, few studies have considered the role that age may play in adult figurative language use.
However, recent research has suggested that age may be an important factor in satire interpretation. Based on a study demonstrating that younger participants reported better ability to concentrate when viewing satire as compared to older participants (Boukes et al., 2015), Skalicky and Crossley (in press) tested the effect of age (among other variables) on participants’ online processing of satirical and non-satirical newspaper headlines. Their results found a significant interaction between reading time and age, in that older participants read satirical headlines much slower than non-satirical headlines when compared to younger participants. These results suggest that reading times between satirical and non-satirical headlines were not significantly different for younger participants, but that a significant difference in reading times between satirical and non-satirical news headlines existed as participant age increased. Skalicky and Crossley (in press) interpreted this finding as a possible familiarity effect, in that younger participants may interact with satire more on a daily basis when compared to older participants, as satirical headlines are now commonly broadcast on social media, and satirical television shows are becoming more common sources of new and political critique among younger people in the United States.

The current research suggests that age does play a role during figurative language comprehension, but mainly from a developmental perspective, and that the ability to understand some forms of figurative language (i.e., verbal irony) does not fully develop until adolescence. Because the participants in this dissertation are all at least 17 years of age, there should be no developmental differences in terms of how they interpret and produce metaphor, sarcasm, and satire. Regardless, among adults age may serve as a proxy for familiarity or exposure to other types of figurative language (i.e., satire), and thus is an important variable to capture when investigating figurative language use.
2.6.2 Language knowledge.

One demographic feature that is often overlooked during investigations of figurative language is the language knowledge and language background of the participants in studies of figurative language use. For instance, most psycholinguistic studies testing sarcasm processing display experimental stimuli in participants’ first language (L1) without also considering whether those participants know a second language (L2) or multiple other languages. However, language knowledge is an important variable to measure during any investigation into language use, as psycholinguistic research into bilingual and multilingual language users has demonstrated that one’s language knowledge is stored concurrently in a shared and integrated lexicon, and that similar linguistic features among languages (e.g., orthography, phonetics, semantics) co-activate words and constructions from all known languages during language input (Basnight-Brown & Altarriba, 2007; Dong, Gui, & Macwhinney, 2005; Kroll & Scholl, 1992; Kroll, Van Hell, Tokowicz, & Green, 2010; Schoonbaert, Duyck, Brysbaert, & Hartsuiker, 2009; Tokowicz, 2014; van Hell & Dijkstra, 2002). Therefore, differences in processing times during figurative language experiments could be related to differences in lexicon size among participants. For instance, two participants may share the same first language, but one participant may know an additional two languages, whereas the other participant is a monolingual speaker. This may result in the monolingual speaker processing the stimuli faster than the multilingual speaker, as the multilingual speaker may have more words and constructions co-activated during the processing of the figurative stimuli, resulting in more activated input that must be considered before arriving at the figurative meaning. On the other hand, the monolingual speaker can more quickly activate the figurative meaning, as they have less activated input from other known languages to consider.
Recently, there have been some studies investigating the role of language knowledge and background during metaphor, sarcasm, and satire use. Regarding metaphor, Piquer-Piriz (2010) tested Spanish L1-English L2 children’s ability to attribute metaphorical senses of *warm* and *cold* to the actions of different speakers (e.g., someone who is nice is *warm*) using English. The children were split into three age groups of six, eight, and ten years old, and the results revealed a developmental pattern similar to that found in L1 metaphor comprehension, in that older children were more accurate in using the metaphorical versions of *warm* and *cold* and were also better able to explain the metaphorical connections between the concepts of temperature and personality. While Piquer-Piriz (2010) demonstrated a similar developmental trajectory in metaphor interpretation among L2 users of English, her analysis did not account for possible language related explanations for differences in her participants’ responses (i.e., did knowledge of Spanish influence participants’ ability to use the metaphorical meanings?). Her study is nonetheless useful because it suggests that, at least among Spanish-English users, figurative language use develops in a similar manner alongside second language proficiency and cognitive development. In a different study examining language proficiency and figurative language ability, Johnson and Rosano (1993) compared a group of English L1 speakers against students enrolled in the first or second semester of an English as a Second Language (ESL) class on metaphor interpretation and production tasks. While Johnson and Rosano (1993) found no differences between the different language groups in terms of understanding the semantic meaning of metaphors, they did find a positive correlation between English proficiency and metaphor production fluency (i.e., ability to produce different metaphorical interpretations for sample metaphors such as *his shirt was a butterfly*), suggesting that metaphor production ability may lag behind comprehension when developing proficiency in a second language.
Other studies have specifically focused on whether a language user’s L1 influences figurative language comprehension in an L2. For example, Golden (2010) performed a study examining metaphorical uses of words in textbooks used in high schools in Norway. In this study, Norwegian L1 and Norwegian L2 speakers were asked to complete multiple choice tests to choose the correct interpretation of both metaphorical and non-metaphorical uses of words such as light and medal. The results showed little difference in accuracy between the L1 and L2 groups when choosing interpretations for the non-metaphorical uses, but there was a significant difference when analyzing answers for the metaphorical data, with the L1 user accuracy at 88.4% and the L2 user accuracy at 67%. In one of the follow up analyses using this data, Golden (2010) further broke down the participants into more specific language groups (i.e., grouped all of the L2 speakers based on their respective L1s) and assessed whether the metaphorical uses of words had similar correlates in the participants’ respective L1s. Results demonstrated that for at least two L1 groups (Urdu and Vietnamese L1 users), accuracy was significantly higher for metaphors that had equivalent forms in the L1 compared to those that did not, providing some evidence suggesting that metaphor comprehension is facilitated in an L2 if the L1 has a corresponding equivalent use.

Conversely, a different study demonstrated that knowledge of metaphors in an L1 can also interfere with L2 metaphor comprehension. Philip (2010) compared metaphors in the academic writing produced by Italian L1-English L2 learners against large Italian and English corpora, and found that attempts to use L1 translation equivalents to create metaphors in English were sometimes hindered because specific, non-metaphorical translations of words were being used in an attempt to produce metaphors. For example, one student wrote heavy depression in an attempt to translate deep depression from Italian. Although the writer knew how to construct this
metaphor in Italian, the translation into English missed its mark because the writer chose another possible translation for pesante (i.e., heavy) that did not carry the same figurative meaning as deep when paired with the word depression in English. Had the metaphor not existed in Italian, the writer may have focused more on the correct English construction.

Interference can also be caused by the relative lack of use of certain types of figurative language in a participant’s L1. For example, while satire is fairly widespread in Europe and the United States (Simpson, 2003), it is used less frequently in Japanese culture (Prichard & Rucynski, 2017). Because of the prevalence of satirical news stories to be shared on social media, Prichard and Rucynski (2017) tested Japanese L1-English L2 speakers’ ability to differentiate between real, satirical, and odd (i.e., real yet strange) news stories in English. Despite the fact that the stories were controlled for vocabulary level and the participants were allowed to consult dictionaries, the Japanese L1 speakers were significantly less accurate when compared to English L1 speakers at categorizing the news stories, which the authors interpreted to be partially a result of the Japanese L1 users having less overall experience with the genre of satire when compared to the English L1 speakers (who were American).

While these studies provide some evidence that language knowledge and proficiency may interfere with ultimate comprehension and production ability, they are unable to provide information about the online processing of figurative language in an L2. Studies that have used psycholinguistic methods to investigate L2 online processing of figurative language have produced results suggesting that bilinguals process figurative language differently in their L2 when compared to their L1 and also when compared to monolingual speakers of their L2, but are still relatively few and have produced sometimes inconsistent findings (Heredia & Cieślicka, 2015). For instance, Vaid, Belem, López, and Martinez (2015) performed two experiments
testing metaphor comprehension which included monolingual English speakers, Spanish L1-English L2 speakers reporting English dominance, and Spanish L1-English L2 speakers reporting dominance between Spanish and English. In Experiment 1, Vaid et al. (2015) asked participants to determine whether an English sentence was literally true or false, and then presented participants with metaphorical phrases, non-metaphorical phrases, and scrambled metaphors (i.e., nonsense metaphors formed from two legitimate metaphors such as *some cats are chains*). Results indicated that all participants were slower to accurately judge metaphorical statements when compared to non-metaphorical true (e.g., *some flowers are daisies*) and non-metaphorical false statements (e.g., *some flowers are cobras*). Additionally, the bilingual speakers showed a significantly higher difference between the metaphorical and non-metaphorical conditions when compared to the monolingual speakers. Thus, the results from Experiment 1 suggested that metaphorical meanings were automatically activated for all participants, but that the activation of the metaphorical meaning caused more interference during the decision process for bilinguals when compared to monolinguals.

In Experiment 2, Vaid et al. (2015) recruited only Spanish L1-English L2 bilinguals, separated them into English dominant versus equally balanced bilinguals, and then repeated the procedure from Experiment 1, but also included translated versions of the metaphors in Spanish. Results from Experiment 2 demonstrated no interference for the English dominant group for either English or Spanish metaphors, and thus failed to replicate the results from Experiment 1. Additionally, the equally balanced group demonstrated an interference effect in Spanish only. Vaid et al. (2015) hypothesized that the mixed language stimuli may be partially responsible for these different findings, as was the direct translation of English metaphors into Spanish, which may have resulted in metaphorical meaning becoming lost in translation. Despite these
inconsistent findings across their experiments, the results of Vaid et al. (2015) do suggest that monolinguals and bilinguals differ in speed of figurative language processing, and also that differences among bilinguals in terms of language dominance also affect this process.

Another aspect of bilingualism that may affect figurative language comprehension ability is the relative amount of experience one has with translation. López, Vaid, Tosun and Rao (2017) tested whether experience with informal translation (i.e., language brokering) would facilitate figurative language comprehension in a second language. López et al. (2017) recruited 80 Spanish L1-English L2 bilinguals with high English proficiency and sorted them into two groups based on self-reported experience with informal translation. During the experiment, participants were shown an adjective (e.g., stinging) combined with a word that would result in either a metaphorical (e.g., stinging insult), non-metaphorical (e.g., stinging insect), or nonsense meaning (e.g., stinging picnic). Participants were to indicate whether the phrase made plausible sense or not. Stimuli was presented in two separate sessions, one in English, and one in Spanish. Their results demonstrated that regardless of language of presentation or brokering experience, all participants responded to non-metaphorical pairings quicker than metaphorical pairings. In addition, language non-brokers were both faster and more accurate when making their decisions in English compared to Spanish, whereas language brokers showed no significant differences between the two languages, suggesting that the language brokering experience served as an equalizing variable on the brokers’ ability to process the stimuli. Thus, the results from this study provides further evidence that differences in language experience can influence speed of figurative language processing.

Taken together, these studies suggest language knowledge is an important variable to consider during any investigation into figurative language use. Equivalent metaphorical
meanings between a language user’s first and second languages may serve to facilitate or inhibit production ability, whereas overall proficiency in an L2 may influence processing speed. In addition to these factors, one’s aptitude for language learning is another important element of language background. Language aptitude aims to measure the innate ability to learn new languages by measuring one’s explicit and implicit ability to process and comprehend new and unfamiliar features of language (e.g., vocabulary, grammar), and therefore may also be predictive of figurative language production and comprehension ability (Dörnyei & Skehan, 2003; Li, 2015).

### 2.7 Summary of Individual Differences

This section has reviewed a number of studies demonstrating that a wide range of individual differences have been shown to influence figurative language use. In regards to cognitive differences, influences among features such as intelligence and working memory capacity have all been shown to interact in some manner with the production and processing of both metaphor and sarcasm, yet these features are typically investigated in isolation or only for one aspect of figurative language use (e.g., metaphor production only). In addition to cognitive differences, demographic features such as gender, age, and language knowledge and background may also play significant roles during figurative language use. No previous studies have investigated all of these features in a comprehensive manner during both the production and processing of metaphor and sarcasm, and very few studies have investigated these features and how they relate to any aspect of satire use.

### 2.8 Figurative Language and Linguistic Features

While individual differences reflect the characteristics of the language user that may influence figurative language use, linguistic features are representative of varying differences
among figurative language itself. Traditionally, the main linguistic feature that has defined figurative language is the dichotomy between literal and figurative meaning (see section above on figurative language processing). However, other linguistic features besides the literal and figurative distinctions have been investigated in metaphor, sarcasm, and satire, and these features may also account for differences in production and comprehension of figurative language.

Computational explorations of figurative language such have celebrated some success in cataloguing the linguistic nature of figurative language. For example, computers have been trained to produce metaphors and similes using linguistic formulas that measure semantic distance between two topics based on semantic dictionaries (Veale, 2012b, 2012a, 2013). Other researchers employ data-mining and machine learning paradigms to determine which linguistic features are capable of distinguishing figurative from non-figurative language. For example, many studies have used Twitter as a source of data in order to locate sets of linguistic features capable of distinguishing sarcastic from non-sarcastic tweets (González-Ibáñez, Muresan, & Wacholder, 2011; Joshi, Sharma, & Bhattacharyya, 2015; Khodak, Saunshi, & Vodrahalli, 2017). Similar studies have compared satirical and non-satirical corpora against one another, including satirical product reviews and newspaper stories (Burfoot & Baldwin, 2009; Horne & Adali, 2017; Skalicky & Crossley, 2015).

While useful in their own right, there are several limitations to computational approaches when trying to determine which of these features may influence production and processing of figurative language among humans. First off, the purpose of many machine learning studies is to develop the best linguistic algorithm that a computer can use to identify a target text type (e.g., sarcasm). Thus, these features are useful for computers, but may not reflect features used by actual language users during the production and processing of figurative language. Secondly, the
sheer number of studies and different sets of linguistic features that have been reported as predictive of sarcasm or satire are exhaustive and beyond the scope of this review, but include linguistic features ranging from numbers of words, punctuation use, positive or negative sentiment, and contextual features such as how many other documents mention similar entities. This suggests that different data sets of sarcasm or satire rely on different linguistic features to differentiate themselves from their non-figurative counterparts, and therefore no single, universal set of linguistic features can be developed to explain the larger constructs of sarcasm or satire. Finally, many studies using Twitter as a corpus rely on user self-tagging of tweets in order to identify instances of sarcasm or satire (i.e., users will post a tweet with the hashtag #sarcasm), calling into question whether the data being used is reflective of the actual construct under study, as it relies on the user’s own intuitive definition of sarcasm. For these reasons, linguistic feature sets identified in computational studies as representative of figurative language are not immediately helpful for locating potential sets of features which may influence actual use of figurative language.

However, computational text analysis methods can still prove useful in identifying important linguistic features when combined with human ratings and carefully controlled stimuli. For instance, Campbell and Katz (2012) tested whether the linguistic features of contextual information surrounding satirical utterances would be representative of theoretical assumptions regarding necessary conditions for satire (e.g., presence of a victim, negative emotion). To test their question, they presented participants with utterances such as I did great on that test framed in minimal contexts. Half of the participants were then asked to fill in the surrounding context so that the utterance would be interpreted sarcastically, whereas the other half were asked to supply the missing contextual details without any specific instructions to make the interpretation
sarcasm. Afterwards, Campbell and Katz (2012) had the contexts rated and categorized into high and low sarcasm inducing contexts and then analyzed the contexts for textual features using the Linguistic Inquiry and Word Count (LIWC) program, which provides numerical measurements of texts based on psychological, emotional, and syntactic variables (Tausczik & Pennebaker, 2010). A comparison between sarcastic and non-sarcastic contexts using the LIWC indices revealed that the sarcastic contexts differed significantly from the non-sarcastic contexts in that they contained more words related to exclusion, sadness, negations and negative emotions, while also containing fewer words related to certainty and past tense when compared to the non-sarcastic contexts. Thus, Campbell and Katz (2012) were able to identify specific linguistic features which distinguished sarcastic from non-sarcastic texts, suggesting that stylistic textual features may also be important in inducing a sense of sarcasm, and that these features should be considered in addition to the extant theories of sarcasm processing.

A similar approach was taken by Skalicky and Crossley (2015), who analyzed a corpus of satirical and non-satirical Amazon.com reviews. Although they did not employ human raters such as Campbell and Katz (2012) did, Skalicky and Crossley (2015) carefully selected a corpus of satirical texts by tapping into the internet phenomenon of satirical product reviews. Specifically, anonymous authors post satirical product reviews for products on Amazon.com and other online shopping websites in order to criticize the product or other entities (e.g., satirizing Mitt Romney’s *binders full of women* gaffe by leaving satirical products reviews for Avery binders). Skalicky and Crossley (2015) constructed a corpus of 375 satirical reviews and 375 non-satirical reviews, taking care to match product type across the corpora. They performed a similar analysis as Campbell and Katz (2012) by using LIWC and also included measures of lexical sophistication using The Tool for the Automatic Assessment of Lexical Sophistication.
The results demonstrated that satire shared some of the same linguistic features identified by Campbell and Katz (2012) as predictive of sarcasm, such as more negative emotion words and tendency to remain in the past tense, but differed in that satirical texts contained more certainty terms and fewer exclusion terms (the opposite of sarcasm). In addition, Skalicky and Crossley (2015) also reported that satirical texts were marked by significantly higher word concreteness when compared to non-satirical texts, suggesting that satirical texts were less lexically sophisticated because they employed less abstract language when compared to non-satirical texts.

These two studies suggest that automatic text analysis methods can shed light on the linguistic properties of figurative language, but further research is needed in order to explore a wider range of lexical properties. Specifically, two types of linguistic features are worthy of investigation because they have been demonstrates as important features of text comprehension in general. These features are lexical sophistication and semantic cohesion.

2.8.1 Lexical sophistication.

Features representative of lexical sophistication (e.g., word frequency) have long been shown to influence the time it takes to recognize individual words in studies of non-figurative language (Ellis, 2002; Forster & Chambers, 1973), and therefore most likely also affect processing times for figurative language, especially in light of Skalicky and Crossley (2015) which provided some evidence that satirical texts are less sophisticated than non-satirical texts. Lexical sophistication is a broad term that aims to define and classify the vocabulary level or a particular text, corpus, or otherwise (Kyle & Crossley, 2015). Many different lexical features tap into the construct of lexical sophistication, such as word frequency (i.e., how often a word appears in a particular reference corpus) or the psycholinguistic properties of words (i.e.,
measures illustrative of cognitive representations of words, such as word abstractness or familiarity). Lexical sophistication can be measured using automatic text analysis tools, such as TAALES 2.0 (Kyle, Crossley, & Berger, 2017), which includes hundreds of linguistic indices that measure the lexical sophistication of single words as well as two- and three-word phrases. Because there is only one examination of figurative language using indices related to lexical sophistication (Skalicky & Crossley, 2015), more research is needed to better understand whether features related to lexical sophistication play a strong role during the production and processing of metaphor, sarcasm, and satire.

2.8.2 Semantic cohesion.

In addition to lexical sophistication, the semantic cohesion of figurative language is important to investigate because metaphor, sarcasm, and satire are all thought to represent some form of incongruity between semantic meanings or contexts. Semantic cohesion is defined as “the presence or absence of explicit cues in the text that allow the reader to make connections between ideas in the text” (Crossley, Kyle, & McNamara, 2016, p. 1228), and has been used in previous theoretical explanations of metaphor processing. Specifically, the internal cohesion between the different conceptual domains in metaphors has been measured using techniques such as Latent Semantic Analysis (LSA), which provides strengths of associations among words in a text by analyzing all the possible semantic contexts each word appears in (Kintsch, 2008). When applied to metaphors, LSA analyses reveal that metaphors that are easier to understand are those with more coherent interpretations. For example, Kintsch and Bowles (2002) noted that participant interpretations of metaphors such as Mosquitos are vampires were more consistently coherent with one another (e.g., blood sucker, sucks blood) as well as more strongly associated with both terms in the metaphor itself. More difficult metaphors, such as Happiness is a ditch
resulted in a greater range of answers among participants that were therefore less cohesive with the words in the metaphor. Kintsch (2008) argued that LSA is thus representative of how the mind processes metaphor. Metaphors with similar concepts are more quickly processed because the mind does not need to search very long for a coherent meaning, whereas less cohesive metaphors require more processing time because the mind must search longer for an interpretation that is semantically coherent with the comparison being made in the metaphor.

Cohesion may also play a role in the production of metaphors. Some research into creativity has reported that participants who perform better on tests of divergent thinking, where a participant is asked to generate as many possible solutions to a prompt as possible, generate ideas that are more conceptually distant from one another (Acar & Runco, 2014). As such, metaphor production quality may be in part due to perceptions of creativity, which in turn may be reflective of the relative semantic distance among concepts used in a metaphor. However, this claim needs to be further empirically validated.

Using a tool similar to TAALES, semantic cohesion can be measured using the Tool for the Automatic Analysis of Text Cohesion (TAACO; Crossley et al., 2016). TAACO includes over one hundred indices which measure cohesion at different levels, including semantic overlap at the word-to-word, sentence-to-sentence, and paragraph-to-paragraph levels. Aside from studies of metaphor, cohesion has not been considered as a factor during the production or processing of sarcasm and satire, and therefore more research is needed in order to better understand the role semantic cohesion may play during the production and processing of metaphor, sarcasm, and satire.
2.8.3 **Summary of linguistic features.**

Studying the linguistic features of metaphor, sarcasm, and satire provides a ripe opportunity to locate additional influences on how these types of figurative language are produced and comprehended. Although the majority of research into linguistic features of figurative language is computational in nature, there is some evidence that obtaining numerical measures of the linguistic properties of figurative language can help further describe the nature of figurative language, which may in turn provide insight into how figurative language is produced and processed.

2.9 **The Pragmatics of Figurative Language**

Finally, one important feature overlooked during investigations of figurative language has been the pragmatic and social-contextual variables associated with its use. In terms of production, figurative language is used for specific pragmatic reasons that go beyond the literal semantic message of the utterance. For instance, similar to how speakers employ indirect speech acts to avoid direct impositions, such as asking someone to close a window by stating *Don’t you think it’s cold in here?*, speakers also use sarcasm in order to deliver praise or criticism in pragmatically strategic ways (Jorgensen, 1996). For example, Dews, Kaplan, and Winner (1995) hypothesized that sarcasm must serve a unique pragmatic function, or else speakers would never feel the need to be sarcastic. Dews et al. (1995) asked participants to listen to or read examples of sarcastic and non-sarcastic criticism or praise. Participants then rated these materials for how humorous and insulting they felt the examples were. The results demonstrated sarcastic praise was rated as more insulting than literal praise, while sarcastic criticism was rated as less insulting than literal criticism. Dews et al. (1995) interpreted these findings as evidence that sarcasm was used for specific pragmatic purposes (i.e., mitigating face threats) different than the non-sarcastic
equivalent of an utterance. In a later study, Dews and Winner (1995) reported similar results and argued that sarcasm functioned to pragmatically “mute” (p. 3) the evaluative meaning typically associated with non-sarcastic criticism or praise (i.e., negative sarcasm was less negative than direct criticism, whereas positive sarcasm was less positive than direct praise).

Much like sarcasm, metaphor and satire also have their respective social functions. Metaphor can be used to explain concepts more directly in a vivid and concise manner, and can also be used to strengthen social relationships, broadcast attitudinal feelings, and serve to politically persuade an undecided voter (Gibbs, 1994, 2011; Lakoff, 2009; Lakoff & Johnson, 2008). Satire’s primary social function is to subtly criticize targets in a manner that evokes humor from a sympathetic audience (Simpson, 2003). Therefore, metaphor, sarcasm, and satire are not just indirect ways of imparting non-figurative meanings, they also accomplish specific pragmatic goals, and comparisons between figurative and non-figurative utterances that do not consider the pragmatic functions of figurative language are therefore excluding important additional information that is included during figurative language use (Colston, 2015; Gibbs & Colston, 2012).

Because figurative language involves imparting specific pragmatic meanings, the interpretation of figurative language also relies on a hearer’s ability to incorporate pragmatic information into the interpretation of figurative meaning. There are a large number of possible sources of pragmatic knowledge that affect figurative language comprehension, such as register and genre familiarity, cultural background, social membership and identity, the ability to recognize and process emotions, and affective perceptions. All of these sources of knowledge can influence how quickly a figurative meaning is interpreted, as well as how likely one is to use figurative language in a specific situation (Gibbs & Colston, 2012). For instance, Deignan et al.
(2013) demonstrated how the metaphorical conceptualization of the word *copy* in reference to cloning takes on a positive connotation within scientific communities, but negative connotations within mass media publications. Additionally, Burgers et al. (2012) analyzed a corpus of sarcasm in six different written genres (commercial and noncommercial advertisements, newspaper columns, cartoons, book and film reviews, and letters to the editor). Their results indicated that while sarcasm existed in each of these written genres, features of the sarcasm, such as the target of the sarcasm, levels of evaluations, and incongruity differed significantly among the different written genres.

Sarcasm interpretation is also influenced by the amount of contextual knowledge available to a hearer. Once hearers are provided with more contextual information about a sarcastic text or utterance, such as the relationship between the speakers, the speakers’ occupations, or the intention of the speaker, participants are better able to directly access the sarcastic meaning (Gibbs, 1986; Ivanko & Pexman, 2003; Pexman et al., 2000). For instance, based on results from a previous study that found occupations such as truck drivers and other blue-collar workers were more likely to evoke expectations of sarcasm, whereas higher status professions such as nursing evoked expectations of metaphor use (Katz & Pexman, 1997), Pexman et al. (2000) tested whether stereotypes regarding different occupations would affect expectations for sarcastic or metaphorical speech. To do so, Pexman et al. (2000) asked participants to read short conversations between two speakers that included one utterance following a metaphorical pattern of *A is a B* but which could also be interpreted sarcastically (e.g., *that comment hit the bull’s-eye*). Additionally, some of the passages were embedded with information about the speaker’s occupation, which varied between occupations associated more strongly with sarcasm (e.g., truck driver) or with metaphor (e.g., nurse). They then asked
participants to read the passages and measured participants’ reading times on the target $A$ is a $B$ phrase in the passage. Their results demonstrated that passages including occupations more inclined towards metaphor use were read significantly faster than those including occupations more strongly associated with sarcasm use. Pexman et al. (2000) interpreted these findings to suggest that readers incorporate all possible pragmatic and contextual information into their interpretation of a speaker’s utterance, and that when the passages contained occupational information more inclined towards sarcasm use, readers spent more time interpreting the target utterance and reinterpreting it in a sarcastic, rather than metaphorical manner.

Neurolinguistic research also provides evidence suggesting hearers’ incorporate social and pragmatic information into their interpretation of sarcasm use. Akimoto et al. (2014) analyzed regions of participants’ brains using functional magnetic resonance imaging (fMRI) while they were presented with fictional situations ending in either an sarcastic, non-sarcastic, or nonsensical ending using pictures of speakers with speech bubbles. Participants read the example while the researchers took measurements of which regions of the brain were activated during processing of the different conditions. The results indicated that when participants were reading the sarcastic situations, portions of the brain related to emotional and social knowledge were more active than during the processing of non-sarcastic language, suggesting greater incorporation of these types of knowledge during the processing of sarcasm.

These studies provide evidence suggesting that comprehension of metaphorical and sarcastic utterances involves incorporation of different sources of pragmatic knowledge. Satire also relies on these sources of knowledge, perhaps to an even greater extent than metaphor and sarcasm. Indeed, many of the studies investigating satire comprehension focus on how participant attitudes, opinions, and perceptions shape satirical comprehension, although they
typically do not employ online processing paradigms. For instance, Pfaff and Gibbs (1997) conducted several experiments testing satirical interpretation by asking participants to read passages from a satirical book mocking extreme political correctness. In Experiment 1, they asked participants to read a satirical passage and then write down a thesis representative of the author’s intentions, as well as describe whether they questioned the author’s motives and any other information about their perceptions of the author. Results demonstrated that when participants first read the passages, only 23% of the readers reported a satirical interpretation. After being informed of the author’s satirical intent, this number increased to 47%. In Experiment 2, Pfaff and Gibbs (1997) repeated the procedure from Experiment 1 and also asked participants to point out which elements of the passages suggested a satirical intent among the author. Analysis of the results revealed that the most common textual signal for satire was repetition and exaggeration. Finally, in Experiment 3, Pfaff and Gibbs (1997) again repeated the procedure from Experiment 1, but also constructed fictional descriptions of the author: a neutral description, a politically conservative description, and a politically liberal description. Participants who were told the author was conservative were more likely to describe the author as satirizing excessive political correctness when compared to the neutral and liberal descriptions. This, much like the stereotypes of occupations investigated in sarcasm comprehension (Katz & Pexman, 1997; Pexman et al., 2000), stereotypical views of conservative and liberal political views in the United States influenced the manner in which participants interpreted the meaning of the passages. As a whole, the results reported by Pfaff and Gibbs (1997) demonstrate that readers were considering representations of the author when reading the passages and interpreting the satirical stories.
Other studies of satire comprehension have reported similar findings. When viewing the satirical television show *The Colbert Report*, which satirizes conservative political pundits in the United States, participants who reported politically conservative views more often interpreted the show as delivering serious conservative viewpoints in a humorous manner, whereas participants who reported politically liberal views interpreted the show as a direct satirical attack on conservative beliefs (LaMarre et al., 2009). In a similar study, Boukes et al. (2015) had participants view original satirical or non-satirical video critiques of a recent policy decision in the Netherlands. Participants were asked before watching the video to provide their opinion regarding the policy decision. They then watched the video and were then asked to rate the video for humor, describe how focused they were during the video, and provide their reaction to the video. Similar to the findings of LaMarre et al. (2009), Boukes et al. (2015) found that participants who agreed with the satirical message rated the satirical videos as funnier when compared to those who did not agree with the satirical message (based on the opinions obtained prior to watching the videos).

Finally, affective perceptions, such as humor, are also important when considering how hearers understand satire. LaMarre et al. (2014) compared comprehension of satire with an overt and obvious humorous delivery against satire with a less overt humorous delivery. Participants who watched the overt satire were more likely to interpret only the humorous intent of the satire, whereas participants watching the less overt satire were more likely to interpret both the underlying satirical message as well as the humor typically associated with appreciation of a satirical message. This suggests that satire that is highly marked by humor may result in only partial comprehension of the humorous message. In addition, Skalicky and Crossley (in press) found that satirical headlines were rated as significantly less familiar, less understandable, less
sincere, and more humorous when compared to non-satirical newspaper headlines, suggesting that readers are rely on a variety of different perceptions to recognize satire. Thus, it appears that humor and other affective perceptions are associated with satire comprehension. However, at least one study has demonstrated that even if a reader is unable to reach a satirical interpretation when exposed to satire, that reader may still nonetheless comprehend the overall meaning intended by the satirical author, albeit in a less convincing and non-humorous manner (A. Johnson et al., 2010).

2.9.1 Summary of pragmatic influences.

There is ample evidence demonstrating that social and contextual knowledge is crucial when using metaphor, sarcasm, and metaphor. Despite this evidence, the influence of these variables is still understudied, especially in psycholinguistic investigations of figurative language production and processing (Colston & Katz, 2005). While it is difficult (if not impossible) to consider every possible social and contextual influence that is present when examining figurative language use, these studies provide a compelling reason to at least avoid examining decontextualized examples of figurative language, as well as further highlight the false equivalence that exists between figurative and literal utterances that are commonly investigated (Gibbs & Colston, 2012).

2.10 Literature Review Summary

There has been a considerable amount of research investigating figurative language. This review has demonstrated that many studies have focused on testing theoretical models of figurative language comprehension, which aim to determine whether figurative language comprehension requires a two-stage process of first identifying then rejecting the literal version of a figurative utterance’s meaning before arriving at the figurative meaning. However, there is
strong evidence suggesting that hearers incorporate a large amount of contextual information available to them when interpreting figurative utterances in naturalistic discourse, and studies that remove this contextual information from stimuli may suffer from a lack of ecological validity when investigating figurative language.

Furthermore, because figurative language is used for specific purposes that impart both semantic and pragmatic meaning, there is more than just a semantic difference between figurative and assumed literal versions of figurative utterances, and comparisons between the two may be misguided and serve to perpetuate a false dichotomy between the two. Therefore, investigations of figurative language processing would benefit from moving beyond comparing literal and figurative utterances that are assumed to be equivalent, and instead focus on other factors that shape figurative language comprehension, such as individual differences, linguistic features, and pragmatic information. At the same time, research would benefit from incorporating different types of figurative stimuli in addition to metaphor and sarcasm, such as satire.

This review has also highlighted the relative lack of studies investigating figurative language production. From the studies that do exist, results suggest that individual differences play a strong role in determining the quality of figurative language that is produced, and that developmental patterns of figurative language use related to age and language proficiency are important considerations. However, because of the focus on figurative language comprehension studies, there is still a relative lack of investigation on figurative language production, and more research is necessary in order to better understand how other factors, such as individual differences, linguistic features, and affective perceptions may shape the process of figurative language production.
3 INDIVIDUAL DIFFERENCES AND LINGUISTIC FEATURES OF METAPHOR AND SARCASM PRODUCTION

3.1 Introduction

The purpose of this study is to investigate the roles of individual differences (e.g., demographic features, language background, and cognitive ability) and linguistic features on figurative language production quality (specifically, metaphor and sarcasm production). As mentioned in Chapter 2, investigations of figurative language production are relatively scarce when compared to investigations of figurative language processing (Benedek et al., 2014; Corts & Meyers, 2002; J. Johnson & Rosano, 1993). Of the studies that have investigated figurative language production, the majority of them have focused primarily on how different cognitive measures (e.g., intelligence) influence metaphor production (Beaty & Silvia, 2013; Benedek et al., 2014; Blasko, 1999; Chiappe & Chiappe, 2007; Pierce & Chiappe, 2008; Silvia & Beaty, 2012), with some studies focusing on sarcasm production (Huang et al., 2015; Ivanko, Pexman, & Olineck, 2004). Individually, these studies have identified links between metaphor production quality and intelligence (Beaty & Silvia, 2013; Silvia & Beaty, 2012), working memory capacity (Blasko, 1999; Chiappe & Chiappe, 2007; Pierce & Chiappe, 2008), and between sarcasm production and abstract thinking (Huang et al., 2015). In addition to cognitive individual differences such as intelligence and working memory capacity, there is also growing evidence that one’s language background (i.e., first language status, second language proficiency) influences figurative language production ability (Golden, 2010; Heredia & Cieślicka, 2015; J. Johnson & Rosano, 1993; Philip, 2010).

Aside from studies investigating participant individual differences, there have also been studies that highlight the benefits of investigating the linguistic features of figurative language as
a method to better understand figurative language use. For example, computational models of metaphor have been successful at modeling the conceptual comparisons used in metaphors through quantitative measures of semantic distance between lexical items, which provides insight into how the mind organizes difference concepts (Kintsch, 2008; Kintsch & Bowles, 2002). Additionally, sarcastic language can be distinguished from non-sarcastic language through lexical features related to sentiment and affect (e.g., negative emotions, words related to certainty; Campbell & Katz, 2012).

As a whole, these studies identify several influences on figurative language use, but more research is needed in order to determine whether these individual differences are predictive when considered all at once (i.e., measuring the same construct), whether the same individual differences apply to metaphor and sarcasm, and whether additional linguistic features, such as lexical sophistication, can predict figurative language production quality. Therefore, the purpose of this study is to further investigate how individual differences and linguistic features influence figurative language production quality. To do so, the first goal of this study is to investigate a wide range of cognitive individual differences (i.e., crystallized intelligence, fluid intelligence, broad retrieval, working memory capacity, need for cognition, abstract thinking, language analytic ability) and demographic features (i.e., language background, gender, age) on human perceptions of figurative language production quality for metaphors and sarcasm, which capture theorized necessary components for metaphor and sarcasm (i.e., conceptual distance and situational incongruity) as well as generalized conceptualizations of creativity (i.e., original and effective examples of figurative language).

The second goal is to investigate how additional linguistic features such as lexical sophistication and semantic cohesion may help provide a better understanding into the linguistic
properties of metaphor and sarcasm. Specifically, semantic cohesion, or the amount of semantic
distance between metaphorical comparisons may serve to increase or decrease human
perceptions of metaphor production quality. Moreover, additional features of lexical
sophistication, such lexical frequency, psycholinguistic properties of words (e.g., concreteness,
familiarity), and semantic diversity may shed further light on perceptions of metaphor and
sarcasm by identifying additional lexical properties associated with figurative language
production quality.

3.2 Research Questions

The following research questions guide this study:

1. What role, if any, do individual differences such as demographic features, cognitive ability,
   and language background play in the production of metaphor and sarcasm?

2. Can human ratings of the quality of figurative language production be predicted using
   linguistic features of lexical sophistication and semantic cohesion?

3.3 Method

The purpose of this study is to investigate the influence of individual differences and
linguistic features on figurative language production quality. To do so, human participants were
recruited to complete a series of metaphor and sarcasm production tasks and to also complete
surveys and tests designed to measure a wide variety of individual differences. Human raters
then rated the participants’ metaphors and sarcasm for figurative language production quality,
which were also analyzed using automatic text analysis tools capable of automatically measuring
lexical sophistication and semantic cohesion. The measures of individual differences and
linguistic features were then used in subsequent statistical analyses in order to investigate
whether any of these features were predictive of the human ratings of figurative language
production quality, which would suggest a meaningful relation between the two and provide further insight into figurative language production.

### 3.4 Participants

A total of 61 participants were included in this study. Table 3.1 displays demographic information for these participants.

#### Table 3.1 Participant demographic information.

<table>
<thead>
<tr>
<th>Gender</th>
<th>Female</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left-Handed</td>
<td>46</td>
<td>15</td>
</tr>
<tr>
<td>Right-Handed</td>
<td>10</td>
<td>51</td>
</tr>
<tr>
<td>NES</td>
<td>29</td>
<td>32</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Age</th>
<th>M</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>25.56</td>
<td>8.341</td>
<td>17</td>
<td>63</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NNES English Age of Onset</th>
<th>M</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>8.125</td>
<td>3.966</td>
<td>1</td>
<td>17</td>
</tr>
</tbody>
</table>

### 3.5 Materials

#### 3.5.1 Demographic survey.

A demographic survey was used to collect information related to participants’ age, sex, handedness, college year, and language background information, including first language status, total number of languages spoken, proficiency of languages, and relative levels of English use. The demographic survey questions are presented in full in Appendix A.

#### 3.5.2 Fluid intelligence test.

Fluid intelligence is a measure of problem-solving ability, and has previously been shown to influence the ability to construct original metaphors (Beaty & Silvia, 2013; Silvia & Beaty, 2012). Fluid intelligence can relate to several different types of problem-solving, such as inductive and deductive problem solving (Flanagan, 2013). For this study, an original test of
Fluid intelligence was designed to measure inductive problem-solving using a letter sets task, similar to the tests used in previous investigations of metaphor production quality (Silvia & Beaty, 2012). An original task was developed because freely available test items only included one or two sample test items, (e.g. Guilford, 1959) which were used as a baseline to develop a full 16-item test. In the letter sets task designed for this study, participants were shown five sets of different four letter combinations (e.g., ABCD), and decided which of the five sets did not follow a pattern that existed in the other four sets. For example, four of the sets followed alphabetical order or contained a consistent consonant-vowel pattern, while a fifth set deviated from the pattern. The letter set task designed for this study employed four different patterns (specific vowel-consonant order, forwards alphabetical order, containing a specific letter, or containing only one vowel), and four sets were designed for each pattern. Therefore, a total of 16 test items were designed, for a maximum total of 16 points possible on the test.

Instructions for this test stated that patterns for the letters were not based on the sounds or shapes of the letters, or whether the letters form words in English. The instructions also included two examples which demonstrated how to choose the correct answer based on alphabetical order. Finally, the instructions clarified that alphabetical order is just one of the possible patterns in the letter sets, and that there were also additional patterns that will occur. After viewing the instructions, the test items were presented in isolation without the instructions. The letter sets task and its instructions are presented in Appendix B.

3.5.3 **Crystallized intelligence test.**

Crystallized intelligence is a measure of one’s declarative background knowledge and has previously been reported to aid in the construction of conventional metaphors (Silvia & Beaty, 2012). In order to measure crystallized intelligence, a 30-question multiple-choice world
knowledge test was given to the participants. The test contained 10 questions each about literature, science, and world history, and each test item included four possible answers (one correct and three incorrect). The test is based on SAT level questions and was originally used to prepare high school students in New York state to complete state-wide standardized examinations, and has previously been used to investigate students’ reading abilities and performance on science examinations (O’Reilly & McNamara, 2007), to examine the effect of prior knowledge on reading strategies (O’Reilly, Best, & McNamara, 2004), and to examine the effect of prior knowledge on satirical headline comprehension (Skalicky & Crossley, in press). The instructions explained the three domains of knowledge included in the test and the total number of items on the test. The total possible score for this test was 30 points. This test and its instructions are presented in Appendix C.

### 3.5.4 Broad retrieval tests.

Broad retrieval measures one’s ability to access lexical knowledge (Flanagan, 2013), and has previously been reported to aid in the construction of novel metaphors (Silvia & Beaty, 2012). Three separate word generation tasks were designed to measure broad retrieval: a synonym generation a task, a letter prompt generation task, and a category generation task, similar to the ones used in previous metaphor research (Beaty & Silvia, 2013). Each task asked the participant to generate as many possible words that follow the task prompt in a single minute. For the synonym generation task, participants generated as many synonyms as possible for the word bad. For the letter prompt task, participants generated as many words as possible that start with the letter m. For the category generation task, participants generated as many words as possible that represent different occupations. Total scores for each of the tasks were the sum of
correct answers produced, and correct spelling was not necessary in order for a word to be counted as correct (Flanagan, 2013).

3.5.5 Working memory capacity test.

Working memory capacity is a measure of information that can be held in the mind while also completing cognitively complex tasks (Cowan, 2005). The executive control function of working memory capacity (i.e., ability to control information while solving tasks) has been shown to influence metaphor production and processing, as well as sarcasm processing (Blasko, 1999; Chiappe & Chiappe, 2007; Kaakinen et al., 2014; Olkonemi et al., 2016; Pierce & Chiappe, 2008). In this study, a customized run span working memory test was developed. These tests have been shown to reliably assess working memory capacity (Foster et al., 2015). During the test, participants viewed a random set of letters, one at a time, and were then asked to remember a random number of letters from the end of the set. In total, five different span recall ranges were used, containing three, four, five, six, or seven letters. Each span recall range included three trials, with one trial including no extra letters, one trial including one extra letter, and one trial including two extra letters beyond the span recall range. For example, in the five letter recall range, participants saw a trial containing five, six, or seven letters total (while always having to recall the last five letters). Participants were told in advance how many letters they must remember, but were never told how many total letters will be shown. The recall ranges are presented in a random order, with each set of three trials in each recall range presented together in random order (i.e., block randomization). The test also included a practice session using a span range of two. Each correctly recalled letter in the correct position was assigned one point, for a total maximum score of 85.
3.5.6 Abstract thinking survey.

Abstract thinking represents the tendency for one to perceive events in either a more concrete or a more abstract manner, and one previous study has demonstrated that exposure to sarcasm can lead to more abstract perceptions of events (e.g., describing the activity of reading as gaining knowledge as opposed to following lines of print; Huang et al., 2015), suggesting that figurative language comprehension and abstract thinking are linked. Abstract thinking was measured using the Behavior Identification Form (BIF; Vallacher & Wegner, 1989), which is a list of 25 actions with two descriptions of each action, and participants were asked to choose which description best identifies their interpretation of the action. One description represents a concrete interpretation of the action, whereas the other description represents an abstract interpretation of the action (see example above describing reading). The BIF was scored by assigning one point for each answer representing the abstract interpretation of an action, for a total maximum score of 25 points. This test and its instructions are presented in Appendix D.

3.5.7 Need for cognition survey.

Need for Cognition (NFC) represents the desire to perform complex cognitive tasks, and there exists some evidence that NFC is important for metaphor (but not sarcasm) processing (Kaakinen et al., 2014; Olkoniemi et al., 2016). NFC was measured using a survey designed to capture participants’ inclinations to perform cognitively complex tasks (Cacioppo et al., 1984). This 18-item survey asked participants to rate their agreement using a scale of -4 (very strong disagreement) to 4 (very strong agreement) for statements such as I like tasks that require little thought once I’ve learned them. Participants’ NFC score was the sum of all answers, with a total score range of -72 (low NFC) to 72 (high NFC). This test and its instructions are presented in Appendix E.
3.5.8 Language aptitude test.

Language aptitude refers to an inherent ability to learn additional languages based on one’s ability to both explicitly and implicitly detect and subsequently comprehend and produce grammatical patterns and sound systems of a language (Dörnyei & Skehan, 2003; Li, 2015). While no previous research has associated language aptitude with figurative language use, explicit aspects of language aptitude (e.g., grammatical sensitivity) may nevertheless be an important predictor of figurative language use because it represents a latent ability to consciously understand the semantic and grammatical relations among words and phrases in a language, which figurative language exploits in unconventional ways. Specifically, the measure of language aptitude used in this study taps into participants’ language analytic ability, which is associated with explicit grammatical sensitivity (i.e., ability to consciously decode grammatical and semantic roles of words in context) and pattern recognition in languages (i.e., ability to detect morphology patterns in a new language and extrapolate them to new contexts; Dörnyei & Skehan, 2003; Li, 2015). Participants’ language analytic ability was measured using the Language Analysis test (Schmitt, Dörnyei, Adolphs, & Durow, 2004). The Language Analysis test presented participants with words and phrases from a fictional language with their respective English translations in a table. Participants were then randomly shown 14 English sentences, one at a time, with four possible translation options (only one of which was correct), and participants were asked to choose the correct translation option for each of the sentences. The translation table was available to participants during each question. Participants were assigned one point for each correct item, for a total maximum score of 14 points. This test and its instructions are presented in Appendix F.
3.5.9 **Metaphor production items.**

Two different metaphor production tasks were developed using stimuli from previous metaphor production experiments (Beaty & Silvia, 2013; Chiappe & Chiappe, 2007), which were also modified to include vocabulary more suitable for second language English users. First, a conventional metaphor task was designed containing 22 different items. Each item consisted of a Topic and a Description. All of the Topics were nouns (e.g., *her family*), and all of the Descriptions were descriptions or properties of those nouns (e.g., *something that keeps her stable and prevents her from drifting into danger*), and participants were instructed to use the Description of the Topic to write a metaphor reflective of the same meaning in the Description, but without reusing any of the words from the Description. In addition, a novel metaphor task was developed, also based on the same previous studies. In this task, participants were presented with two scenarios: the most boring class they have been in, and the most disgusting item they have ever eaten or drank. For each scenario, participants were instructed to produce a metaphor that described their feelings during that scenario and were also provided with an example of how to start their metaphors (e.g., *Being in that class was like ____*). The full list of metaphor production stimuli for the conventional and novel metaphors and participant instructions are presented in Appendix G.

3.5.10 **Sarcasm production items.**

Twelve different drawn cartoons were adapted or created to serve as sarcasm production prompts. Four of these items were black and white cartoons used by Huang et al. (2015) to prompt sarcastic responses. These four cartoons were taken from the Rosenzweig Picture Frustration Study, originally designed to assess patient responses to frustrating situations in order to diagnose aggression (Rosenzweig, 1945). Each of the black and white cartoons is a single-
panel cartoon which depicts a frustrating situation with more than one speaker (e.g., one person’s car broke down and thus two people missed their train). The person responsible for the frustration is shown saying something, whereas the victim of the frustration is presented with a blank speech bubble. Four additional items were created by revising four single-panel *Bizarro!* comics. *Bizarro!* is a single-panel comic strip created by Dan Piraro that is syndicated daily in print newspapers across the United States. *Bizarro!* comics typically depict absurd or otherwise unlikely situations for the purpose of humor, social commentary, or both (www.bizzaro.com). The specific *Bizarro!* comics used in this study are four desert island comics, which each depicted two people stranded on a small desert island in the middle of an ocean. The original cartoons all contained a single speech bubble for one of the speakers, which was made blank for the purposes of this study. Finally, an additional four sarcasm production items were developed by creating original comics online using a free comic creation website (www.pixton.com). These four comics are each comprised of three panels, and each involve two speakers. In each comic, the first two panels set up an initial situation (e.g., a young man is recruited to join the army and is guaranteed to travel the world in an exciting manner by a military recruiter), while the final panel includes one of the speakers with an empty speech bubble in a situation designed to prompt a sarcastic response (e.g., the young man ends up peeling potatoes instead of traveling the world). For each of the twelve comics, participants were instructed to imagine they were the speaker with the empty speech bubble and to write something sarcastic they would say if they were in that situation. The full list of sarcastic production stimuli and instructions are presented in Appendix H.
3.6 Procedure

3.6.1 Online survey.

Participants first completed an online survey from a location and computer of their choice using the online survey platform Qualtrics (Qualtrics, Provo, UT), which included the demographic and language background questionnaires and the fluid intelligence, broad retrieval, abstract thinking, need for cognition, and language aptitude tests (see Appendices A, B, D, E, and F). Each participant first read and acknowledged an online version of the informed consent document (also presented in Appendix A). Participants were instructed to allot a single hour for the test and to reduce all outside distractions such as music, television, or other people. They were also instructed to complete the survey using a laptop or desktop computer (and not using a tablet or mobile phone).

After acknowledging the informed consent and instructions, all participants first completed the demographic survey information. Participants who indicated their first language was not English were asked additional questions related to their English proficiency. Afterwards, the individual differences tasks listed above were presented in a randomized order (with questions within each test also presented in a random order). Each individual difference test first displayed the instructions listed in the respective Appendices. All participants reported the online survey required approximately 30 to 45 minutes to complete.

3.6.2 Laboratory session.

After completing the online survey, participants scheduled a time to complete the laboratory portion of the experiment, which included the working memory capacity test, crystallized intelligence test, and metaphor and sarcasm production tests. On average, participants completed the laboratory experiment two days after completing the online survey.
portion, with some participants completing both on the same day, and four completing the laboratory session over a week after the survey (maximum time was 11 days after the survey for one participant, the remainder all completed the laboratory session less than 8 days after the online survey). Upon arriving for the laboratory session, participants first reviewed and signed a printed version of the informed consent form they initially encountered during the online survey. Participants were then seated in a soundproof room with a desktop computer in order to complete the crystallized intelligence (i.e., 30-question multiple world knowledge choice test; Appendix C) and working memory capacity tests. Participants completed the world knowledge test during the laboratory session in order to reduce the temptation to locate answers to the questions online. All participants completed the world knowledge test first. Before starting the world knowledge test participants were verbally instructed by the experimenter that they were to take a multiple choice test where each answer contained four possible answers (only one of which was correct) related to science, world history, and literature. Participants were instructed to make their best educated guess should they not be sure of their answer, and to read the instructions in full before initiating the test.

After finishing the world knowledge test, participants then completed the working memory capacity test using E-Prime software (Psychology Software Inc., Sharpsburg, PA) on the same desktop computer. Participants were first verbally instructed on how the working memory capacity test is administered, provided a brief example of a letter span, and then told to read the instructions in full and complete the practice session before initiating the test. After finishing the world knowledge and working memory capacity tests, participants then completed the metaphor and sarcasm production tasks using E-Prime software on the same desktop computer. Participants were first asked by the experimenter if they were familiar with the
concepts of metaphor and sarcasm. If a participant indicated uncertainty with either term, the experimenter provided examples until the participant reported they were now certain of what the terms meant. The experimenter then verbally described the metaphor and sarcasm production tasks and asked the participants to be as creative as possible while also aligning their answers as best they could to the definitions of metaphor and sarcasm supplied in the instructions. Finally, the experimenter asked if the participants had any remaining questions, and answered them if so. Once participants indicated they had no questions and were ready to proceed, the experimenter left the participant in the soundproof room to complete the tasks, reminding the participant to read the instructions in full before initiating the task. The metaphor and sarcasm tasks were then presented to participants in a random order. Once participants completed the laboratory session, they were given a $25 Amazon.com gift card for their participation.

3.6.3 Metaphor production tasks.

During the metaphor production tasks, participants were first provided with a definition of metaphor and several examples of metaphors (none of the examples were repeated in the stimuli). The definition of metaphor provided to participants was *a metaphor is a comparison between two things in order to help describe something* (see Appendix G for full instructions and examples). Once the instructions were read, participants began the task, with the metaphor production trials presented in a random order. During each trial, participants were presented with a screen that displayed the Topic and Description in clearly marked areas of the screen, with a blank text box for the participants to type their metaphor using the keyboard. Additionally, the definition of metaphor was presented on the screen for participants during each trial, as was a reminder of their task and the need to avoid repeating words listed in the Description. Figure 3.1 displays an example of one of the metaphor production item stimuli screens participants were
shown. After typing their metaphor into the answer box, participants pressed the Enter key which then moved to an inter-trial screen instructing participants to push the space bar to see the next metaphor prompt. After completing all 22 conventional metaphor prompts, participants then completed the two novel metaphor situations in a randomized order, also providing their answers by typing their metaphor into the text box.

![Figure 3.1 Sample conventional metaphor screen from E-Prime](image)

For the novel metaphors, participants were provided with example sentences to help them begin their metaphors (see Figure 3.2 for an example). The written participant answers and time spent creating their answers (in milliseconds) was recorded by the E-Prime software.
3.6.4 Sarcasm production tasks.

During the sarcasm production section, participants were provided with a definition of sarcasm and given several different examples of sarcasm. The definition of sarcasm provided to participants was *sarcasm is a form of indirect language. When someone is being sarcastic, they mean something different than what they literally said* (see Appendix H for full instructions). Once the instructions were read, participants began the task. Each trial involved one of the 12 comics randomly displayed above a text box, with a reminder asking participants to supply a sarcastic comment for the situation depicted in the comic (see Figure 3.3 for an example). After typing their sarcastic statement into the answer box, participants pressed the Enter key which then moved to an inter-trial screen instructing participants to push the space bar to see the next sarcasm prompt. After providing an answer for all 12 comics, participants completed the sarcasm production experiment. The written participant answers and time spent creating their answers (in milliseconds) was recorded E-Prime.
Figure 3.3 Sample sarcasm prompt using three-panel comic from E-Prime

3.6.5 Human ratings.

An analytic rubric was created in order to obtain measures of figurative language production quality for the metaphor and sarcasm answers provided by the participants. The rubric contained separate sections for metaphor and sarcasm, and was designed to capture quality measures specific to each type of figurative language as well as general aspects of creativity itself based off previous rubrics designed to measure creativity in metaphor production (Beaty & Silvia, 2013; Silvia & Beaty, 2012). Specifically, the rubric for both metaphors and sarcastic answers was comprised of three separate subscales designed to capture metaphor or sarcasm quality based on participants’ ability to develop accurate, effective, and original examples of metaphor and sarcasm. Accuracy was related to theoretical definitions of metaphor and sarcasm, while effectiveness and originality were related to theoretical definitions of creativity.
Accordingly, the metaphor section included the subscales Conceptual Distance, Novelty, and Mirth. The Conceptual Distance subscale was designed to capture the relative conceptual distance between the two entities contained in a metaphor (i.e., between the noun listed in the Topic and the other information listed in the metaphor answer), which captures the theoretical definition of metaphor (Gibbs, 1994; Glucksberg, 2001). Although metaphors do not need to include large gaps in conceptual domains in order to be defined as a metaphor, the ability to create descriptive links between seemingly disparate concepts is fundamental to metaphor production (Kintsch, 2008; Kintsch & Bowles, 2002), and therefore metaphors with greater conceptual distance may also be more creative.

The Novelty subscale was designed to capture the originality of the metaphor based on other examples provided by participants as well as general knowledge of metaphors commonly used in English, capturing the originality aspect of cognitive definitions of creativity (Runco & Jaeger, 2012). The Mirth subscale was designed to capture whether the metaphor was interesting, humorous, or clever (or any combination of these elements). The concept of Mirth was used to capture the emotional reaction typically associated with humor, wherein one can experience slight amusement to intense hilarity arising from humorous or playful stimuli (Martin, 2007), which provided an additional measure of effectiveness.

The sarcasm section was modeled on the metaphor rubric but did not include Conceptual Distance because that is not a feature relevant to sarcasm. The sarcasm section thus included the subscales Novelty, and Mirth, and, the subscale Incongruity was added. The Incongruity subscale was designed to capture whether the literal form of the sarcastic comment clashed in some way with the intended meaning of the comment, capturing essential features of sarcasm based on theoretical definitions of verbal irony (Colston, 2017; Colston & Gibbs, 2007; Gibbs & Colston,
2012), much like the *Conceptual Distance* subscale for metaphors. The *Novelty* and *Mirth* subscales used the same definitions as the metaphor subscales described above for the same reasons. Each subscale was measured using a range of one through six, with a score of one meaning the answer did not meet the criterion in any way and a score of six meaning the answer met the criterion in every way. The analytic rubric with benchmarks for each rating level is presented in full in Appendix I.

Two human raters were recruited to provide ratings of the participants’ metaphor and sarcasm answers using this analytic rubric. Both raters were undergraduate majors in applied linguistics in their fourth and final year. The first rater was a 22 year old male who spoke English as a native language, Romanian fluently as a heritage language, and Japanese at intermediate proficiency. The second rater was an approximately 22 year old female who spoke English as a native language. The raters first trained on a corpus of metaphors and sarcastic replies collected from previous pilot sessions using the same stimuli before rating the data obtained here. After initial ratings, a third rater (i.e., the author) adjudicated any disagreements of two points or greater for all of the subscales, and also removed any participant answers that were not examples of sarcasm or metaphor. Table 3.2 displays the inter-rater reliability measures (Pearson correlations and Cohen’s Kappa) before and after adjudication. After adjudication, the raters’ scores were averaged to provide a single score per subscale per item, and a total of 1304 metaphors and 716 sarcasms remained.

| Table 3.2 Inter-rater reliability scores for human ratings of metaphor and sarcasm. |
|---------------------------------------------|------|------|
| *Metaphor Initial*                         |     |     |
| Conceptual Distance                        | 0.735 | 0.709 |
| Novelty                                    | 0.604 | 0.593 |
| Mirth                                      | 0.605 | 0.597 |

| *Metaphor Adjudicated*                     |     |     |
| Conceptual Distance                        | 0.872 | 0.872 |
| Novelty                                    | 0.862 | 0.854 |
82

<table>
<thead>
<tr>
<th>Sarcasm Initial</th>
<th>r</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incongruity</td>
<td>0.557</td>
<td>0.530</td>
</tr>
<tr>
<td>Novelty</td>
<td>0.502</td>
<td>0.453</td>
</tr>
<tr>
<td>Mirth</td>
<td>0.411</td>
<td>0.373</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sarcasm Adjudicated</th>
<th>r</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incongruity</td>
<td>0.867</td>
<td>0.835</td>
</tr>
<tr>
<td>Novelty</td>
<td>0.831</td>
<td>0.783</td>
</tr>
<tr>
<td>Mirth</td>
<td>0.819</td>
<td>0.777</td>
</tr>
</tbody>
</table>

3.6.6 Linguistic features.

The metaphors and sarcastic responses produced by the participants were analyzed for lexical sophistication and semantic cohesion using the automatic text analysis tools TAALES 2.0 (Kyle et al., 2017) and TAACO 2.0 (Crossley et al., 2016), respectively. Before the analysis, participants’ answers were separated into individual text files and processed through an automatic spelling correction program developed in the Python computer language. This program functions by comparing each word in each text against a pre-existing list of correctly spelled words using the autocorrect Python package. If a word does not appear on the list, the program identifies the closest possible match based on orthographic probabilities (i.e., the most likely word based on valid letter combinations in English) and replaces it with a correctly spelled word. A total of 156 spelling corrections were made to the files. These corrections were manually reviewed for accuracy, and any change that was incorrect was manually corrected (45 changes were manually corrected). Once this process was completed, the files were then processed using TAALES and TAACO.

From TAALES 2.0, several indices representative of lexical sophistication were calculated. To date, very few studies have investigated lexical sophistication in the context of figurative language, aside from one study reporting that satirical product reviews were less concrete than non-satirical product reviews (Skalicky & Crossley, 2015). As mentioned in
Chapter 2, there is a need to perform more investigation into lexical sophistication and figurative language in order to better determine if these features interact with perceptions of figurative language quality. Therefore, this study includes broad measures of lexical sophistication related to lexical frequency, psycholinguistic properties of words, and word exposure in order to investigate and report any initial links between figurative language production quality and lexical sophistication.

First, measures of psycholinguistic properties of words were gathered because these measures represent cognitive representations of lexical items and can be used to assess the relative sophistication of lexical items (Kyle & Crossley, 2015). Specifically, these measures were word Familiarity, Concreteness, Imageability, and Meaningfulness. Word Familiarity represents how familiar one is with a specific word, with more familiar words being words that are also more commonly encountered, making familiarity similar to word frequency (Crossley & Skalicky, in press). Word Concreteness refers how perceptible an entity associated with a particular word is (Brysbaert, Warriner, & Kuperman, 2014). For example, the word dog is more concrete than the word music. Word Imageability represents the ease of conjuring a mental image of a word, with words like tree being more imageable than words such as abatement (Salsbury, Crossley, & McNamara, 2011). Word Meaningfulness represents how many different associations to other words a particular word has. For example, a word such as tree has more associations (e.g., branch, leaf, wood) than a word such as savant, which activates fewer associations (Salsbury et al., 2011). Measures of word Imageability, Familiarity, and Meaningfulness were all calculated based on the MRC Psycholinguistics Database norms (Coltheart, 1981), which is a curated compilation of previous rating studies for these features. Word Concreteness values were calculated using the Brysbaert Concreteness norms (Brysbaert et
al., 2014), which were derived from human ratings of word concreteness using online crowdsourcing.

In addition to those indices, linguistic features related to word exposure and use were also calculated, as these represent the relative frequency of occurrence and use for certain words, which is also a measure of lexical sophistication. These indices were spoken word frequency, semantic diversity, and age of acquisition. Spoken word frequency was calculated using counts from the spoken portion of the Corpus of Contemporary American English (COCA; Davies, 2008). Semantic Diversity represents the number of different word contexts a particular word typically occurs in, and thus represents specificity of word meanings. Semantic Diversity was calculated for each word using the norms published by Hoffman, Lambon Ralph, and Rogers (2013). To calculate Semantic Diversity, Hoffman et al. (2013) separated the British National Corpus into chunks of 1,000 words, and then analyzed the total number of these 1,000 word contexts any particular word occurred in, as well as the semantic similarity of each word to all of the other words in those contexts. The end result is that words with higher Semantic Diversity can be used in more contexts and have more variable meanings than those with lower Semantic Diversity. Finally, Age of Acquisition (AoA) values represent human intuition regarding the age when they first learned a particular word. AoA values based on Kuperman et al., (2012) were used, which were collected using a large number of human raters via online crowdsourcing. The number of words in the participant utterances was relatively low but also relatively stable across figurative language types (Sarcasm $M = 10.61$, $SD = 5.88$, Conventional Metaphor $M = 8.95$, $SD = 4.62$, Novel Metaphor $M = 13.81$, $SD = 6.68$, All Metaphors $M = 9.36$, $SD = 5.01$, All Responses $M = 9.77$, $SD = 5.35$), and therefore all of the linguistic indices were calculated based
on content words only in order to analyze words carrying the majority of semantic information in the responses.

TAACO 2.0 was used in order to calculate semantic overlap between prompts and participant answers for the metaphors. Distance between concepts used in metaphors has been accurately modeled using measures of semantic association, such as Latent Semantic Analysis (Kintsch, 2008; Kintsch & Bowles, 2002), and therefore a measure of semantic distance was included in this study in order to determine if distance between concepts influences human perceptions of metaphor production quality. To do so, the participants’ metaphors were grouped by prompt and analyzed separately using the source text analysis option in TAACO 2.0. This option allows the user to load in a source text as a reference text for other texts to be compared against for semantic and cohesive similarity or differences. For each group of metaphors, the Description provided to the participants was loaded as the source text, and the participant’s metaphor were analyzed to gather the amount of semantic overlap between participants’ answers and the prompts using the Word2vec measure in TAACO 2.0. Word2vec models the semantic direction and magnitude of words as they relate to other words (known as vectors). By modeling words as vectors, Word2vec assumes words more closely grouped together are more semantically related than those that are further apart and employs predictive modeling in order to calculate the semantic relations among words in a text.

3.6.7 Language background.

Participants were assigned to either native-English speaker (NES) or non-native English speaker (NNES) groups based on their answers to the survey questions. Initial attempts to further separate participants into more discrete categories (e.g., total number of languages, overall proficiency of additional languages, relative levels of bilingualism) resulted in a high number of
imbalanced groups (i.e., many groups with low N sizes), and therefore the simple NES/NNES binary was employed. Because this binary contrast between participants masks some of the more nuanced features of language knowledge, such as individual exposure to English and influence of one’s first language on English knowledge, two additional variables were calculated. First, participants’ self-reported age of English onset was included as a separate language background variable, with NES speakers assigned an onset age of zero. Additionally, a numeric measure of linguistic distance between a participants’ first language and English was also included. When learning an additional language, research has demonstrated that learners will rely on knowledge of their first language while learning an additional language (Miller & Chiswick, 2005; Odlin, 2003; Ringbom & Jarvis, 2011). Therefore, a measure of language distance was assigned to each participant indicating the distance between their first language and English using the framework devised by Miller and Chiswick (2005). Miller and Chiswick (2005) assigned a score ranging between one and three for languages other than English which is reflective of the overall difference between that language’s linguistic features when compare to English. A lower score indicates more distance from English when compared to a higher score. In order to include native English speakers in this calculation, the scores were first inversely transformed by dividing the language distance score by one, so that languages further away from English were represented with higher scores, and languages closer to English represented with lower scores. Native English speakers were assigned a language distance score of zero.

3.7 Statistical Analysis

Several analyses were first conducted to prepare the data for subsequent statistical analyses designed to answer the research questions using R statistical software (R Core Team, 2017). First, dimension reduction via Principal Component Analysis using the \textit{prcomp} function
in the *psych* package (Revelle, 2017) was used to collapse the word generation tasks into a single weighted score representative of broad retrieval. The same method was also used to investigate whether the human ratings of metaphor and sarcasm could be collapsed into single, weighted scores of figurative language production quality. Next, descriptive statistics and correlations using the *cor.test* function in the base *stats* package in R were calculated for the individual differences tests, the linguistic features of the metaphors and sarcastic responses produced by the participants, and the human ratings of the metaphor and sarcastic responses in order to examine statistical assumptions of distribution and linear relation to the dependent variables.

Finally, a series of linear mixed effects models (LME) were constructed in order to systematically test the research questions guiding this study. LMEs are constructed and interpreted in a similar manner to multiple linear regression models, in that a number of predictor variables (i.e., independent variables) are examined to determine whether they explain a significant amount of the variance in a dependent variable. In an LME, predictor or independent variables are commonly named *fixed effects*, and can include either independent variables related to testing a hypothesis and also variables which a researcher may want to control for (e.g., initial language proficiency, age; Baayen, Davidson, & Bates, 2008). Additionally, LMEs allow for the inclusion of *random effects*, which are most commonly the participants and items used in a study. By modeling participants and items as random effects, LME models calculate individual regression slopes for each participant and each item (as opposed to partially pooled means as used in traditional multiple regression analyses). This has the advantage of allowing LMEs to account for non-independence of data (i.e., a repeated measures design), and also provides a high level of statistical power to find even small effects. Finally, LME models also allow for the inclusion of *random slopes*, which account for variance in random effects in light of particular
fixed effects. Random effects are constructed to account for between participant and item variation, while random slopes are constructed to account for within participant and item variation (Barr, Levy, Scheepers, & Tily, 2013).

There are several ways to analyze the output of LME models. One method advocated for hypothesis testing is to fit a fully defined model with all fixed effects, random effects, and random slopes of interest, including all possible interaction terms (Barr et al., 2013), whereas others advocate for the use of model comparisons to determine which random effects and random slopes are necessary for a particular model (Baayen et al., 2008). Regardless of which approach is chosen, the coefficients of significant main effects and interactions are interpreted in order to determine which independent variables exert a significant effect on the dependent variable.

All LME models for this study were constructed in R using the lme4 (Bates, Mächler, Bolker, & Walker, 2015) and lmerTest packages (Kuznetsova, Brockhoff, & Christensen, 2016). In addition, measures of effect sizes for the models were obtained through the MuMIn package (Nakagawa & Schielzeth, 2013), which provides two measures of explained variance: a marginal $R^2$ for the fixed effects only, and a condition $R^2$ for both the fixed and random effects. The marginal $R^2$ best represents the $R^2$ that would be obtained through a traditional linear model analysis, whereas the conditional $R^2$ is specific to LME as it includes variance captured in the random effects structure.

Models were constructed for both participant individual differences and linguistic features for the metaphors and the sarcastic responses. In each model, the metaphor or sarcasm production score obtained from the human ratings was entered as the dependent variable. For the models testing participant individual differences, participants’ age, sex, English age of onset,
Abstract Thinking, Need for Cognition, Fluid Intelligence, Crystallized Intelligence, Broad Retrieval, Working Memory Capacity, and Language Aptitude scores were entered as fixed effects (i.e., predictor variables). In addition, metaphor type (conventional vs. novel) was entered as a fixed effect for the metaphor models, and sarcastic prompt type (three-panel comic, desert island, or black and white picture) was fit for the sarcasm models. Participant time spent creating the metaphor or sarcastic response as well as the trial order of the prompt were also entered as fixed effects. For metaphors, interactions were fit between English age of onset and individual differences, between metaphor type and individual differences, and between metaphor production time and individual differences. For the sarcastic responses, interactions were fit between English age of onset and individual differences, between sarcasm prompt type and individual differences, and between sarcastic responses production time and individual differences. Only significant interactions were retained in the final models in order to properly interpret main effects. Participants and items were entered as random effects for both metaphors and sarcasm, with a random slope of metaphor type entered on subjects for metaphors. A random slope of sarcasm prompt type entered on subjects for sarcastic responses was attempted, but failed to converge. For the models testing linguistic features, MRC Familiarity, MRC Imageability, MRC Meaningfulness, Brysbaert Concreteness, Kuperman Age of Acquisition, COCA Spoken Word Frequency, and Semantic Diversity were entered as the predictor variables. In addition, word2vec source similarity and metaphor type were entered as predictor variables for the metaphors. Participants and items were entered as crossed random effects. All numerical predictors in all models were z-scored (i.e., centered and scaled) before being input into the models in order to properly interpret interactions and assist with model convergence.
3.8 Results

3.9 Dimension Reduction

3.9.1 Broad retrieval.

Results from the three different word generation tasks designed to measure broad retrieval were analyzed using Principal Components Analysis (PCA). Pearson correlations among the synonym generation ($M = 5.72, SD = 2.89$), letter prompt generation, ($M = 14.43, SD = 3.69$) and occupation generation tasks ($M = 10.30, SD = 4.57$) demonstrated these three variables were all significantly and positively correlated with each other, with $r$ sizes ranging from .386 to .423, indicating suitable levels of correlation for PCA (Levshina, 2015). A subsequent Bartlett’s Test using the `cortest.bartlett` function from the `psych` package (Revelle, 2017) further confirmed suitable correlations among the three variables for dimension reduction $\chi^2(3) = 25.510, p < .001$, and a Kaiser-Meyer-Olkin test of sampling adequacy using the `KMO` function in the `psych` package reported values between .65 and .68 for the three variables, suggesting adequate ability to extract unique components from the data (Field, 2013). The PCA reported three components, with the first component explaining 60% of the variance and including all three word generation measures, with an eigenvalue of 1.343. Eigenvalues for the other two components were below one, suggesting those components should not be retained (Levshina, 2015). This was also supported by visual inspection of the scree plot. Therefore, only the first component was retained. Loadings into this component for the three-word generation tasks were .585 for the synonym generation task, .566 for the letter prompt generation task, and .581 for the occupation generation task. Based on these loadings, a single, weighted score for the word generation tasks was calculated by first multiplying each word generation task score by its
respective loading and then summing these three weighted scores per participant. This new weighted value was used to represent broad retrieval in all statistics reported below.

3.9.2 Sarcasm ratings.

The human ratings of sarcasm for the three subscales (Incongruity, Novelty, and Mirth) were analyzed using PCA for the remaining 716 sarcastic responses after adjudication. Pearson’s correlations among the Incongruity ($M = 4.53$, $SD = .806$), Novelty ($M = 3.118$, $SD = 1.022$), and Mirth ($M = 3.190$, $SD = 1.039$) subscales indicated that the Novelty and Mirth subscales were strongly and positively correlated with each other ($r = .848$, $p < .001$), that Novelty and Incongruity were not significantly correlated ($r = .030$, $p = .408$), and that Mirth and Incongruity demonstrated a significant yet weak positive correlation ($r = .149$, $p < .001$). A Bartlett’s Test confirmed these correlations were suitable among the three subscales for dimension reduction, $\chi^2(3) = 943.422$, $p < .001$, but results from a KMO test reported values between .21 and .49, suggesting poor ability to extract unique components from the data. The PCA reported three components, with the first component explaining 62% of the variance and containing the Novelty and Mirth subscales. The second component explained 33% of the variance and contained the Incongruity subscale. Results from the correlations and PCA suggest that human ratings of Novelty and Mirth captured a single construct, while human ratings of Incongruity represented a different construct, suggesting that these ratings could not be collapsed into a single, weighted score. Therefore, the ratings for Novelty and Mirth were averaged, and the ratings for Incongruity were retained in their original manner, resulting in two dependent variables for the sarcastic responses per item. The Incongruity and averaged Novelty/Mirth scores correlated at $r = .094$, $p = .012$, demonstrating a significant yet weak positive correlation.
3.9.3 Metaphor ratings.

The human ratings of metaphor for the three subscales (Conceptual Distance, Novelty, and Mirth) were analyzed using PCA for the remaining 1304 metaphors after adjudication. Pearson’s correlations among the Conceptual Distance ($M = 4.563$, SD = .875), Novelty ($M = 3.388$, SD = 1.209), and Mirth ($M = 3.253$, SD = 1.204) subscales indicated that the Novelty and Mirth subscales were strongly and positively correlated with each other ($r = .899$, $p < .001$), and also that both of these subscales significantly and positively correlated with the Conceptual Distance subscale at $r = .347$ and .348, both $p < .001$. A Bartlett’s Test confirmed these correlations were suitable among the three subscales for dimension reduction, $\chi^2(3) = 2330.705$, $p < .001$, but results from a KMO test reported values between .54 and .94, suggesting less than adequate ability to extract unique components from the data. The PCA reported three components, with the first component explaining 71% of the variance and containing the Novelty and Mirth subscales, while the second component explained an additional 26% of the variance and containing only the Conceptual Distance subscale. In a similar fashion to the sarcasm ratings, results from the correlations and PCA suggested that the human ratings of Novelty and Mirth were capturing a single construct. The PCA also suggested that the human ratings of Conceptual Distance were capturing a separate construct, suggesting that these ratings could not be collapsed into a single, weighted score. Therefore, human ratings of Novelty and Mirth for each metaphor were averaged, and the human ratings of Conceptual Distance were retained in their original manner, resulting in two separate dependent variables for the metaphors. The Conceptual Distance and averaged Novelty/Mirth scores correlated at $r = .357$, $p < .001$, demonstrating a significant, positive correlation with a medium effect size.

---

2 Separate analyses of the conventional and novel metaphors resulted in the same dimensions, and thus all metaphors were included in one single analysis.
3.10 Descriptive Statistics

3.10.1 Individual differences.

Table 3.3 displays descriptive statistics and measures of internal consistency when applicable (i.e., Chronbach’s alpha [\(\alpha\)]) for each of the individual difference measures\(^3\).

Specifically, Chronbach’s alpha measures the reliability of results obtained from different tests using a series of correlations among all individual items of a test. A higher Chronbach’s alpha suggests higher reliability, in that the same participants would receive approximately the same scores were they to repeat the test again. As Table 3.3 displays, internal consistency among the different individual differences tests ranged from 0.75 to 0.84, demonstrating adequate internal consistency for test items used for all measures of individual differences.

<table>
<thead>
<tr>
<th>Measures</th>
<th>M</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
<th>Range</th>
<th>(\alpha)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract Thinking</td>
<td>13.69</td>
<td>5.09</td>
<td>3</td>
<td>25</td>
<td>0-25</td>
<td>0.81</td>
</tr>
<tr>
<td>Need for Cognition</td>
<td>19.11</td>
<td>17.02</td>
<td>-11</td>
<td>64</td>
<td>-72-72</td>
<td>0.84</td>
</tr>
<tr>
<td>Language Aptitude</td>
<td>8.77</td>
<td>2.90</td>
<td>2</td>
<td>13</td>
<td>0-14</td>
<td>0.78</td>
</tr>
<tr>
<td>Fluid Intelligence</td>
<td>9.41</td>
<td>3.10</td>
<td>3</td>
<td>14</td>
<td>0-16</td>
<td>0.75</td>
</tr>
<tr>
<td>Crystallized Intelligence</td>
<td>19.75</td>
<td>4.80</td>
<td>9</td>
<td>27</td>
<td>0-30</td>
<td>0.78</td>
</tr>
<tr>
<td>WMC</td>
<td>42.15</td>
<td>11.27</td>
<td>15</td>
<td>62</td>
<td>0-85</td>
<td>N/A</td>
</tr>
<tr>
<td>Broad Retrieval</td>
<td>17.67</td>
<td>4.83</td>
<td>9.69</td>
<td>34.51</td>
<td>0-N</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Chronbach’s alpha (\(\alpha\)) was computed in R using the `alpha` function from the `psych` package (Revelle, 2017). No value is displayed for WMC because the experiment did not save individual item level scores and only outputted a total final score per participant. Broad Retrieval is a weighted score derived from a PCA, and thus also has no score of internal consistency. Range represents the possible score range for each test, and there is no upper bound for Broad Retrieval because there was no upper limit to the number of words participants could generate.

3.10.2 Human ratings.

Table 3.4 shows the descriptive statistics for the human ratings of metaphor and sarcasm after rater adjudication and dimension reduction. For metaphors, the conceptual distance and

\(^3\) Chronbach’s alpha is not reported for WMC because the experiment did not save individual item level scores and only outputted a total final score per participant.
novelty/mirth ratings for all participants correlated at \( r = .357, p < .001 \), while the sarcasm incongruity and novelty/mirth ratings for all participants correlated at \( r = .094, p = .011 \).

Table 3.4 Descriptive statistics for human ratings of metaphors and sarcastic responses.

<table>
<thead>
<tr>
<th></th>
<th>English L1</th>
<th>Non-English L1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Metaphor</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conceptual Distance</td>
<td>4.638</td>
<td>4.497</td>
</tr>
<tr>
<td>SD</td>
<td>0.799</td>
<td>0.932</td>
</tr>
<tr>
<td>Novelty/Mirth</td>
<td>3.397</td>
<td>3.254</td>
</tr>
<tr>
<td>SD</td>
<td>1.209</td>
<td>1.142</td>
</tr>
<tr>
<td><strong>Sarcasm</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Incongruity</td>
<td>4.581</td>
<td>4.482</td>
</tr>
<tr>
<td>SD</td>
<td>0.777</td>
<td>0.830</td>
</tr>
<tr>
<td>Novelty/Mirth</td>
<td>3.128</td>
<td>3.179</td>
</tr>
<tr>
<td>SD</td>
<td>1.017</td>
<td>0.965</td>
</tr>
</tbody>
</table>

Note. Rating scale ranged from 1-6.

3.11 Correlations

3.11.1 Individual differences.

Table 3.5 displays Pearson’s correlations among the different individual differences. A total of eight significant correlations among the measures were reported, with absolute \( r \) values ranging between .252 and .430, suggesting no evidence of strong multicollinearity among these measures.

3.11.2 Individual differences and production measures.

Table 3.6 displays Pearson’s correlations among the participant individual differences and human ratings for metaphors and sarcasms. Only correlations reporting at least a small effect size (i.e., absolute \( r \geq 1 \)) are considered meaningful and significant. Three variables significantly correlated with the novelty/mirth ratings of metaphors, while six variables significantly correlated with either the incongruity or novelty/mirth ratings of sarcastic responses. Despite being reported as significant, none of these correlations exceeded a small effect size (absolute \( r > .30 \); Cohen, 1992), suggesting relatively weak correlations among these features and the human ratings of metaphor and sarcasm.
Table 3.5 Pearson's correlations among the cognitive individual differences measures.

<table>
<thead>
<tr>
<th></th>
<th>WMC</th>
<th>F. Int.</th>
<th>C. Int.</th>
<th>AT</th>
<th>NFC</th>
<th>LA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fluid Intelligence</td>
<td>0.128</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crystallized Intelligence</td>
<td><strong>0.318</strong></td>
<td>0.189</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Abstract Thinking</td>
<td>-0.067</td>
<td>-0.091</td>
<td><strong>-0.252</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Need For Cognition</td>
<td>0.235</td>
<td>0.037</td>
<td>0.248</td>
<td>0.154</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Language Aptitude</td>
<td>0.161</td>
<td><strong>0.430</strong></td>
<td>0.039</td>
<td>-0.039</td>
<td><strong>0.251</strong></td>
<td></td>
</tr>
<tr>
<td>Broad Retrieval</td>
<td><strong>0.321</strong></td>
<td><strong>0.299</strong></td>
<td><strong>0.345</strong></td>
<td>-0.225</td>
<td>0.224</td>
<td><strong>0.261</strong></td>
</tr>
</tbody>
</table>

Bolded values represent significant correlations: $p = .012^a$, $p = .012^b$, $p < .001^c$, $p = .019^d$, $p = .050^e$, $p = .006^f$, $p = .050^g$, $p = .042^h$.

Table 3.6 Correlations among participant individual differences and human ratings of metaphor and sarcasms.

<table>
<thead>
<tr>
<th></th>
<th>Metaphor</th>
<th>Sarcasm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Conceptual</td>
<td>Novelty/Mirth</td>
</tr>
<tr>
<td>Abstract Thinking</td>
<td>-0.070</td>
<td>-0.005</td>
</tr>
<tr>
<td>Need for Cognition</td>
<td>-0.017</td>
<td>-0.031</td>
</tr>
<tr>
<td>Fluid Intelligence</td>
<td>0.032</td>
<td>0.044</td>
</tr>
<tr>
<td>Crystallized Intelligence</td>
<td>0.091</td>
<td><strong>0.158</strong></td>
</tr>
<tr>
<td>Broad Retrieval</td>
<td>0.041</td>
<td>0.097</td>
</tr>
<tr>
<td>Working Memory Capacity</td>
<td>-0.050</td>
<td>0.018</td>
</tr>
<tr>
<td>Language Aptitude</td>
<td>0.025</td>
<td>-0.025</td>
</tr>
<tr>
<td>Age</td>
<td>-0.006</td>
<td>0.009</td>
</tr>
<tr>
<td>Socio Economic Status</td>
<td>-0.041</td>
<td>0.006</td>
</tr>
<tr>
<td>English Age of Onset</td>
<td>-0.057</td>
<td>-0.064</td>
</tr>
<tr>
<td>Language Distance</td>
<td>-0.080</td>
<td><strong>-0.100</strong></td>
</tr>
<tr>
<td>Production Times</td>
<td>0.050</td>
<td><strong>0.290</strong></td>
</tr>
</tbody>
</table>

Bolded values represent significant correlations: $p < .001^a$, $p = .002^b$, $p = .005^c$, $p = .006^d$, $p = .006^e$

3.11.3 Linguistic features and production measures.

Table 3.7 displays Pearson’s correlations among the linguistic features and human ratings for metaphors and sarcasms. Three linguistic variables correlated significantly with either the conceptual or novelty/mirth ratings of metaphors while two linguistic variables correlated significantly with the novelty/mirth ratings of sarcasm. None of these correlations exceeded a small effect size.

Table 3.7 Correlations among linguistic features and human ratings of metaphors and sarcasms.

<table>
<thead>
<tr>
<th></th>
<th>Metaphors</th>
<th>Sarcasm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

…
3.12 Linear Mixed Effects Models

Analyses of the data for metaphors and sarcasm before fitting the models revealed the presence of extreme positive outliers for the production time variable for both metaphors and sarcasm (i.e., length of time writing metaphors or sarcasm), which was corrected through logarithmic transformation. Additionally, all of the individual differences and linguistic features predictor variables were further analyzed for multicollinearity using their variance inflation factor (VIF), which is a measure of the strength of linear relationships among independent variables. A higher VIF value indicates a strong relation among predictors, leading to inflated $p$ values and therefore more difficulty in locating true significant effects. A conservative VIF threshold of 2.5 was used in order to account for even moderate collinear relations among the predictors, which can still affect the output of linear statistical models (Zuur, Ieno, & Elphick, 2010). Results from the VIF analysis reported Language Distance had a VIF greater than 2.5, which was a result of a significant, positive correlation with English age of onset ($r = .791$), and was thus removed from the analysis. Additionally, the categorical variable of NES/NNES were collinear with English age of onset, in that all NES had an English age of onset of zero. Therefore, the NES/NNES categorical variable was also removed from further analysis. None of the remaining individual differences measures reported VIF values higher than 2.5 and were...

<table>
<thead>
<tr>
<th></th>
<th>Conceptual</th>
<th>Novelty/Mirth</th>
<th>Incongruity</th>
<th>Novelty/Mirth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source Similarity (word2vec)</td>
<td>-0.053</td>
<td>-0.024</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>MRC Familiarity</td>
<td>0.029</td>
<td>-0.004</td>
<td>0.060</td>
<td>-0.009</td>
</tr>
<tr>
<td>MRC Imageability</td>
<td>0.082</td>
<td>-0.029</td>
<td>0.019</td>
<td>-0.014</td>
</tr>
<tr>
<td>MRC Meaningfulness</td>
<td>-0.007</td>
<td>0.016</td>
<td>-0.017</td>
<td>0.040</td>
</tr>
<tr>
<td>Age of Acquisition</td>
<td><strong>-0.151</strong>a</td>
<td>-0.005</td>
<td>0.032</td>
<td><strong>0.102</strong>d</td>
</tr>
<tr>
<td>Byrsbaert Concreteness</td>
<td>0.076</td>
<td>0.006</td>
<td>-0.001</td>
<td><strong>0.133</strong>a</td>
</tr>
<tr>
<td>COCA Spoken Frequency</td>
<td>0.004</td>
<td><strong>-0.272</strong>a</td>
<td>-0.007</td>
<td>-0.029</td>
</tr>
<tr>
<td>Semantic Diversity</td>
<td><strong>-0.117</strong>a</td>
<td>-0.015</td>
<td>0.023</td>
<td>-0.064</td>
</tr>
</tbody>
</table>

Bolded values represent significant correlations: $p < .001^a$, $p = .007^b$
retained in models for metaphors and sarcasms. In regards to the linguistic features predictors, the features for the metaphors also all reported VIF values < 2.5, and thus all linguistic features were retained in the models testing human ratings of metaphor. For the sarcastic responses, MRC Imageability reported a VIF of 2.95 and was thus removed from models testing human ratings of sarcastic responses. Because a total of eight different LME models are reported, a brief discussion occurs after each LME model, which is then followed by brief summaries and a general discussion.

3.13 Individual Differences in Metaphor Production

3.13.1 Results: Metaphor conceptual distance ratings.

An LME model with the conceptual distance scores as the dependent variable and participant individual differences (age, sex, age of English onset, Abstract Thinking, Need for Cognition, Fluid Intelligence, Crystallized Intelligence, Broad Retrieval, Working Memory Capacity, Language Aptitude), trial order, production time, and metaphor type (novel. vs. conventional) as predictor variables included one significant main effect and two significant interactions. First, there was a significant, positive main effect for metaphor production time, in that participants who spent more time creating their metaphor received higher human ratings of conceptual distance. Second, there was a significant interaction between metaphor type and Working Memory Capacity and a significant interaction between metaphor type and Crystallized Intelligence. These interactions suggested that higher levels of Working Memory Capacity resulted in higher conceptual distance ratings for novel metaphors compared to conventional metaphors, and that higher ratings of Crystallized Intelligence (i.e., performance on the world knowledge test) resulted in lower ratings of conceptual distance for novel metaphors when compared to conventional metaphors. These interactions are plotted in Figures 3.4 and 3.5. This
model reported a marginal $R^2$ of .035 and a conditional $R^2$ of .208. Table 3.8 displays the coefficients, standard error, $t$ and $p$ values for this full model.

**Table 3.8 Linear mixed effects model predicting human ratings of conceptual distance scores of participant metaphor production.**

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>SE</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>4.637</td>
<td>0.172</td>
<td>27.018</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Age</td>
<td>0.013</td>
<td>0.030</td>
<td>0.417</td>
<td>0.678</td>
</tr>
<tr>
<td>Sex: Male</td>
<td>0.020</td>
<td>0.065</td>
<td>0.301</td>
<td>0.765</td>
</tr>
<tr>
<td>Socioeconomic Status</td>
<td>-0.024</td>
<td>0.051</td>
<td>-0.469</td>
<td>0.642</td>
</tr>
<tr>
<td>Age of English Onset</td>
<td>-0.032</td>
<td>0.034</td>
<td>-0.964</td>
<td>0.340</td>
</tr>
<tr>
<td>Metaphor Type: Novel</td>
<td>0.061</td>
<td>0.279</td>
<td>0.219</td>
<td>0.828</td>
</tr>
<tr>
<td>Abstract Thinking</td>
<td>-0.035</td>
<td>0.030</td>
<td>-1.166</td>
<td>0.250</td>
</tr>
<tr>
<td>Need for Cognition</td>
<td>-0.021</td>
<td>0.031</td>
<td>-0.669</td>
<td>0.507</td>
</tr>
<tr>
<td>Fluid Intelligence</td>
<td>0.006</td>
<td>0.029</td>
<td>0.193</td>
<td>0.848</td>
</tr>
<tr>
<td>Crystallized Intelligence</td>
<td>0.105</td>
<td>0.032</td>
<td>3.257</td>
<td>0.002</td>
</tr>
<tr>
<td>Broad Retrieval</td>
<td>0.028</td>
<td>0.030</td>
<td>0.933</td>
<td>0.357</td>
</tr>
<tr>
<td>Working Memory Capacity</td>
<td>-0.103</td>
<td>0.032</td>
<td>-3.247</td>
<td>0.002</td>
</tr>
<tr>
<td>Language Aptitude</td>
<td>0.014</td>
<td>0.031</td>
<td>0.447</td>
<td>0.657</td>
</tr>
<tr>
<td>Production Time*</td>
<td>0.073</td>
<td>0.025</td>
<td>2.873</td>
<td>0.004</td>
</tr>
<tr>
<td>Trial Order</td>
<td>-0.002</td>
<td>0.004</td>
<td>-0.549</td>
<td>0.583</td>
</tr>
</tbody>
</table>

**Significant Interactions**

| Metaphor Type: Novel | WMC* | 0.185 | 0.081 | 2.280 | 0.023 |
| Metaphor Type: Novel | Crystallized Intelligence* | -0.239 | 0.080 | -2.996 | 0.003 |

* = Significant effect. SE = Standard Error. Baselines for categorical variables are Sex = Female, Metaphor Type = Conventional. All numerical predictors were z-scored in order to center and scale them before being entered into the model.

The primary finding from this model is that two cognitive measures of individual differences interacted with the two metaphor types (novel vs. conventional) in different ways.

First, Working Memory Capacity interacted significantly with metaphor type, in that higher Working Memory Capacity scores resulted in significantly higher conceptual distance scores for novel metaphors when compared to conventional metaphors, and that lower amounts of Working Memory Capacity resulted in the opposite effect, with lower Working Memory Capacity resulted in higher conceptual scores for conventional metaphors when compared to novel metaphors (see Figure 3.4). The interaction in Figure 3.4 suggests that producing novel metaphors with words
and concepts more distantly related to the information in the metaphor prompt required more executive control when compared to conventional metaphors. This makes intuitive sense, as the construction of conventional metaphors naturally allows for participants to draw upon pre-existing knowledge of clichéd and commonly used metaphorical comparisons, whereas asking participants to develop a novel metaphor requires them to first recall a time from their past experience and then describe that experience metaphorically.

![Figure 3.4 Interaction between metaphor type and WMC (conceptual distance)](image)

*Figure 3.4 Interaction between metaphor type and WMC (conceptual distance)*

Effect of interaction between metaphor type (novel vs. conventional) and Working Memory Capacity on conceptual distance ratings. Higher levels of Working Memory Capacity resulted in significantly higher conceptual distance for novel metaphors when compared to conventional metaphors.

Participants who generated metaphors with greater conceptual distance are potentially cycling through a greater number of concepts which are more distantly related, requiring a greater amount of executive control and inhibition of irrelevant information. Moreover, these findings are similar to previous studies that identified Working Memory Capacity as a significant predictor of metaphor production quality, in that higher levels of Working Memory Capacity
resulted in better ability to both produce and explain metaphors (Chiappe & Chiappe, 2007; Pierce & Chiappe, 2008), but differ in that the current study controlled for metaphor type (i.e., conventional vs. novel). As such, the current findings suggest that Working Memory Capacity may be an important component of metaphor production quality (when defined as conceptual distance), but only for metaphors that require participants to construe an original event in a metaphorical manner.

Crystallized Intelligence also significantly interacted with metaphor type when predicting conceptual distance scores, but in the opposite manner as WMC. In other words, higher amounts of Crystallized Intelligence (i.e., greater performance on the world knowledge test) resulted in significantly higher conceptual distance scores for conventional metaphors when compared to novel metaphors (see Figure 3.5).

![Figure 3.5 Interaction between metaphor type and Crystallized Intelligence (conceptual distance)](image)

*Figure 3.5 Interaction between metaphor type and Crystallized Intelligence (conceptual distance)*

Effect of interaction between metaphor type (novel vs. conventional) and Crystallized Intelligence on conceptual distance ratings of metaphors Higher amounts of Crystallized Intelligence (i.e., better performance on the world knowledge test) resulted in significantly higher conceptual distance scores for conventional metaphors when compared to novel metaphors.
Again, these findings make intuitive sense. The measure of Crystallized Intelligence used in this study best represents the static, declarative style of knowledge that is important to one’s culture (Flanagan, 2013). Accordingly, possessing a greater breadth and depth of general knowledge naturally contains exposure to more instances of metaphors encountered in previous discourse, media as well as a greater repository of concepts in general. At the same time, when participants are asked to develop novel metaphors, they are less able to rely on their pre-existing knowledge as compared to when they are asked to develop conventional metaphors.

Finally, metaphor production time was also a significant, positive predictor of conceptual distance scores. This finding suggests that participants who spent longer creating their metaphor were better able to produce metaphors that were more conceptually distant from the metaphor prompt.

### 3.13.2 Results: Metaphor novelty/mirth scores.

A linear mixed effects model with the averaged metaphor novelty and mirth scores as the dependent variable and the same fixed effects predictors used in the previous model included two significant main effects and three significant interactions. First, males produced metaphors that were rated as more original and mirthful when compared to females. Second, Language Aptitude was a significant negative predictor of novelty/mirth scores, suggesting that participants who scored higher on the Language Aptitude test created metaphors that were rated as significantly less novel/mirthful. In terms of the interactions, there was a significant interaction between metaphor type and Need For Cognition, such that higher levels of Need for Cognition resulted in significantly higher novelty/mirth ratings for novel metaphors when compared to conventional metaphors (see Figure 3.6). There was also a significant interaction between Broad Retrieval
Scores and Production Times, suggesting that higher Broad Retrieval scores resulted in significantly higher novelty/mirth ratings as Production Time also increased (see Figure 3.7).

Finally, there was a significant interaction between English age of onset and Crystallized Intelligence, suggesting that participants with a lower English Age of Onset and higher Crystallized Intelligence created metaphors that were rated as significantly higher for novelty/mirth when compared to participants with a higher English Age of Onset. There were no other significant main effects or interactions. This model reported a marginal $R^2$ of .180 and a conditional $R^2$ of .318 Table 3.9 displays the coefficients, standard error, $t$ and $p$ values this model.

Table 3.9 Linear mixed effects model predicting human ratings of novelty/mirth scores of participant metaphor production

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>SE</th>
<th>$t$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>3.074</td>
<td>0.105</td>
<td>29.211</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Age</td>
<td>0.040</td>
<td>0.052</td>
<td>0.756</td>
<td>0.452</td>
</tr>
<tr>
<td>Sex: Male*</td>
<td>0.341</td>
<td>0.122</td>
<td>2.790</td>
<td>0.008</td>
</tr>
<tr>
<td>Age of English Onset</td>
<td>-0.051</td>
<td>0.061</td>
<td>-0.833</td>
<td>0.409</td>
</tr>
<tr>
<td>Metaphor Type: Novel</td>
<td>0.078</td>
<td>0.254</td>
<td>0.306</td>
<td>0.762</td>
</tr>
<tr>
<td>Abstract Thinking</td>
<td>0.040</td>
<td>0.057</td>
<td>0.709</td>
<td>0.482</td>
</tr>
<tr>
<td>Need for Cognition</td>
<td>-0.073</td>
<td>0.060</td>
<td>-1.230</td>
<td>0.225</td>
</tr>
<tr>
<td>Fluid Intelligence</td>
<td>-0.009</td>
<td>0.055</td>
<td>-0.172</td>
<td>0.864</td>
</tr>
<tr>
<td>Crystallized Intelligence</td>
<td>0.201</td>
<td>0.059</td>
<td>3.428</td>
<td>0.001</td>
</tr>
<tr>
<td>Broad Retrieval</td>
<td>0.172</td>
<td>0.057</td>
<td>3.007</td>
<td>0.004</td>
</tr>
<tr>
<td>Working Memory Capacity</td>
<td>-0.096</td>
<td>0.054</td>
<td>-1.762</td>
<td>0.084</td>
</tr>
<tr>
<td>Language Aptitude*</td>
<td>-0.133</td>
<td>0.058</td>
<td>-2.303</td>
<td>0.026</td>
</tr>
<tr>
<td>Production Time*</td>
<td>0.409</td>
<td>0.033</td>
<td>12.506</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Trial Order</td>
<td>0.008</td>
<td>0.005</td>
<td>1.647</td>
<td>0.100</td>
</tr>
</tbody>
</table>

**Significant Interactions**

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>SE</th>
<th>$t$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metaphor Type: Novel</td>
<td>NFC*</td>
<td>0.247</td>
<td>0.097</td>
<td>2.538</td>
</tr>
<tr>
<td>Broad Retrieval</td>
<td>Production Time*</td>
<td>0.092</td>
<td>0.031</td>
<td>2.918</td>
</tr>
<tr>
<td>Age of English Onset</td>
<td>Crystallized Intelligence*</td>
<td>-0.126</td>
<td>0.055</td>
<td>-2.311</td>
</tr>
</tbody>
</table>

* = Significant effect. SE = Standard Error. Baselines for categorical variables are Sex = Female, Metaphor Type = Conventional. All numerical predictors were z-scored in order to center and scale them before being entered into the model.
A greater number of main effects and interactions were reported for metaphor novelty/mirth scores when compared to the conceptual distance model. In regards to the main effects, Language Aptitude was a significant, negative predictor of metaphor novelty/mirth ratings for both novel and conventional metaphors. This finding suggests that participants better at explicit pattern recognition specific to languages and grammatical sensitivity produced less novel and less mirthful metaphors. Reasons for this are not immediately clear, because Language Aptitude is similar to Fluid Intelligence (i.e., problem solving), which has been shown previously to be a significant positive predictor of metaphor production ability (Silvia & Beaty, 2012). However, the measure of Fluid Intelligence (i.e., the letter sets task) used in this study relied on inductive reasoning, whereas the Language Analysis Test was more related to explicit, deductive reasoning, because participants were provided with a set of language rules to analyze, decode, and extrapolate during the test. Therefore, it may be that deductive reasoning (as measured through language analytic ability) is not as useful when attempting to produce creative metaphors. However, more research is needed in order to fully understand this effect.

Another significant influence reported in the model was that that males produced metaphors that were judged to be significantly more novel and mirthful when compared to females. This may be further evidence of previously reported gender differences between males and females in regards to figurative and humorous language use. For instance, one study investigating casual conversation among college students reported that men used sarcasm significantly more than women (Gibbs & Colston, 2012). Moreover, both genders are thought to use figurative and humorous language for different communicative goals, with women focusing on social group maintenance and men focusing on creating positive self-images (Martin, 2007). It may be that in this data, males focused on creating metaphors that were clever or humorous in
order to further their positive self-image, and thus resulted in significantly higher novelty/mirth scores on metaphors when compared to females. This finding is similar to research investigating responses to incomprehensible humor, where men were more invested in overtly displaying their ability to understand humor (even when they did not understand it), whereas females were less invested in demonstrating their ability to understand humor and significantly more likely to directly admit their inability to understand a humorous message (Bell, 2013). However, more research, such as post-hoc interviews or think aloud protocols are needed in order to fully understand the motivations behind metaphor creation.

There were also three significant interactions reported. First, there was a significant interaction between Need for Cognition and metaphor type (see Figure 3.6). This interaction indicated that higher amounts of Need for Cognition contributed to significantly higher novelty/mirth scores for novel metaphors when compared to conventional metaphors, suggesting that higher levels of Need for Cognition aided in the ability to produce original and mirthful novel metaphors.
Figure 3.6 Interaction between metaphor type and Need for Cognition (metaphor novelty/mirth)
Effect of interaction between metaphor type and Need for Cognition on conceptual distance ratings. Increasing levels of Need for Cognition led to increased novelty/mirth ratings for novel metaphors when compared to conventional metaphors.

This finding suggests that participants who were more willing to engage in a cognitively complex task produced metaphors that were perceived to be more novel and mirthful when producing a novel vs. a conventional metaphor. This finding is similar to the above result demonstrating that Working Memory Capacity aided in producing novel metaphors with higher conceptual scores compared to conventional metaphors. Because producing novel metaphors requires more complex cognition than a conventional metaphor, it follows that participants who were more pre-disposed towards taking on cognitively complex tasks would produce more creative metaphors, as they would expend more cognitive effort and concentration during the task.

Additionally, there was a significant interaction between Broad Retrieval and Production Times (see Figure 3.7). This interaction demonstrates that while participants who spent longer creating metaphors generally produced metaphors that were rated higher for novelty and mirth,
this effect was facilitated by additional increases in Broad Retrieval ability (i.e., performance on word generation tasks).

Figure 3.7 Interaction between Broad Retrieval and Production Time (metaphor novelty/mirth)
Effect of interaction between Broad Retrieval and Production Time on novelty/mirth ratings for metaphors. As Production Time increased, participants with higher Broad Retrieval scores produced metaphors with significantly higher novelty/mirth scores compared to those with lower Broad Retrieval scores.

Broad Retrieval represents the ability to access concepts and lexical items, and this finding suggests that production time may have partially been a result of the need amongst participants to locate additional concepts and lexical items with which to create their metaphor. Participants with greater Broad Retrieval ability would be better able to locate more concepts and lexical items with greater time, and having access to a wider range of concepts and lexical items naturally allows for more options when producing a creative metaphor, which in turn allows for metaphors that are either more novel, mirthful, or both.

Finally, there was a significant interaction between English age of onset and Crystallized Intelligence. This interaction indicated that participants with a higher English age of onset
benefitted less from higher levels of Crystallized Intelligence when compared to participants with a lower English age of onset (see Figure 3.8).

![Graph showing interaction between English age of onset and Crystallized Intelligence](image)

**Figure 3.8 Interaction between English age of onset and Crystallized Intelligence (metaphor novelty/mirth)**

Effect of interaction between English Age of Onset and Crystallized Intelligence on novelty/mirth ratings for metaphors. Participants with higher Crystallized Intelligence and a lower English Age of Onset produced metaphors with significantly higher novelty/mirth scores compared to those with higher English Age of Onset.

English age of onset is a strong predictor of one’s ability to attain native-like performance in English (DeKeyser, 2000, 2012), but cannot be taken as a relative measure of one’s English proficiency. In this data, participants with a lower English age of onset created metaphors with lower novelty/mirth scores if they also had lower performance on the Crystallized Intelligence test, which suggests these participants benefitted more from greater pre-existing knowledge, which includes one’s cultural knowledge (Flanagan, 2013). Alternatively, participants with the highest English age of onset did not benefit from greater performance on the Crystallized Intelligence test to the same degree as those with lower English age of onset. Thus, it may be that lower English age of onset in this data goes hand-in-hand with cultural knowledge related to
English speaking communities, and may have captured a dichotomy between the native and non-native English speakers, as it was collinear with the NES/NNES categorical comparison mentioned earlier. In other words, a higher English age of onset in this data may index relative familiarity with American culture through an English medium. Accordingly, participants in this data with a lower English age of onset but higher Crystallized Intelligence may not have the same level of access to pre-existing, cultural knowledge specific to the English speaking communities that is a result of lived experience in the United States, and therefore resulted in less access to knowledge that may result in a more creative metaphor. Therefore, this interaction may demonstrate a divide between generalized knowledge (as measured via Crystallized Intelligence) and specific, cultural knowledge (as measured via English age of onset).

3.13.3 Summary of individual differences in metaphor production.

These results add to previous similar research investigating individual differences and metaphor use. Working Memory Capacity was previously identified as a significant predictor of metaphor comprehension and production ability, and the results here mirror those findings (Chiappe & Chiappe, 2007; Pierce & Chiappe, 2008), with the new caveat that Working Memory Capacity may only apply to aspects of conceptual distance in novel metaphors, and may actually decrease the ability to produce creative conventional metaphors. Moreover, other measures of intelligence (i.e., Crystallized Intelligence and Broad Retrieval) were also significant predictors of metaphor production scores. While these results at first seem to diverge from previous reports that Fluid Intelligence is an important component of metaphor production quality (Beaty & Silvia, 2013; Silvia & Beaty, 2012), those previous studies did not test for Crystallized Intelligence, although Beaty and Silvia (2013) recommended investigating the role of Crystallized Intelligence in metaphor production. Moreover, Working Memory Capacity and
measures of Fluid Intelligence are strongly related, and in that light the results here align with Silvia and Beaty (2013), who found that Fluid Intelligence was a significant predictor only for novel metaphors (and not conventional metaphors), similar to the current results which found that Working Memory Capacity was only a significant predictor of conceptual distance scores for novel metaphors.

Furthermore, much like previous studies of metaphor production ability, time spent producing the metaphors was a significant, positive predictor, with participants’ metaphors more conceptually distant than the prompts when spending more time creating the metaphor. The significant interaction between Broad Retrieval and production time for the novelty/mirth scores further suggests that spending longer on metaphor production is at least partially a function of searching for concepts and lexical items with which to create the metaphors.

These results also identify other variables that influence metaphor production ability that have not been previously reported. The finding that Language Aptitude is a significant negative predictor of metaphor novelty/mirth may tap into specific, explicit language-related analytic ability that serves to inhibit creative language ability, although more research is needed. Moreover, English age of onset interacted with Crystallized Intelligence in a manner that suggests Crystallized Intelligence may have also captured one’s pre-existing cultural knowledge. The additional finding that Need for Cognition is also a predictor of creative novel metaphor production ability demonstrates that participants’ willingness to engage in certain tasks is another important consideration when examining metaphor production ability. While Need for Cognition may not predict naturalistic metaphor production ability, modeling participants’ willingness to complete tasks is at least an important variable to control for during laboratory experiments.
3.14 Individual Differences in Sarcasm Production

3.14.1 Results: Sarcasm incongruity ratings.

An LME model with sarcasm incongruity scores as the dependent variable and participant individual differences (age, sex, age of English onset, Abstract Thinking, Need for Cognition, Fluid Intelligence, Crystallized Intelligence, Broad Retrieval, Working Memory Capacity, Language Aptitude), sarcasm prompt type (three-pane comic, desert island, or black and white), trial order, and production time as fixed effects included four significant interactions, as well as a significant main effect for trial order. First, there was a significant interaction between English age of onset and Broad Retrieval, in that higher Broad Retrieval scores resulted in significantly lower incongruity scores for participants with a higher English age of onset when compared to those with lower English age of onset (see Figure 3.9). There was another significant interaction between production time and Crystallized Intelligence, such that higher Crystallized Intelligence scores resulted in significantly higher incongruity scores as production time also increased (see Figure 3.10). Additionally, there was a significant interaction between English age of onset and sarcasm prompt, and a significant interaction between Crystallized Intelligence and sarcasm prompt. Follow up comparisons among the three levels of the sarcasm prompt type revealed that higher amounts of Crystallized Intelligence resulted in significantly lower ratings of incongruity for sarcastic responses produced for the three-panel comic prompts when compared to both the black and white ($p < .001$) as well as the desert island ($p = .022$) single-panel comics (see Figure 3.11). In regards to the interaction between English age of onset and sarcasm prompt type, follow up multiple comparisons revealed that a higher English age of onset resulted in significantly lower incongruity ratings for the three-panel comics when compared to the black and white single-panel comics ($p = 002$; see Figure 3.12). The significant
main effect for trial order suggested participants produced sarcastic responses that were rated as less incongruous as they proceeded through the experiment. There were no other significant main effects or interactions. This model reported a marginal $R^2$ of .094 and a conditional $R^2$ of .248.

Table 3.10 displays the coefficients, standard error, $t$ and $p$ values this model.

**Table 3.10 Linear mixed effect model predicting human ratings of incongruity of participant sarcastic responses**

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>SE</th>
<th>$t$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>4.693</td>
<td>0.111</td>
<td>42.234</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Age</td>
<td>-0.081</td>
<td>0.053</td>
<td>-1.531</td>
<td>0.132</td>
</tr>
<tr>
<td>Sex: Male</td>
<td>0.105</td>
<td>0.119</td>
<td>0.880</td>
<td>0.383</td>
</tr>
<tr>
<td>Age of English Onset</td>
<td>0.047</td>
<td>0.071</td>
<td>0.652</td>
<td>0.516</td>
</tr>
<tr>
<td>Abstract Thinking</td>
<td>-0.047</td>
<td>0.053</td>
<td>-0.896</td>
<td>0.375</td>
</tr>
<tr>
<td>Need for Cognition</td>
<td>0.085</td>
<td>0.054</td>
<td>1.553</td>
<td>0.127</td>
</tr>
<tr>
<td>Fluid Intelligence</td>
<td>0.049</td>
<td>0.055</td>
<td>0.902</td>
<td>0.371</td>
</tr>
<tr>
<td>Working Memory Capacity</td>
<td>-0.018</td>
<td>0.054</td>
<td>-0.343</td>
<td>0.733</td>
</tr>
<tr>
<td>Crystallized Intelligence</td>
<td>0.100</td>
<td>0.070</td>
<td>1.428</td>
<td>0.156</td>
</tr>
<tr>
<td>Production Time</td>
<td>-0.055</td>
<td>0.034</td>
<td>-1.618</td>
<td>0.106</td>
</tr>
<tr>
<td>Sarcasm Prompt: Comic</td>
<td>-0.203</td>
<td>0.118</td>
<td>-1.712</td>
<td>0.121</td>
</tr>
<tr>
<td>Sarcasm Prompt: Desert Island</td>
<td>-0.039</td>
<td>0.119</td>
<td>-0.331</td>
<td>0.748</td>
</tr>
<tr>
<td>Language Aptitude</td>
<td>0.026</td>
<td>0.058</td>
<td>0.451</td>
<td>0.654</td>
</tr>
<tr>
<td>Broad Retrieval</td>
<td>-0.094</td>
<td>0.059</td>
<td>-1.595</td>
<td>0.117</td>
</tr>
<tr>
<td>Trial Order*</td>
<td>-0.023</td>
<td>0.008</td>
<td>-2.896</td>
<td>0.004</td>
</tr>
</tbody>
</table>

**Significant Interactions**

| Crystallized Intelligence | Production Time* | 0.085 | 0.034 | 2.481 | 0.013 |
| Crystallized Intelligence | Sarcasm Prompt: Comic* | -0.305 | 0.071 | -4.286 | < .001 |
| Crystallized Intelligence | Sarcasm Prompt: D.I. | -0.139 | 0.072 | -1.926 | 0.055 |
| Age of English Onset       | Broad Retrieval*   | -0.149 | 0.066 | -2.266 | 0.028 |
| Age of English Onset       | Sarcasm Type: Comic* | -0.224 | 0.072 | -3.130 | 0.002 |
| Age of English Onset       | Sarcasm Type: Desert Island | -0.135 | 0.072 | -1.887 | 0.060 |

* = Significant effect. SE = Standard Error, D.I. = Desert Island. Baselines for categorical variables are Sex = Female, Sarcasm Prompt = Black and White. All numerical predictors were $z$-scored in order to center and scale them before being entered into the model.

As a whole, four significant interactions comprise the significant predictors for incongruity ratings of participant sarcasm responses. First, there was a significant interaction between Broad Retrieval and English age of onset (see Figure 3.9). The interaction in Figure 3.9 indicates that participants with a higher English age of onset as well as higher Broad Retrieval
scores produced sarcastic responses rated as significantly lower for incongruity when compared to participants with a lower English age of onset.

*Figure 3.9 Interaction between English age of onset and Broad Retrieval (sarcasm incongruity)*

Effect of interaction between English age of onset and Broad Retrieval scores on sarcasm incongruity ratings. Increasing levels of Broad Retrieval resulted in significantly lower incongruity ratings for sarcastic responses produced by participants with a higher English age of onset when compared to those with lower English age of onset.

Therefore, although participants with a higher English age of onset could achieve high Broad Retrieval scores during the word generation tasks, the increased access did not help produce sarcastic replies that were highly incongruous. This finding is difficult to interpret, but one reason may be that greater access to *English* lexical items caused these participants to create more contextually appropriate replies for their sarcastic answers which were also grammatically and semantically correct, lowering perceptions of incongruity among the raters as the answers were less incongruous within the context of the prompt. Again, if English age of onset was indexing relative experience with culture through an English medium, it may be that increased access to lexical items alone was not enough if the participants did not possess the requisite pragmatic knowledge to employ these items in an incongruous manner.
In addition, there was a significant interaction between Crystallized Intelligence and sarcasm production time, in that if production time increased, participants with higher Crystallized Intelligence scores produced sarcastic responses with significantly higher incongruity scores compared to those with lower Crystallized Intelligence scores (see Figure 3.10). In other words, much like the significant interaction between production time and Broad Retrieval for the novelty/mirth scores of metaphor, participants who spent longer producing their sarcastic responses may have done so for the purpose of searching through stored knowledge for concepts and information that would aid in the production of a sarcastic response.

Figure 3.10 Interaction between production time and Crystallized Intelligence (sarcasm incongruity)
Effect of interaction between production time and Crystallized Intelligence on sarcasm incongruity ratings. As production times increased, higher amounts of Crystallized Intelligence resulted in significantly higher incongruity ratings.

In this case, Crystallized Intelligence may also contain a participants’ understanding of sarcasm and previous encounters with sarcastic responses, and by providing a more prototypical sarcastic response, participants may thus have been rated higher for incongruity, a key
component of sarcasm. There were also two interactions between the sarcasm prompt type and participant individual differences. First, there was a significant interaction between Crystallized Intelligence and prompt type, in that higher levels of Crystallized Intelligence resulted in significantly lower ratings of incongruity for the three-panel comics when compared to the single-panel black and white and single-panel desert island comics (see Figure 3.11).

![Figure 3.11 Interaction between prompt type and Crystallized Intelligence (sarcasm incongruity)](image)

Effect of interaction between sarcasm prompt type and Crystallized Intelligence on sarcasm incongruity ratings. Higher amounts of Crystallized Intelligence resulted in significantly lower incongruity ratings for the three-panel comic prompts when compared to both the single panel black and white and single panel desert island prompts.

As stated previously, Crystallized Intelligence is a measure of one’s static, declarative knowledge, and this interaction suggests that such knowledge did not aid in producing sarcastic responses that were incongruent with the situation presented in the prompt when responding to the three-panel comics when compared to the other two prompt types. When considering the prompts themselves, each prompt type included differing levels of contextual information, with the three-panel comics providing the most contextual information of all the prompts in that they
established a prior situation for which to make a sarcastic comment on (see Figure 3.3 and Appendix H). Conversely, both the black and white and desert island comics provided relatively less contextual information prior to the sarcastic utterance, with the desert island comics providing almost no contextual information (see Figure 3.12 below for an example). Therefore, depending on the prompt given, participants may have a higher need to rely on their background information (i.e., Crystallized Intelligence) in order to intuit what may have led to the situation that prompted a sarcastic response.

Figure 3.12 Example of desert island comic used as sarcastic stimuli

Accordingly, higher amounts of Crystallized Intelligence would be more useful when responding to prompts with lower amount of contextual information provided (i.e., desert island and black and white comics), and be less useful with prompts including the highest amount of contextual information (i.e., three-panel comics), which is reflected by the interaction plotted in Figure 3.11.
The other interaction with prompt type was with English age of onset, in that a higher English age of onset resulted in lower ratings of incongruity for the three-panel comics when compared to the single-panel black and white comics (see Figure 3.13).

![Figure 3.13 Interaction between prompt type and English age of onset (sarcasm incongruity)](image)

Effect of interaction between sarcasm prompt type and English age of onset on sarcasm incongruity ratings. Higher English age of onset resulted in significantly lower incongruity ratings for the three-panel comic prompts when compared to the single panel black and white prompts.

This interaction suggests that at lower levels of English age of onset, which most likely captured the native English speakers in this data, there was little difference in incongruity ratings among the three prompt types. However, as English age of onset rose, incongruity ratings for the three-panel comic prompts became significantly lower when compared to the black and white comic prompts. Much like the previous interaction with Crystallized Intelligence, this interaction may have to do with the amount of contextual information provided in the prompts. Intuitively, it follows that all participants would benefit more from prompts containing the most contextual information available (i.e., the three-panel comics), but these findings suggest otherwise, and
instead indicate that the black and white comics created an ideal prompt for participants with a higher English age of onset to create incongruous sarcastic responses with. This finding is difficult to interpret, but it may be that the black and white comics provided just enough contextual and pragmatic information for participants with a higher English age of onset, whereas the desert island comics had too little and the three panel comics had too much. For example, the comic depicted in Figure 3.14 provides a relatively straightforward scene, in which it is made clear that one person was inconvenienced through the actions of another.

![Figure 3.14 Example black and white comic used as a sarcasm prompt](image)

In comparison, the three-panel comics all involved situations that are more pragmatically complex, such as a classmate not sharing the burden on an assignment, or a braggart being beaten during a sporting competition, and therefore participants with higher English age of onsets may not possess the pragmatic knowledge necessary in order to fully unpack the pragmatic implications behind the actions in the comics, as they may be specific to American cultural experiences. However, more research is needed in order to fully understand this finding.
3.14.2 Results: Sarcasm novelty/mirth ratings.

An LME model with sarcasm novelty/mirth scores as the dependent variable and the same participant individual differences used in the previous model as predictor variables reported five significant, positive main effects. First, production time was a significant, positive predictor, suggesting that participants who spent longer writing their sarcastic responses produced sarcastic responses which were rated higher for novelty/mirth. Second, Broad Retrieval was a significant, positive predictor, suggesting participants with higher Broad Retrieval scores (i.e., generated more words during the three word generation tasks) produced sarcastic responses with higher novelty/mirth ratings. Third, there was a significant difference among the three sarcasm prompt types, with sarcastic responses produced for the single-panel desert island comics receiving significantly higher novelty/mirth scores when compared to sarcastic responses produced after three-panel comic prompts. Fourth, each increase in participants’ age led to significantly higher perceptions of novelty/mirth. Finally, trial order was a significant, positive predictor, in that ratings of novelty/mirth increased as trial order increased, suggesting that participants produced more novel and mirthful examples as the experiment proceeded. There were no other significant main effects or interactions. This model reported a marginal $R^2$ of .151 and a conditional $R^2$ of .270. Table 3.11 displays the coefficients, standard error, t and p values for this model.

Table 3.11 Linear mixed effect model predicting human ratings of novelty/mirth of participant sarcastic responses

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>SE</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>2.810</td>
<td>0.123</td>
<td>22.823</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Age*</td>
<td>0.127</td>
<td>0.058</td>
<td>2.175</td>
<td>0.034</td>
</tr>
<tr>
<td>Sex: Male</td>
<td>0.117</td>
<td>0.132</td>
<td>0.885</td>
<td>0.381</td>
</tr>
<tr>
<td>Age of English Onset</td>
<td>0.095</td>
<td>0.063</td>
<td>1.503</td>
<td>0.139</td>
</tr>
<tr>
<td>Abstract Thinking</td>
<td>0.084</td>
<td>0.058</td>
<td>1.437</td>
<td>0.157</td>
</tr>
<tr>
<td>Need for Cognition</td>
<td>-0.027</td>
<td>0.060</td>
<td>-0.448</td>
<td>0.656</td>
</tr>
<tr>
<td>Fluid Intelligence</td>
<td>-0.045</td>
<td>0.060</td>
<td>-0.746</td>
<td>0.459</td>
</tr>
<tr>
<td>Working Memory Capacity</td>
<td>0.072</td>
<td>0.059</td>
<td>1.218</td>
<td>0.229</td>
</tr>
<tr>
<td>Crystallized Intelligence</td>
<td>0.106</td>
<td>0.063</td>
<td>1.679</td>
<td>0.099</td>
</tr>
</tbody>
</table>
Results for the sarcasm novelty/mirth ratings suggested that features related to participants and the prompts all influenced perceptions of novelty/mirth. First, higher amounts of Broad Retrieval resulted in significantly higher ratings of novelty/mirth. Much like the ratings of novelty/mirth for the metaphor responses, the Broad Retrieval findings suggest that increased access to concepts and lexical items provided access to a greater range of options with which to create an example of sarcasm. Second, longer time spent creating a sarcastic response also resulted in significantly higher ratings of novelty/mirth. Because no other variables interacted with production time, this effect may simply represent the notion that participants who spent longer to processing the sarcastic stimuli or thinking about their answer were able to generate more creative sarcastic responses. In addition, participants that were older produced sarcastic responses that were deemed to be more novel and mirthful, which may suggest that the natural experience that comes with being older allows for better ability to be creative when producing sarcastic responses. Finally, there was a significant effect of prompt, with answers produced after the desert island comics receiving significantly higher novelty/mirth ratings when compared to the three-panel comic prompts. When considering the prompts, the desert island comics allowed for the largest amount of creative license, as they placed almost no constraint on the sarcastic reply that could be produced. Alternatively, the three-panel comics placed a large amount of constraint on the participants, as they were instructed to produce answers that aligned with the
definition of sarcasm, allowing for relatively little creative freedom. Therefore, answers in the desert island comics could vary to a greater degree when compared to the three-panel comics, and thus may have affected the raters’ perceptions of novelty simply because of the wider range of answers that were provided.

3.14.3 Summary of individual differences in sarcasm production.

Compared to the results for metaphor production, not as many predictor variables were able to account for human ratings of incongruity or novelty/mirth in the sarcastic responses produced by the participants. However, some similarities do exist, in that Crystallized Intelligence and Broad Retrieval were both positive and significant predictors of either the sarcasm incongruity or novelty/mirth scores. These results suggest greater access to pre-existing, static knowledge, concepts, and lexical items was associated with the ability to produce sarcastic responses with greater incongruity between the sarcastic statement and the situation depicted in the stimuli, whereas greater access to concepts and lexical items through Broad Retrieval helped with both incongruity and novelty/mirth scores. However, because of the interaction between Broad Retrieval and English age of onset, these findings suggest that differences in sarcasm production may be tied to the ability to access concepts and words in English and pragmatically apply them in a sarcastic manner, which may be more difficult for those with a higher English age of onset as these were the non-native English speakers that may have less experience in an American cultural context.

Additionally, the different sarcastic prompts interacted significantly with several variables, suggesting that the amount of contextual information provided in the prompt is an important element when examining sarcasm. This should not be surprising, as much of the research into verbal irony and sarcasm has always placed a strong emphasis on how context
shapes verbal irony interpretation (Attardo, 2000; Colston & Gibbs, 2007; Gibbs, 1986; Gibbs & Colston, 2012), and these results provide evidence that context is taken into consideration when producing sarcastic utterances.

3.15 Linguistic Features and Metaphor Production

3.15.1 Results: Metaphor conceptual distance ratings.

An LME model with metaphor conceptual distance as the dependent variable and linguistic features related to lexical sophistication (MRC Familiarity, MRC Imageability, MRC Meaningfulness, Age of Acquisition, Brysbaert Concreteness, COCA Spoken Word Frequency, and Semantic Diversity) and source overlap (word2vec), along with metaphor type (conventional vs. novel) as predictor variables reported three linguistic indices as significant predictors of the conceptual distance ratings. First, Brysbaert Concreteness was a significant, positive predictor of conceptual distance scores, suggesting that higher levels of Word Concreteness resulted in significantly higher ratings of conceptual distance. Age of Acquisition and Semantic Diversity were significant negative predictors of the conceptual distance score, suggesting that words used in metaphor production that were learned at a higher age and those used in a greater number of contexts resulted in lower ratings of conceptual distance. There were no other significant main effects. This model reported a marginal $R^2$ of .041 and a conditional $R^2$ of .249. Table 3.12 reports the coefficients, standard errors, $t$ and $p$ values for this model.

Table 3.12 Linear mixed effects model predicting conceptual distance scores for metaphors using linguistic features

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>SE</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>4.559</td>
<td>0.089</td>
<td>51.179</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Metaphor Type: Novel</td>
<td>-0.228</td>
<td>0.399</td>
<td>-0.571</td>
<td>0.575</td>
</tr>
<tr>
<td>Source Similarity (word2vec)</td>
<td>0.010</td>
<td>0.032</td>
<td>0.324</td>
<td>0.746</td>
</tr>
<tr>
<td>MRC Familiarity</td>
<td>0.015</td>
<td>0.027</td>
<td>0.569</td>
<td>0.570</td>
</tr>
<tr>
<td>MRC Imageability</td>
<td>-0.011</td>
<td>0.039</td>
<td>-0.277</td>
<td>0.782</td>
</tr>
<tr>
<td>MRC Meaningfulness</td>
<td>-0.034</td>
<td>0.034</td>
<td>-0.999</td>
<td>0.318</td>
</tr>
<tr>
<td>Age of Acquisition*</td>
<td>-0.123</td>
<td>0.035</td>
<td>-3.533</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Brysbaert Concreteness*</td>
<td>0.102</td>
<td>0.039</td>
<td>2.610</td>
<td>0.009</td>
</tr>
<tr>
<td>-------------------------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>COCA Spoken Word Frequency</td>
<td>0.027</td>
<td>0.031</td>
<td>0.877</td>
<td>0.380</td>
</tr>
<tr>
<td>Semantic Diversity*</td>
<td>-0.106</td>
<td>0.035</td>
<td>-2.993</td>
<td>0.003</td>
</tr>
</tbody>
</table>

* = Significant predictor. SE = Standard Error. Baseline for Metaphor Type = Conventional. All predictors were centered and scaled before being entered into the model. All indices represent measures of content words only.

Three linguistic features related to lexical sophistication were significant predictors of the metaphor conceptual distance ratings. First, metaphors containing words with higher average Age of Acquisition (AoA) scores received significantly lower conceptual distance ratings. Words with a higher AoA are those that are self-reported to be learned later in life based on human judgments, and therefore represent less frequent and more sophisticated vocabulary. This suggests that more sophisticated language in terms of AoA scores was not necessary in order to construct metaphors with higher conceptual distance between the entities being described in the metaphors.

For example, the following metaphor had an average AoA of 8.9 and a conceptual distance score of one: Some professors are geniuses like a supercomputer. The prompt for this metaphor was Some professors are very smart. The word genius has an AoA of 7.21 and the word supercomputer has an AoA of 12.44, and these two words contributed significantly to the higher AoA score. Moreover, the word genius is conceptually similar to the prompt (i.e., very smart), and does not allow for any alternative conceptual interpretations. Indeed, genius is essentially a synonym of smart, and thus represents the same concept, and the inclusion of supercomputer also contains concepts related to intelligence, further amplifying the notion of smartness evoked by the word genius. Conversely, the following metaphor has an average AoA of 3.5 and a conceptual distance score of five: That book is worth my arm and leg in response to the prompt Some property is very valuable. In this metaphor, the words arm, leg, and book all have AoA scores of less than four, and thus contribute to a relatively low AoA rating.
Furthermore, there is greater conceptual distance between a variety of concepts in this metaphor, with the words *arm* and *leg* perhaps conceptualized as *high value currency*, but only if one is aware of the idiomatic use of the expression *costs an arm and a leg*. Unlike the *genius* metaphor with high AoA, the words *arm* and *leg* are also not more sophisticated synonyms of any words in the prompt.

In addition to AoA, metaphors with higher Semantic Diversity scores also received significantly lower conceptual distance scores. Words with higher Semantic Diversity are words with less specific meanings that can be used a wider variety of word contexts (Hoffman et al., 2013), which means words with higher Semantic Diversity are less sophisticated and more semantically ambiguous, (Hoffman & Woollams, 2015). Therefore, it follows that metaphors containing more semantically ambiguous words may not be directly referencing specific concepts to make a metaphorical comparison with. For example, the metaphor *Some teenagers are like neutrons in an atom* in response to the prompt *some teenagers are always moving around and never stopping* received a conceptual distance score of five and a Semantic Diversity score of 1.47. Conversely, the metaphor *Certain memories can be recalled as if they were occurring right before you* in response to the prompt *certain memories are very detailed* received a conceptual distance score of three and a Semantic Diversity score of 2.144. The first metaphor contains very specific and constrained words (i.e., neutrons), whereas the second metaphor contains very ambiguous and frequent words (e.g., *they, right, you*) and does not index specific concepts, but rather evokes a scene for the reader to imagine themselves in.

In a similar fashion, metaphors with higher average Word Concreteness received significantly higher conceptual distance scores. Words with higher concreteness scores are those that represent specific concepts and represent experience-based, rather than language-based
concepts (Brysbaert et al., 2014), suggesting that metaphors containing more concrete language were better able to make explicit conceptual comparisons that were also judged to be conceptually distant. For example, the metaphor *Her beauty is always an advantage* in response to the prompt *Her beauty gets her things* had an average word concreteness of 2.175 and a conceptual distance score of one, whereas the *Her smile pulls people towards her like bees to honey* in response to the prompt *Her smile attracts others* had an average word concreteness of 4.6 and a conceptual distance score of 4.5. In the first metaphor, the word *advantage* is less concrete than the words used in the second metaphor (e.g., *bees, honey*) and evokes less specific conceptual information, as the word *advantage* does not encode any specific advantage, but rather the overall idea or concept of an advantage.

These findings suggest that the human raters’ perceptions of conceptual distance in the metaphors were influenced by the use of specific words in the metaphors. This may be because metaphors with more specific word usage were better able to evoke conceptual comparisons that were more distantly related, making it easier for the raters to identify the size of the conceptual comparison in the metaphor. Conversely, metaphors with higher AoA scores may have tended to use synonyms with the same overall semantic meaning, leading to lowered perceptions of conceptual distance among the human raters. That being said, this model explained a total of 4.1% of the variance in conceptual distance scores, suggesting that these linguistic features only account for a relatively small amount of the variation in conceptual distance scores and that they did not play a strong role in the human raters’ conceptual rating decisions. Indeed, there are counter-examples of metaphors with high Semantic Diversity also receiving high conceptual distance scores, and so on. Moreover, many of the participants reproduced words from the prompt to start their metaphors, and these words would also influence the linguistic
measurements of the participant metaphors in a way that prompts with lower Semantic Diversity would lead to metaphors with lower Semantic Diversity, and so on.

3.15.2 Results: Metaphor novelty/mirth scores.

An LME model with the averaged metaphor novelty/mirth score of human ratings the dependent variable and the same linguistic features related to lexical sophistication and source overlap used in the previous model as predictor variables reported three linguistic indices as significant predictors of the novelty/mirth ratings. First, MRC Imageability was a significant, negative predictor of the novelty/mirth ratings, suggesting that metaphors including more imageable words resulted in lower ratings of novelty/mirth. Second, word2vec source similarity was also a significant, negative predictor of novelty/mirth, suggesting that metaphors containing higher semantic overlap with the source text received lower ratings of novelty/mirth. Third, COCA Spoken word frequency was also a significant, negative predictor of novelty/mirth ratings, suggesting that metaphors containing words with higher spoken word frequency resulted in significantly lower ratings of novelty/mirth. There were no other significant main effects or interactions. This model reported an $R^2$ of .075 and a conditional $R^2$ of .294. Table 3.13 reports the coefficients, standard errors, $t$ and $p$ values for this model.

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>SE</th>
<th>$t$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>3.292</td>
<td>0.101</td>
<td>32.604</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Metaphor Type: Novel</td>
<td>0.165</td>
<td>0.388</td>
<td>0.425</td>
<td>0.676</td>
</tr>
<tr>
<td>Source Similarity (word2vec)*</td>
<td>-0.127</td>
<td>0.041</td>
<td>-3.127</td>
<td>0.002</td>
</tr>
<tr>
<td>MRC Familiarity</td>
<td>0.064</td>
<td>0.035</td>
<td>1.830</td>
<td>0.068</td>
</tr>
<tr>
<td>MRC Imageability*</td>
<td>-0.106</td>
<td>0.050</td>
<td>-2.120</td>
<td>0.034</td>
</tr>
<tr>
<td>MRC Meaningfulness</td>
<td>0.003</td>
<td>0.043</td>
<td>0.064</td>
<td>0.949</td>
</tr>
<tr>
<td>Age of Acquisition</td>
<td>-0.065</td>
<td>0.045</td>
<td>-1.451</td>
<td>0.147</td>
</tr>
<tr>
<td>Brysbaert Concreteness</td>
<td>-0.067</td>
<td>0.050</td>
<td>-1.347</td>
<td>0.178</td>
</tr>
<tr>
<td>COCA Spoken Word Frequency*</td>
<td>-0.314</td>
<td>0.041</td>
<td>-7.660</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Semantic Diversity</td>
<td>-0.040</td>
<td>0.045</td>
<td>-0.895</td>
<td>0.371</td>
</tr>
</tbody>
</table>
There were three significant negative predictors of the metaphor novelty/mirth scores, suggesting that metaphors received higher novelty/mirth ratings if they included more sophisticated language and also included less semantic overlap with the metaphor prompt. Specifically, higher levels of both Spoken Word Frequency and Word Imageability resulted in significantly lower ratings of novelty/mirth for metaphors. Both of these features are components of lexical sophistication, and the direction of their influence on the novelty/mirth ratings indicates that more lexically sophisticated metaphors received higher novelty/mirth scores. For example, the metaphor *Friendship binds us like superglue* had an average word frequency of 4.16 and a novelty/mirth score of four, whereas the metaphor *Her family is like a house* had an average word frequency of 4516.5 and a novelty/mirth score of two. *House* is a much more frequent word than *superglue*, and the comparison between friendship and superglue is intuitively more clever than a comparison between a family and a house. Therefore, metaphors with higher amounts of lexical sophistication most likely incorporated higher amounts of interesting or novel lexical items, leading to increased perceptions of novelty, mirth, or both.

In addition to lexical sophistication, metaphors that contained greater semantic overlap with the metaphor prompt (as measured through word2vec) received significantly lower novelty/mirth scores. This finding makes intuitive sense because metaphors that were more closely related to the metaphor prompt were most likely those that were more cliché or did not make more distant comparisons. On the other hand, the word2vec measure may also capture the extent to which participants relied on the language from the metaphor prompt. For example, the metaphor *Some relationships are like working in a research lab and having a project fail*
received a novelty/mirth score of five and a semantic overlap score of -0.17. The only words repeated in this metaphor from the prompt are some relationships, while the rest of the metaphor includes words outside of the prompt as well as words less semantically related to the prompt.

Conversely, the metaphor The earth is full of people working like bees received a novelty/mirth score two and a semantic overlap score of 0.68. Unlike the previous metaphor, this metaphor almost completely repeats the metaphor prompt word for word (i.e., the earth is full of busy people) and only includes three original words. Therefore, it is most likely that the word2vec source similarity overlap measure captured the extent to which participants relied on language from the metaphor prompt, with participants who relied less on the prompt language creating metaphors perceived by the raters as more novel, mirthful, or both. Much like the model predicting metaphor conceptual distance ratings, the linguistic features predicting the metaphor novelty/mirth scores only explained a relatively small amount of variance in rater scores (7.5%), suggesting that linguistic features were just one small influence on the human ratings of novelty and mirth.

3.16 Linguistic Features and Sarcasm Production

3.16.1 Results: Sarcasm incongruity ratings.

An LME model predicting incongruity ratings of the sarcastic responses using linguistic features (MRC Familiarity, MRC Meaningfulness, Age of Acquisition, Brysbaert Concreteness, COCA Spoken Word Frequency, and Semantic Diversity) reported that MRC Meaningfulness was a significant, negative predictor of incongruity ratings, suggesting that sarcastic responses with more average associations to other words resulted in lower ratings of incongruity. There were no other significant main effects or interactions. This model reported a marginal $R^2$ of .020.
and a conditional $R^2$ of .198. Table 3.14 displays the coefficients, standard error, $t$ and $p$ values for this model.

**Table 3.14 Linear mixed effects model predicting incongruity scores for sarcastic responses using linguistic features**

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>SE</th>
<th>$t$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>4.396</td>
<td>0.099</td>
<td>44.278</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Sarcasm Prompt: Black and White</td>
<td>0.216</td>
<td>0.129</td>
<td>1.676</td>
<td>0.128</td>
</tr>
<tr>
<td>Sarcasm Prompt: Desert Island</td>
<td>0.175</td>
<td>0.129</td>
<td>1.352</td>
<td>0.209</td>
</tr>
<tr>
<td>MRC Familiarity</td>
<td>0.063</td>
<td>0.035</td>
<td>1.806</td>
<td>0.071</td>
</tr>
<tr>
<td>MRC Meaningfulness*</td>
<td>-0.067</td>
<td>0.032</td>
<td>-2.079</td>
<td>0.038</td>
</tr>
<tr>
<td>Age of Acquisition</td>
<td>0.034</td>
<td>0.030</td>
<td>1.130</td>
<td>0.259</td>
</tr>
<tr>
<td>Brysbaert Concreteness</td>
<td>0.027</td>
<td>0.034</td>
<td>0.780</td>
<td>0.436</td>
</tr>
<tr>
<td>COCA Spoken Frequency</td>
<td>-0.003</td>
<td>0.033</td>
<td>-0.103</td>
<td>0.918</td>
</tr>
<tr>
<td>Semantic Diversity</td>
<td>-0.026</td>
<td>0.036</td>
<td>-0.729</td>
<td>0.466</td>
</tr>
</tbody>
</table>

* Significant effect. SE = Standard error. Baseline for Sarcasm Prompt = Three Panel Comic. All numerical predictors were z-scored before being entered into the model.

There was only one significant linguistic predictor of the sarcasm incongruity ratings. The results demonstrated that sarcastic responses with higher average Word Meaningfulness (i.e., greater number of word associations) received significantly lower incongruity ratings. When examining the sarcastic responses, it is not immediately clear why a greater number of word associations would result in lower perceptions of incongruity. For example, the following sarcastic reply *Oh, thank you! I appreciate your effort* in response to a comic of a man splashing another man with a puddle while driving and explaining that he had tried to avoid the puddle received an incongruity score of 4.5 and had an average Word Meaningfulness score of 318.5. Conversely, the sarcastic reply *I hate traveling* in response to a comic of a young man being promised exciting adventures in the military but only to end up peeling potatoes received an incongruity score of two and had an average Word Meaningfulness score of 517. One potential difference is that the first example uses more conversational language when compared to the second example, and thus contains words with fewer other word associations. Regardless, this
model only accounted for 2% of the variance in incongruity scores, suggesting that this linguistic feature played a small role in raters’ perceptions of incongruity.

Indeed, in terms of incongruity, raters were tasked with judging whether the speaker’s statements clashed in some way with the surrounding context provided in the sarcasm comic prompt. For the first example, the man who has been recently splashed by a puddle proceeds to thank the driver who splashed him, and the statement was deemed to be highly incongruous, as human experience would dictate that an apology is little relief from being splashed by a puddle from a passing car. In the second example, the young man’s comment was deemed to be less incongruous than the first example, receiving an average score of two. This most likely was the result of the raters being less certain about the young man’s true feelings. While the comic set up a context where the young man desired to travel the world, the reality was the travelling equated to either travelling and peeling potatoes, or not travelling and peeling potatoes (see Figure 3.15). Either way, it was entirely plausible that the young man had changed his mind and legitimately hated travelling within the context of the military, and thus his statement could have been entirely congruent with the context, resulting in a lower incongruity score.

Figure 3.15 Sarcastic comic stimuli related to travelling the world in the military
3.16.2 Results: Sarcasm novelty/mirth ratings.

An LME model predicting novelty/mirth ratings of the sarcastic responses using the same linguistic features as the previous model included one significant main effect and two significant interactions. First, Age of Acquisition was a significant, positive predictor of the novelty/mirth scores, suggesting that sarcastic responses including words that were acquired later in life led to higher ratings of novelty/mirth. Second, there was a significant interaction between MRC Familiarity and sarcasm prompt type and a significant interaction between Brysbaert Concreteness and sarcasm prompt type. Multiple comparisons for the sarcasm prompt types revealed significant differences between the desert island prompt and three-panel comic prompt for both of these features. There were no other significant main effects or interactions. This model reported a marginal $R^2$ of .068 and a conditional $R^2$ of .268. Table 3.15 displays the coefficients, standard error, $t$ and $p$ values for this model.

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>SE</th>
<th>$t$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>2.965</td>
<td>0.125</td>
<td>23.661</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>MRC Familiarity</td>
<td>0.007</td>
<td>0.044</td>
<td>0.154</td>
<td>0.877</td>
</tr>
<tr>
<td>Sarcasm Prompt: Black and White</td>
<td>0.190</td>
<td>0.162</td>
<td>1.175</td>
<td>0.272</td>
</tr>
<tr>
<td>Sarcasm Prompt: Desert Island</td>
<td>0.377</td>
<td>0.162</td>
<td>2.328</td>
<td>0.046</td>
</tr>
<tr>
<td>Age of Acquisition*</td>
<td>0.114</td>
<td>0.038</td>
<td>3.013</td>
<td>0.003</td>
</tr>
<tr>
<td>Brysbaert Concreteness</td>
<td>0.038</td>
<td>0.063</td>
<td>0.607</td>
<td>0.544</td>
</tr>
<tr>
<td>COCA Spoken Frequency</td>
<td>0.030</td>
<td>0.039</td>
<td>0.763</td>
<td>0.446</td>
</tr>
<tr>
<td>Semantic Diversity</td>
<td>-0.065</td>
<td>0.043</td>
<td>-1.507</td>
<td>0.132</td>
</tr>
<tr>
<td>MRC Meaningfulness</td>
<td>0.001</td>
<td>0.039</td>
<td>0.018</td>
<td>0.985</td>
</tr>
</tbody>
</table>

**Significant Interactions**

- MRC Familiarity: Sarcasm Prompt: Black and White | 0.196 | 0.120 | 1.638 | 0.102
- MRC Familiarity: Sarcasm Prompt: Desert Island* | 0.428 | 0.128 | 3.351 | 0.001
- Concreteness: Sarcasm Prompt: Black and White | 0.057 | 0.088 | 0.652 | 0.515
- Concreteness: Sarcasm Prompt: Desert Island* | 0.210 | 0.085 | 2.472 | 0.014

* Significant effect. SE = Standard error. Baseline for Sarcasm Prompt = Three Panel Comic. All numerical predictors were z-scored before being entered into the model.
Sarcastic responses containing greater levels of average AoA received significantly higher novelty/mirth ratings. This finding provides some evidence suggesting that sarcastic responses which are more lexically sophisticated are perceived as more creative, because higher amounts of AoA tend to suggest higher levels of lexical sophistication. For example, the sarcastic reply of *at least we have water* for one of the desert island comics received a novelty/mirth score of 2.25 and had an average AoA score of 3.04, whereas the sarcastic reply *you have surgical precision behind the wheel* in response to the puddle splash comic received a novelty/mirth score of 4.75 and had an average AoA of 7.45. The second example’s use of *surgical precision* represents less frequent words when compared to the first example, which in turn provides a higher likelihood that the author of the second sarcastic response coined an answer that was unique when compared to the other participants, subsequently increasing perceptions of novelty and perhaps mirth among the human raters. Thus, the AoA results suggest that using more lexically sophisticated language could be one strategy for producing more creative sarcastic responses.

Additionally, two lexical features interacted with prompt type in a similar manner, in that there were significant differences between the desert island prompt and the three-panel comic prompt for both features. These interactions are plotted in Figure 3.16 and Figure 3.17, and demonstrate that increasing levels of MRC Familiarity and Brysbaert Concreteness significantly increased perceptions of novelty/mirth for sarcastic replies made in response to the desert island prompts when compared to the three-panel comic prompts.
**Figure 3.16 Interaction between MRC Familiarity and prompt (sarcasm novelty/mirth)**
Effect of interaction between MRC Familiarity and sarcasm prompt on raters’ perceptions of novelty/mirth for sarcastic responses. Higher levels of MRC Familiarity, resulted in significantly higher perceptions of novelty/mirth for sarcastic responses produced after desert island prompts when compared to those produced after three-panel comic prompts.

**Figure 3.17 Interaction between Brysbaert Concreteness and prompt (sarcasm novelty/mirth)**
Effect of interaction between Brysbaert Concreteness and sarcasm prompt on raters’ perceptions of novelty/mirth for sarcastic responses. Higher levels of Brysbaert Concreteness resulted in significantly higher perceptions of
novelty/mirth for sarcastic responses produced after desert island prompts when compared to those produced after three-panel comic responses.

Higher levels of both MRC Familiarity and Brysbaert Concreteness suggest less lexically sophisticated language, because words that are more familiar correlate with more frequently used words, and words that are and more concrete represent concepts that are more easily retrieved due to their encoding as both a lexical item (e.g., car) as well as the visual concept of that same item (e.g., a concept of a car). Words that are less concrete may have weaker encodings due to being more abstract (e.g., the word *upkeep* may only have a strong lexical representation and a weak visual representation), which in turn represents a more complex idea and thus more sophisticated language. As was previously reported in the individual differences analysis, the desert island comics received significantly higher novelty/mirth ratings when compared to the three-panel comics, perhaps due to the greater contextual freedom the participants had when crafting sarcastic statements for desert island prompts, which could in turn lead to greater perceptions of novelty on the part of the raters due to a wider range of possible answers. Because there was less contextual information available in the desert island prompts, it may be that sarcastic responses including less sophisticated language (i.e., more concrete concepts that are more familiar) were better able to index specific ideas indicative of sarcastic meaning for the desert island prompts when compared to the three-panel comic prompts, where contextual information could fill in semantic gaps for the raters. Much like the other models, these features accounted for a relatively small amount of variance in the raters’ scores (6.8%), again suggesting that linguistic features played a small yet significant role in raters’ perceptions of creativity among the sarcastic responses.
3.17 General Discussion

Although figurative language such as metaphor and sarcasm is widely studied in the fields of linguistics and cognitive science, the purpose behind many of these studies is to investigate participant processing and comprehension of figurative language (Gibbs & Colston, 2012). Compared to investigations of figurative language processing and comprehension, studies exploring production of figurative language are relatively few. Of those that do exist, the vast majority focus on production of metaphor, and have reported that participant individual differences such as intelligence, working memory capacity, and language background influence the quality of metaphors produced during metaphor production tasks (Blasko, 1999; Chiappe & Chiappe, 2007; Golden, 2010; J. Johnson & Rosano, 1993; Philip, 2010; Pierce & Chiappe, 2008; Silvia & Beaty, 2012). The purpose of this study was to further investigate figurative language production by simultaneously investigating a wider number of individual differences and their relation to figurative language production quality for both metaphor and sarcasm. In addition to participant individual differences, the role of lexical sophistication and semantic cohesion was also examined in order to gain a better understanding into how different linguistic properties of figurative language influence human perceptions of figurative language quality.

To do so, a total of 61 participants completed a series of metaphor and sarcasm production tasks and also undertook surveys and tests measuring a wide battery of individual differences. Human raters then rated each metaphor and sarcastic response for features related to figurative language production quality. For metaphors, the human raters assigned scores for conceptual distance, novelty, and mirth. For the sarcastic responses, the human raters assigned scores for incongruity, novelty, and mirth. The metaphors and sarcastic responses were also analyzed for linguistic features related to lexical sophistication, and the metaphors were further
analyzed for source overlap between the metaphor prompts and the participant answers. A series of linear mixed effects models were then constructed in order to determine whether any of these features were significant predictors of the human ratings of metaphor and sarcasm production quality.

Dimension reduction of the human ratings provided two separate measures of figurative language quality for both the metaphors and the sarcastic responses. For both types of responses, the ratings of novelty and mirth were highly correlated and loaded strongly into a single construct. Ratings of conceptual distance for the metaphors and ratings of incongruity for the sarcastic responses loaded into separate constructs that were not strongly associated with the novelty/mirth scores, and thus a total of four separate dependent variables were developed. Results from a series of linear mixed effects models identified several cognitive, demographic, and linguistic features that influenced human perceptions of figurative metaphor and sarcasm quality based on these different aspects of figurative language production quality.

In regards to metaphor production, these results and those from previous studies support the notion that metaphor production ability is aided through different measures of cognitive ability (e.g., Crystallized Intelligence, Working Memory Capacity). Moreover, this study adds to the current knowledge base by identifying Need for Cognition, English age of onset, and gender as additional influences on metaphor production quality. The effect sizes for the two metaphor LME models suggest that participant individual differences are much stronger predictors of variance in novelty/mirth scores (18.0%) when compared to conceptual distance scores (3.5%). Moreover, much like previous studies have reported, differences exist in metaphor production depending on if participants are asked to produce relatively creative metaphors or if they can rely on clichéd, generalized knowledge of metaphors.
Results for the sarcastic responses also identified the importance of individual differences on sarcasm production ability. Specifically, higher levels of Crystallized Intelligence, Broad Retrieval, and age led to higher incongruity or novelty/mirth ratings, and English age of onset interacted with these features to highlight the role that familiarity with American culture may have played on sarcasm production quality. In addition, these features also interacted with sarcasm prompt type, with prompts including more or less amounts of contextual information constraining the sarcastic responses in ways that led to facilitation or inhibition of these individual differences. Finally, the effect sizes of the sarcasm models mirrored the results from the metaphor models in that that the individual difference predictors explained much more variance in novelty/mirth scores (15.1%) when compared to the incongruity scores (9.4%).

Differences in the amounts of variance explained (i.e., effect sizes) for these models may relate to the underlying constructs the dependent variables represented. Because the human ratings of conceptual distance, incongruity, and novelty/mirth diverged so strongly and did not load into single components for the metaphors or the sarcastic responses, this suggests that features theoretically related to definitions of metaphors and sarcasm (i.e., conceptual distance and incongruity) are not necessarily related to other aspects of figurative language quality, such as perceived creativity of the figurative language as operationalized through measures of originality and mirth. As mentioned in the methods section, the novelty and mirth subscales were included in order to capture aspects of participants responses that further related to theoretical definitions of creativity in general, which define creativity as a solution to a problem that is both original and effective (Runco & Jaeger, 2012).

Another possible explanation for the differences in effect sizes for the models was that there was a strong ceiling effect observed for conceptual distance and incongruity ratings (i.e.,
the majority of the metaphors and sarcastic responses had very high conceptual distance or incongruity ratings). This may have been a result of the instructions provided to participants, who were informed to match their examples of metaphor or sarcasm as closely as possible to the definitions provided. Conversely, there was not a strong ceiling effect for the novelty/mirth scores for both the metaphors and sarcastic responses, and these scores also contained greater variation, with lower means and greater standard deviations when compared to the conceptual distance and incongruity scores (Table 3.4). While the novelty category was relatively straightforward in that it captured the relative originality of a participants’ answer, mirth was a catch-all category designed to capture any aspects of the participant answers that the raters found interesting, clever, or humorous. Because the novelty and mirth scores were so strongly correlated for both metaphors and sarcastic responses, this suggests that it was difficult for the human raters to disassociate perceptions of novelty and mirth, and that these two constructs are strongly associated with one another. In other words, metaphors and sarcastic responses that are humorous, clever, or interesting are also those that are original. Furthermore, participant answers with lower conceptual distance or incongruity ratings were commonly also the answers that were judged to not be examples of metaphor or sarcasm and were thus excluded from analysis. Therefore, it may be that the novelty and mirth ratings were better measures of creativity for the metaphors and sarcastic responses in this data.

Accordingly, further investigations of metaphor and sarcasm production quality should continue to carefully consider the operationalization of creativity and metaphor quality when studying figurative language production. Specifically, while previous studies of metaphor production used subjective scoring, where raters assigned a single score which attempted to incorporate remoteness, originality, and cleverness of the metaphors (Beaty & Silvia, 2013;
Silvia & Beaty, 2012), the results here suggest that combining these aspects into a single score may place too strong an emphasis on the theoretical understanding of metaphor (i.e., conceptual distance), and under emphasize human perceptions of novelty and mirth, which may themselves be better measures of creativity. Moreover, much like the relation between conceptual distance and creativity for metaphors, perceived creativity of sarcastic responses should be careful to avoid placing too strong of an emphasis on theoretical constructs of sarcasm (i.e., incongruity).

Additionally, future studies employing a similar analytic rubric assessing figurative language production quality should consider refining the manner in which some of the subscales are defined and how raters are trained, as the initial adjudication between the two raters in this study was relatively poor for all of the subscales aside from Metaphor Conceptual Distance. Specifically, measures of novelty and mirth (which were strongly correlated and collapsed into a single construct) may have been primarily capturing perceptions of humor, and research into the assessment of humor appreciation may provide some guidance in regards to revising the training in identifying novelty and mirth. For instance, previous studies investigating the assessment of perceptions of humor in verbal jokes and cartoons have reported that the appreciation of humor can be split into two factors related to appreciation of humorous *structure* and appreciation of humorous *content* (Ruch, 1992). Therefore, disagreements or differences in the raters’ mirth ratings may reflect mismatches in appreciation (or lack of appreciation) or humor structure and humor content in the participants’ metaphors and sarcastic responses. Moreover, related research has demonstrated significant differences exist in the appreciation of humor based on individual personality differences, such as one’s tolerance for ambiguity, level of political conservatism, and sensation seeking (Hofmann & Ruch, 2017; Ruch, 1992), and differences among these features of the two raters may have further contributed to differences in humor perception.
Based on this research, future analytic rubrics measuring humor could employ better defined subscales attending to humorous perceptions based on structure as well as content. Additionally, it may also be helpful to have human raters complete the standardized test of humor appreciation that the above research is derived from. This test is known as the 3WD test and measures responses to different jokes and cartoons that have been pre-selected to differ based on structure and content (Ruch, 1983). Choosing raters that have similar performance on the 3WD test could help reduce differences between raters in perceptions of mirth in figurative language.

Initial agreement between raters was also relatively low for perceptions of Sarcasm Incongruity. As explained earlier, the definition of incongruity the raters used was defined in terms of how strongly the intended meaning of an utterance clashed with the context of the prompt. The low agreement for this scale may reflect the overall difficult scholars have had with providing concrete definitions for sarcasm (Colston & Gibbs, 2007), as well as distinguishing it as a separate construct verbal irony (Colston, 2017). There are also reports that regional variation exists in folk definitions of sarcasm, with residents of the northern United States viewing sarcasm as significantly more humorous when compared to residents of the southern United States (Dress, Kreuz, Link, & Caucci, 2008), which leaves open the possibility for other individual differences to affect how raters perceive sarcasm which are currently unknown. Moreover, even though the raters were treated to identify incongruity, novelty, and mirth as separate constructs, it may have been difficult for them to truly separate these constructs during the ratings. The low agreement between the two raters was addressed in the current study through the use of a third adjudicator more familiar with the theoretical definitions associated with satire (i.e., the author), and that strategy may be necessary for future ratings studies as well.
That being said, one final opportunity to improve rater reliability is to employ large-scale rater training using frameworks such as the Verbal Irony Procedure (VIP), which has been designed as a method to identify and extract instances of verbal irony from previously existing corpora (Burgers, van Mulken, & Schellens, 2011). Although the VIP puts forth a procedure to identify whether an utterance is ironic or not, it may be possible to modify the VIP to include a likert scale instead of a binary scale which asks raters to indicate how ironic an utterance is (rather than if it is just ironic or not).

Finally, it should be noted that the interpretation of these results is not evaluative (e.g., concluding that males are better at metaphor than females), but rather intended to identify different measurable variables that may influence figurative language production. For instance, the ability to create a more novel and mirthful metaphor does not mean that someone is better at language, but rather suggests that one is better at the specific language task employed in this study.

In regards to the linguistic features analysis, variables representative of lexical sophistication played a small yet significant role in explaining variance among rater perceptions of conceptual distance, incongruity, and novelty/mirth among the metaphors and sarcastic responses. Results for the metaphor conceptual distance scores suggested that more specific words resulted in higher ratings of conceptual distance, perhaps because more specific words are better able to encode specific concepts. Results for the sarcasm incongruity scores reported a small, negative effect for Word Meaningfulness, suggesting that sarcastic replies employing words with more associations to other words were deemed less incongruous, which may be a result of the use of more or less conversational language in participants’ responses.
In terms of the novelty/mirth scores, a clearer picture emerged, with responses including higher amounts of features indicative of lexical sophistication received significantly higher novelty/mirth scores, although lexical sophistication interacted significantly with sarcasm prompt in that higher amounts of Word Concreteness and Familiarity (i.e., lower lexical sophistication) resulted in higher perception of novelty/mirth for desert island prompts when compared to the three-panel comic prompts, further highlighting the influence of contextual information in the prompt on participants’ answers. At the same time, the amount of variance explained in all of the models using linguistic indices was relatively low, ranging from 0.2% for incongruity scores among the sarcastic replies to 7.5% for the novelty/mirth scores in the metaphors.

Linguistic features explained more variance in the metaphors, which is most likely a result of the linguistic context in which metaphors operate. Specifically, the understanding of a metaphor requires the possessing of conceptual information encoded in the metaphor. However, in order to understand a sarcastic reply, one must be more aware of the surrounding social and pragmatic context. Echoing contextual information is not necessary in many sarcastic responses, as it is known knowledge already available to those within the situation. Furthermore, a simple thank you can be taken as sarcastic in the right contexts. As mentioned in Chapter 2, many computational investigations of sarcasm struggle with locating linguistic features capable of modeling this contextual incongruity, and the results here attest to the continuing difficulty of modeling sarcasm linguistically. Finally, one limitation present in this data is that the answers produced by the participants were generally short, which in turn could easily bias some of the lexical measurements used, as all of them reported average scores for all the content words in an answer.
3.18 Conclusion

This study has identified several cognitive, social, and linguistic variables which can account for significant amounts of variance in human rater’s perceptions of creativity among examples of metaphor and sarcasm. In most cases, variables related to participant individual differences explained a much greater amount of variance in raters scores when compared to linguistic variables, and several consistent features were identified as important to both metaphor and sarcasm production. In line with previous studies, this research demonstrates that cognitive measures related to intelligence, such as the ability to recall previously known knowledge, concepts, and lexical items had a generally positive effect on human ratings of participant figurative language production. Additionally, gender emerged as an important variable, which may index social knowledge and experience that goes beyond cognitive measures of intelligence. English age of onset was also an important feature, which may have served as an implicit proxy comparing native versus non-native English speakers in this data set, which in turn may reflect relative levels of experience with American culture. Finally, linguistic features related to lexical sophistication suggest that a greater vocabulary can facilitate perceptions of creativity, most likely to do the purposeful use of less frequent and more specific vocabulary.

This study adds knowledge to the field of figurative language by providing more information about features that affect figurative language production. Further research should continue to explore these and other features, and also consider the ecological validity of figurative language production in a laboratory setting. For example, while participants were given contextual information for which to make metaphorical or sarcastic statements, this may not reflect their actual, day-to-day use of figurative language. Regardless, the results here are promising in that specific cognitive, demographic, and linguistic features of metaphor and
sarcasm have been identified as exerting influence on figurative language production ability, and further serve to highlight the need to consider features beyond literal and figurative meaning when investigating figurative language.

4 INDIVIDUAL DIFFERENCES AND LINGUISTIC FEATURES OF SATIRE PROCESSING

4.1 Introduction
The purpose of this study is to investigate the roles of individual differences (e.g., demographic features, language background, and cognitive ability) and linguistic features on satirical text processing and comprehension. As mentioned in previous chapters, much of the focus in figurative language research has been on processing and comprehension of figurative language, with many studies investigating the processing of metaphorical and sarcastic utterances (Gibbs & Colston, 2012). Conversely, satire, a more subtle form of discourse employing verbal irony to impart a critical and subversive figurative meaning, is relatively understudied, especially from the perspective of psycholinguistics and cognitive linguistics (Simpson, 2003; Skalicky & Crossley, in press). Unlike other types of figurative language, satire is more difficult to isolate into single utterances which can be clearly defined as satirical language. Instead, satire relies on large discourse incongruities among background knowledge, genre knowledge, and discursive practices (Simpson, 2003). Indeed, of the few studies that do exist, relative levels of both specific and general background knowledge appear to play a role in one’s ability to both understand and appreciate a satirical message (Boukes et al., 2015; A. Johnson et al., 2010; LaMarre et al., 2009; Skalicky & Crossley, in press). Therefore, it is probable that other cognitive individual differences may also affect satire processing and comprehension.
In addition to cognitive differences, demographic features and affective perceptions may also influence satire processing. For instance, there is evidence that age influences satirical use, with two studies demonstrating that relative youth facilitates satire comprehension and appreciation (Boukes et al., 2015; Skalicky & Crossley, in press). Whether this is due to younger people having more daily exposure to satire or more familiarity with the typical genres satire is used in (e.g., satirical television news such as The Daily Show) still remains unclear, but these findings do highlight the importance of taking age into consideration when investigating satire. Additionally, satire comprehension has been linked to affective perceptions of sincerity and humor, with studies showing readers who are aware of an author’s satirical intent (i.e., lack of sincerity) more easily understand the satirical message in a text (Pfaff & Gibbs, 1997), while other studies demonstrate that perceptions of humor are strong indicators that a reader has understood and appreciated a satirical message (A. Johnson et al., 2010; Simpson, 2003; Skalicky & Crossley, in press).

Furthermore, satirical texts have been shown to differ from non-satirical texts for measures related to lexical sophistication (Skalicky & Crossley, 2015). Specifically, Skalicky and Crossley (2015) found that satirical Amazon.com product reviews contained higher levels of word concreteness when compared to non-satirical reviews, and were also marked by lower incidences of positive words and higher incidences of negative words. Therefore, because satirical texts appear to contain specific linguistic features which differ from non-satirical texts (at least in the context of Amazon.com production reviews), it may be the case that these linguistic features influence the processing and comprehension of satirical meaning. However, no research to date has investigated the role of linguistic features during satirical processing and comprehension.
It appears that only one study has investigated satire processing using psycholinguistic techniques, and that study only focused on decontextualized satirical newspaper headlines (Skalicky & Crossley, in press). Because satirical text normally requires a larger context to impart a satirical meaning when compared to other examples of figurative language (Simpson, 2003), it follows that any investigation of satirical processing and comprehension should examine the phenomenon from a larger, text level. Accordingly, generalized theories of text processing, such as situation models (van den Broek & Helder, 2017; Zwaan, 1999; Zwaan & Radvansky, 1998), may provide a useful baseline with which to examine satire. Situation models posit that during the reading process, readers construct a mental representation of a text. Situation models are continuously updated as a text is read through a combination of passive reading strategies and more deliberate, reader-initiated comprehension strategies (van den Broek & Helder, 2017). While passive strategies allow readers to effortlessly comprehend a text with little deliberate reconciliation necessary, semantic incongruities, lack of background knowledge, and other incongruities in text structure or meaning require a switch to deliberate reader-initiated processes, where a reader searches for additional inferential clues in order to reconcile the meaning of a text (van den Broek & Helder, 2017). Because satire is inherently reliant on large, inferential meaning based on recognition of an ironic incongruity (Simpson, 2003), it follows that satirical texts should require higher levels of reader-initiated processes, which may manifest themselves through more difficulty during text processing.

Therefore, there is a need to further investigate satire processing and comprehension in order to better understand the roles that individual differences, affective perceptions, and linguistic features have on satirical processing and comprehension. Moreover, it is important to do so using longer texts that contain more contextual clues related to the satirical message, which
may provide stronger ecological validity into the reading process. By doing so, a better understanding of satire can be obtained, which also contributes to the study of figurative language as a whole.

4.2 Research Questions

The following research questions guide Study 2:

1. Are there significant differences in affective perceptions between satirical and non-satirical texts?

2. Do individual differences, affective perceptions, and measures of lexical sophistication and semantic cohesion influence the processing of satirical texts?

3. Do individual differences, affective perceptions, measures of lexical sophistication and semantic cohesion, and processing time interact with satirical meaning comprehension?

4.3 Method

The purpose of this study is to investigate the influence of individual differences and linguistic features on the processing and comprehension of satirical texts. To do so, human participants were recruited to read and respond to a series of satirical and non-satirical texts presented to them on a computer screen. Participants also completed tests and surveys designed to measure a wide amount of individual differences (e.g., background knowledge, working memory capacity). The texts were also analyzed for linguistic features related to lexical sophistication (complexity of language) and semantic cohesion. These measures were all used in subsequent statistical analyses in order to investigate if individual differences or linguistic features were predictive of the processing or comprehension of satirical texts.
4.4 Participants

The same participants recruited for Study 1 also participated in Study 2 (see Chapter 3 for a full description of participants’ demographic features).

4.5 Materials and Study Design

4.5.1 Demographic survey and individual differences tests.

The same participants recruited to complete Study 1 also participated in Study 2. Therefore, all demographic and individual differences survey and test material obtained for Study 1 were also used in this study (see Chapter 3).

4.5.2 Satirical and non-satirical texts.

Ten satirical and ten non-satirical texts were chosen from the satirical newspaper The Onion and the non-satirical online publication Science Daily. This was done to control for possible genre differences that may influence the recognition of satirical and non-satirical texts during the experiment. The genre of the texts selected was identified broadly as a scientific report article, in which a short news story is published describing the results of a recent scientific study. The Onion regularly publishes satirical versions of this genre. For example, in June of 2003, the Onion published an article titled Study Finds Jack Shit, which reported that a fictional team of researchers at John Hopkins hospital studying cholesterol had found absolutely nothing (i.e., jack shit) after five years of research. While it is certainly possible to find no significant results after a five-year study, the dissemination of those results would be done in a more serious, professional, and academic manner. In comparison, all of the articles on Science Daily are short published summaries of real research that has been conducted.

An initial analysis of the scientific report articles published by The Onion demonstrated that all of its scientific report articles followed a specific and consistent structure containing
specific moves, or units of discourse that perform specific communicative functions for a specific genre (Swales, 1990, 2004). The *Onion* texts contained a five-move structure, summarized below in Table 4.1.

**Table 4.1 Move structure of Onion scientific report articles genre**

<table>
<thead>
<tr>
<th>Move</th>
<th>Function</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Headline</td>
<td>Provide summary of research findings</td>
<td>Study: More Couples Delaying Divorce Until Kids Old Enough To Remember Every Painful Detail CHICAGO—In a new study published this week in <em>The American Journal Of Sociology</em>, researchers reported that parents throughout the United States are increasingly opting to delay divorce until their children are old enough to remember each and every traumatizing detail. “What we found is that more and more couples are deliberately holding off on dissolving their unhappy marriages until their children are 9 or 10, the approximate age at which they’re cognitively capable of retaining every unbearably painful moment,” said study co-author Anna Dasgupta, adding that children at that stage of maturation will generally have the ability to recall for the rest of their lives the moment their dad told them he was moving out. “And by not rushing the announcement, parents ensure that their children have accumulated at least some memories of happier times, such as Christmases and birthday parties when the whole family was together, which they will use as sources of self-torment in the broken homes of their adolescence.” The study also noted that by postponing their divorce, parents helped ensure their children had sufficiently developed their sense of agency enough to blame themselves for everything.</td>
</tr>
<tr>
<td>Introduction</td>
<td>Describe location and/or publisher of study, repeat main findings</td>
<td></td>
</tr>
<tr>
<td>Direct Quote 1</td>
<td>Introduce main researcher and a quote about findings</td>
<td></td>
</tr>
<tr>
<td>Direct Quote 2</td>
<td>Another direct quote from main researcher elaborating on findings</td>
<td></td>
</tr>
<tr>
<td>Indirect Quote</td>
<td>Indirect quote suggesting limitations, applications, or future research</td>
<td></td>
</tr>
</tbody>
</table>

The *Science Daily* articles also contained these same moves as displayed in Table 4.1, but were typically much longer than the *Onion* reports and contained a greater number of paragraphs, quotes, and elaboration on the study. To better match satirical and non-satirical texts, the following procedure was carried out. First, approximately 20 *Onion* articles were initially identified by using keyword searches for *study* or *research* on the main *Onion* website. Because previous research into satire has suggested the political views of participants strongly influences
their comprehension of satire (Boukes et al., 2015; LaMarre et al., 2009; Lee & Kwak, 2014), articles containing any overt political content were avoided. After collecting this initial sample, searches were conducted on the Science Daily website for research articles reporting on the same general topics (e.g., environment, health) as in the Onion articles. While it was not possible to find exact topic matches in some instances, the texts were matched as closely as possible to a shared topic or theme (e.g., an Onion article talking about stress increasing during family vacations was paired with a Science Daily article discussed the effects of stress on pregnancy).

Once a matching number and type of Science Daily articles were collected, the Science Daily articles were modified to conform to the five move structure presented in Table 4.1 by deleting and rearranging paragraphs in the articles, with care taken to maintain cohesion through the article. Additionally, some Science Daily articles did not include a research location or academic journal in their introduction move, and thus fictional yet plausible names for academic journals were included in some of the articles, along with legitimate locations where the research could have been conducted. Finally, each article from both The Onion and Science Daily was separated into five distinct paragraphs, one for each move listed in Table 4.1. The complete list of Science Daily articles can be found in Appendix J.

Two sub-experiments were then created by forming two groups of 10 texts each, with each group containing five satirical and five non-satirical texts. Text topics did not repeat within each group so that no group would contain a satirical and non-satirical text discussing the same topic. Table 4.2 displays this grouping arrangement. These texts were then uploaded as stimuli into two separate online Qualtrics surveys.

<table>
<thead>
<tr>
<th>Version A</th>
<th>Version B</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Satirical</td>
<td>Non-Satirical</td>
<td>Consumer Spending</td>
</tr>
<tr>
<td>Satirical</td>
<td>Non-Satirical</td>
<td>Health Problems</td>
</tr>
</tbody>
</table>
4.5.3 Linguistic features.

Next, the texts were analyzed using TAALES 2.0 (Kyle et al., 2017) and TAACO (Crossley et al., 2016) for features related to lexical sophistication and semantic cohesion, including word concreteness, word imageability, word familiarity, age of acquisition, word frequency (based on COCA newspaper subsection), and paragraph-to-paragraph semantic overlap. These features were chosen in order to perform an initial investigation into whether or not the processing and comprehension of satirical texts is influenced by features related to lexical sophistication and semantic cohesion. Chapter 3 provides explanations for word concreteness, imageability, familiarity, frequency, and age of acquisition variables, and all of these variables are different representations of the overall lexical sophistication of a text. Paragraph-to-paragraph overlap compares each paragraph to the subsequent paragraph for repeated words based on word lemmas, with higher levels of paragraph-to-paragraph overlap indicating more cohesion between adjacent paragraphs.

Any texts within each group (i.e., Onion or Science Daily) containing outliers for these features were replaced until there were no visible outliers in the boxplots in order to reduce the variance of linguistic features within the two text groups (the texts were also controlled for total number of words). Afterwards, Welch t-tests were conducted between the Onion and Science Daily texts for each linguistic feature in order to ensure no significant differences existed.
between the two groups. When necessary, words in the *Science Daily* texts were modified using synonyms or equivalent phrases in order to reduce any significant differences between the two text groups for any particular features. This process was repeated until 10 texts representing approximately similar topics were chosen for each group, where none or very few outliers were visibly present for any of the features within the groups, and no significant differences existed between the satirical and non-satirical texts for any of the features. Table 4.3 displays descriptive statistics for the linguistic features in the texts. The final versions of the texts are displayed in Appendix J.

| Table 4.3 Mean and standard deviation for linguistic features of satirical and non-satirical texts |
|-------------------------------------------------|------------------|------------------|
| Linguistic Feature                              | Satirical | Non-Satirical |
| Number of words                                 | 199.500  | 193.100        |
| Word Familiarity                                | 573.128  | 567.435        |
| Word Concreteness                               | 372.892  | 366.320        |
| Word Imageability                               | 400.725  | 393.359        |
| Age of Acquisition                              | 6.712    | 6.932          |
| Word Frequency (COCA Newspaper)                 | 640.453  | 700.537        |
| Paragraph-to-paragraph semantic overlap         | 0.097    | 0.121          |

Welch t-tests between satirical and non-satirical texts demonstrated no significant differences for any of these features. Word familiarity, concreteness, and imageability calculated using MRC norms (Coltheart, 1981). Age of acquisition norms based on Kuperman et al. (2012).

4.5.4  Comprehension and affective perceptions questions.

Two types of questions were developed for the texts: one to test direct comprehension of the satirical or non-satirical meaning, and another set of questions designed to measure affective perceptions of the text (i.e., participant emotional reactions to the texts). The satirical comprehension question was open-ended and designed to capture satirical or non-satirical comprehension by asking *What was the author’s intended meaning?* The affective response questions asked participants to rate each text on a four-point scale for three different features: sincerity, humor, and positivity. Specifically, participants were asked to indicate whether they
Strongly Disagree, Disagree, Agree, or Strongly Disagree with the following statements: *The text was sincere*, *The text was funny*, and *The text was positive*. These questions were included after each of the satirical and non-satirical texts in the Qualtrics surveys.

### 4.5.5 Genre background survey.

A survey was developed to measure participants’ overall familiarity and exposure to news and satirical genres. The survey first asked participants to answer how many days per week they engaged with news via physical newspapers, television, social media, and directly on the Internet. The survey also queried how long they typically spend each time they are engaged with this genre. Additionally, participants were asked to rate their familiarity for satirical television shows (e.g., *The Colbert Report*) and newspapers (e.g., *The Onion*) using a four-point scale. The questions for this survey are displayed in Appendix K.

### 4.6 Procedure

#### 4.6.1 Online survey.

The same participants recruited to complete Study 1 also participated in Study 2. Therefore, the online survey results obtained from Study 1 measuring demographic features (e.g., language background) and participant individual differences (e.g., fluid intelligence, need for cognition) were used for Study 2. See Chapter 3 for a full description of this survey procedure, and Appendix A for the survey question items.

#### 4.6.2 Laboratory session.

Participants completed the satire processing experiment during the same laboratory session described in Chapter 3. After completing the world knowledge and working memory capacity tests described in Chapter 3, half of the participants completed the production experiment reported in Chapter 3 before the processing experiment reported here, while the other
half completed the processing experiment before the production experiment. Participants were free to take any necessary breaks between the different experiments. Total time for the entire laboratory session lasted between 90 minutes and two hours for all participants.

4.6.3 Satire processing experiment.

Participants were randomly and equally assigned to one of the two versions of this experiment. Once participants were ready, they were directed to a soundproof room containing a desktop computer with a keyboard and mouse, where they initiated the experiment by first reading a series of instructions on the computer screen. While the first page of instructions was displayed (see Appendix J), the researcher verbally summarized the instructions for participants by informing them that they were to read 10 different news stories from the internet. Participants were instructed to read the story from start to end, just as they would any other story on the internet. Participants were then informed that after reading the story, they would be asked to write their interpretation of the author’s intended meaning (i.e., the satirical comprehension question described above) and were instructed that one or two sentences would be sufficient. Participants were then informed that they would use the mouse to rate the news story for whether they felt the story was sincere or not, whether they felt it was humorous or not, and whether they felt it had a positive or negative mood (i.e., the affective response questions described above). The researcher then informed participants that once they had finished this process for all 10 texts, they would complete a short survey asking about their reading habits (i.e., the genre background survey described above). The researcher then answered any questions about the experiment that the participants had. Once participants indicated they were ready and had no further questions, the researcher informed them that they would immediately see their first news story on the next screen, and to only proceed once they were ready to read the first story. The researcher then left
the participant alone in the soundproof room to complete the experiment. While the informed consent explained to the participants that the purpose of the experiment was to study figurative language processing, the participants were not informed ahead of time that any of the texts were satirical.

Participants read the five satirical and five non-satirical texts in a random order. Each trial began with the text fully displayed on the screen. Texts were displayed on a 22-inch monitor using 16-point, black Lucida Sans Unicode font on a white background. The title of each text was centered and underlined, and each paragraph of the text was left justified and single-spaced. Additionally, a reminder to press the Enter key when participants were done reading was displayed on the bottom of each text. Each text fit entirely on the screen with no need to scroll down using the mouse. Figure 4.1 displays an example of a text stimuli screen for a satirical text. Participants pushed the Enter key on the keyboard when they were done reading and were then required to answer the first comprehension question, which asked them to type their interpretation of the author’s intended meaning into a text box. After providing their answer and clicking the continue button, participants then used the mouse to complete the second set of comprehension questions by rating the text for sincerity, humor, and positivity using the four-point scale described above. After providing their ratings, the next trial began. Participants proceeded this way until all ten trials were completed. After the tenth text, participants immediately began the Genre Background Knowledge survey using the same desktop computer.

Once participants completed the Genre Background Knowledge survey, the processing experiment was complete. Qualtrics recorded the total time spent on each text, providing an approximate measure of total reading time per text. Qualtrics also recorded the answers provided by the participants for the authors’ intended meaning and ratings of sincerity, humor, and
positivity. Average time for participants to read and rate all texts and also complete the Genre Background Knowledge survey was 24.51 minutes ($SD = 7.39$). Participants were given a $25 Amazon.com gift card for their participation.

4.6.4 Comprehension question ratings.

Participants’ answers to the *What was the author’s intended meaning* comprehension question were analyzed only for the satirical texts (since the purpose of this experiment is to measure satirical comprehension). Answers were coded yes/no to reflect whether or not the participants indicated comprehension of satirical meaning or intent on the part of the author. Satirical comprehension was reflected through answers that explicitly mentioned satire, humor, sarcasm, or mockery, or if they directly mentioned the implied satirical message. Non-satirical comprehension was reflected through rote repetition of the purported main point of the article, typically manifested as a repetition of the article’s headline. For example, one *Onion* article reported on a fictional study stating that 750,000 Americans die each year during their first
attempt to exercise. The body of the article included claims that 225,000 Americans die within three minutes of jogging for the first time, and that 60% of Americans who go to a gym for the first time perish after their first abdominal exercise. Thus, the article is mocking the relative poor state of health stereotypically associated with Americans, but is doing so through the use of a fictional scientific study which reports exaggerated and fictional statistics (see Appendix J).

Representative examples of comprehension responses that suggested the participants understood the satirical meaning of this article are “A veiled criticism of the public health crisis of obesity in the U.S. through the veil of humor” and “Americans are obese, and we don't try to correct it, or when we do we complain way too much and then stop, resuming our obese habits.” Both of these responses indicate the participant was able to comprehend the article’s satirical message pointed at the American health crisis and attitudes towards exercise, and do not suggest the participants believed that 750,000 Americans are truly dying each year due to exercising. Because the comprehension question was relatively open-ended (i.e., What was the author’s intended meaning), some participants only reported on the purpose of the article, wherein they simply stated “humor” or “sarcasm” as their answer. These responses were also coded as evidence of satire recognition.

Representative examples of comprehension answers that suggested the participants did not understand the satirical meaning are “750000 people each year in America die from first attempt to get back in shape” and “Those who haven't exercised in years are at a higher risk of injury or death if they try to get back into shape.” Both of these examples demonstrate no comprehension of the satirical message because they repeat the purported results of the fictional study reported in the article, that Americans are dying in droves each year due to attempts at getting back into shape. The entire data set was first coded by the researcher, and then a random
sample of 20% of the data was coded by a second researcher, with agreement between the two reaching 95%.

4.6.5 Genre background information.

Participants’ answers to the genre background knowledge survey were then analyzed in order to determine which of the survey answers would serve as appropriate predictor variables in the subsequent statistical analyses. In general, participants indicated greater familiarity with written satire (e.g., The Onion) when compared with television satire (e.g., The Colbert Report), with 49 of the 59 participants indicating they were at least somewhat familiar with written satire and 30 of the 59 participants indicating they were at least somewhat familiar with television satire. In terms of engagement with different news sources, results suggest over half of the participants never obtained news information from physical newspapers or by watching television, and almost all participants indicated they used social media or direct online sources as their news sources. Participants indicated the typical time spent engaging with various news sources was less than 45 minutes per session. This information is summarized below in Table 4.4.

Based on these results, individual levels of familiarity with written and televised satire were converted to a continuous scale ranging from one to four, with a value of one representing not familiar at all and a value of four representing very familiar. In addition, two other scores were calculated to represent participants’ weekly exposure to news online using their responses for the direct online and social media news exposure. For each participant, the number of days they indicated they read news sources directly online or through social media were multiplied by the time they typically spent engaging with that form of media. Because participants only answered in ranges (e.g., between 15 and 30 minutes), these answers were converted to the
median value for each time point, so that answers of less than 15 minutes were changed to 7.5 minutes, answers of between 15 and 30 minutes were changed to 22.5 minutes, answers of between 30 and 45 minutes were changed to 37.5 minutes, and answers between 45 and 60 minutes were changed to 52.5 minutes. Answers indicating more than 60 minutes were retained at 60 minutes. These four variables (familiarity with written satire, familiarity with televised satire, time spent with news directly online, time spent with news on social media) were then used in subsequent statistical analyses of satirical text comprehension and processing.

Table 4.4 Results from Genre Background Knowledge survey.

<table>
<thead>
<tr>
<th>Level of Familiarity</th>
<th>Written Satire</th>
<th>Television Satire</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very familiar</td>
<td>20</td>
<td>15</td>
</tr>
<tr>
<td>Somewhat familiar</td>
<td>19</td>
<td>15</td>
</tr>
<tr>
<td>Somewhat unfamiliar</td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>Not familiar at all</td>
<td>11</td>
<td>24</td>
</tr>
</tbody>
</table>

Average number of days per week spent engaging with various news media

<table>
<thead>
<tr>
<th>News Source</th>
<th>M</th>
<th>SD</th>
<th>Number of 0s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Print News</td>
<td>0.51</td>
<td>0.77</td>
<td>36</td>
</tr>
<tr>
<td>Directly Online</td>
<td>4.24</td>
<td>2.45</td>
<td>4</td>
</tr>
<tr>
<td>Social Media</td>
<td>5.07</td>
<td>2.60</td>
<td>7</td>
</tr>
<tr>
<td>Watch Television</td>
<td>1.27</td>
<td>1.80</td>
<td>32</td>
</tr>
</tbody>
</table>

Number of participants indicating time per day engaging with news sources

<table>
<thead>
<tr>
<th>Time in minutes</th>
<th>Print News</th>
<th>Directly Online</th>
<th>Social Media</th>
<th>Television</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than 15</td>
<td>10</td>
<td>25</td>
<td>18</td>
<td>3</td>
</tr>
<tr>
<td>Between 15 and 30</td>
<td>12</td>
<td>19</td>
<td>17</td>
<td>12</td>
</tr>
<tr>
<td>Between 30 and 45</td>
<td>0</td>
<td>9</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>Between 45 and 60</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>More than 60</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

*Number of 0s* represents total number of participants who indicated they did not engage with that type of news media at all.

4.7 Statistical Analysis

Before running the statistical analyses, the integrity of the data was checked along with multicollinearity and adherence to statistical assumptions. In regards to data integrity, two participants demonstrated significant inattention to the texts and affective ratings during the
experiment (either skipping texts or selecting all of the same answers for the affective ratings), and their data were discarded. Therefore, the final sample size for the data is 59 participants.

Next, the predictor variables were assessed for multicollinearity through visual examination of correlations as well as using variance inflation values (VIF; see Chapter 3 for a detailed explanation of VIF). All of the individual differences measures, affective perceptions, and genre background knowledge measures reported no strong multicollinearity (i.e., absolute $r > .7$, VIF < 2.5). However, several of the linguistic features demonstrated strong multicollinearity, which suggests they measured the same construct. Specifically, there were strong, positive correlations between Word Concreteness and Word Imageability ($r = .962$) and between Word Familiarity and Age of Acquisition ($r = .821$). After removing Word Concreteness and Familiarity, the resulting four linguistic features (Age of Acquisition, Word Frequency, Word Imageability, and Paragraph-to-Paragraph Overlap) demonstrated no significant multicollinearity and were retained for subsequent analyses.

Finally, the reading time data was visually analyzed for normality and outliers. Visual analysis of boxplots and histograms suggested the reading times for the satirical and non-satirical texts were positively skewed with the presence of extreme positive outliers ($M = 72.45$ seconds, $SD = 36.97$). Therefore, any trial with a reading time longer than 2.5 standard deviations from the mean (164.88 seconds) was removed. This removed a total of 11 trials. Then, any trials representing extremely short reading times (less than 20 seconds) were removed, resulting in the removal of three more trials. The resulting data was approximately normally distributed but still contained the presence of positive outliers, and thus was transformed using logarithmic transformation with an added constant of one. The untransformed mean for the remaining 572
trials was 69.83 seconds (SD = 28.13) for all of the texts, 70.41 seconds (SD = 28.06) for the non-satirical texts (N = 286), and 69.25 seconds (SD = 28.24) for the satirical texts (N = 286).

Several statistical analyses were then conducted in order to systematically answer the research questions of this study. First, the reading times and affective perceptions of the satirical and non-satirical texts were compared using Welch t-tests and LME\(^4\) modeling in order to answer the research questions for this study. Welch t-tests compared affective ratings between the two text types to answer the first research question (*Are there significant differences in affective perceptions between satirical and non-satirical texts?*).

An LME model then measured the influence of individual differences, affective perceptions, and linguistic features on the processing times for satirical and non-satirical texts to answer the second research question (*Do individual differences, affective perceptions, and measures of lexical sophistication and semantic cohesion influence the processing in satirical texts?*).

The fixed effects for this model included participant individual differences (i.e., Abstract Thinking, Need for Cognition, Crystallized Intelligence, Fluid Intelligence, Broad Retrieval, Working Memory Capacity, Language Aptitude, handedness, age, gender, English age of onset, genre background knowledge), affective perceptions of humor, sincerity, and positivity, and linguistic features related to lexical sophistication and cohesion (Age of Acquisition, Word Frequency, Word Imageability, and Paragraph-to-Paragraph Overlap), as well as a fixed effect of text order. Participants and items were entered as crossed random effects, with a random slope of text type entered for participants. Interactions were tested between all of the fixed effects and text type, and only significant interactions were retained in the final model.

\(^4\) See Chapter 3, Methods section for an in-depth explanation of LME models.
A second model was fit in order to determine whether participant individual differences, affective perceptions, linguistic features, and processing time were predictive of participants’ satirical meaning comprehension, providing an answer to the third research question (*Do individual differences, affective perceptions, measures of lexical sophistication and semantic cohesion, and processing time interact with satirical meaning comprehension?*). Because satirical comprehension was coded as a binary decision (i.e., yes or no), a logistic generalized linear mixed effects model was used with a logit link in order to determine the likelihood of satirical comprehension based on the predictor variables. Satirical comprehension was entered as the dependent variable using dummy coding (0 = no, 1 = yes), and participant individual differences (e.g., intelligence measures, genre familiarity), affective perceptions (humor, positivity, sincerity), linguistic features of lexical sophistication and semantic cohesion, and processing time were included as fixed effects, with crossed random effects of subjects and items. Odds ratios (i.e., a measurement of how likely an outcome is to occur, in this case the presence or absence of satirical comprehension) were calculated through exponentiation of the coefficients, providing a measure of effect size for each predictor variable in the models (Levshina, 2015).

4.8 Results

4.8.1 Affective ratings.

Table 4.5 reports the descriptive statistics and results from Welch t-tests comparing the three affective ratings for each text type. As can be seen, there were significant differences between satirical and non-satirical texts for measures of sincerity, humor, and positivity. Satirical texts were rated as significantly less sincere, significantly more humorous, and significantly less positive than the non-satirical texts. Measures of effect sizes (Cohen’s D) reported the a medium
effect size for sincerity, a large effect size for humor, and a small effect size for positivity
(Cohen, 1992).

Table 4.5 Descriptive statistics and Welch t-test results for affective ratings of sincerity, humor, and positivity

<table>
<thead>
<tr>
<th></th>
<th>Satirical</th>
<th>Non-Satirical</th>
<th>Welch t-test</th>
<th>Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>Sincerity</td>
<td>2.42</td>
<td>0.92</td>
<td>3.22</td>
<td>0.66</td>
</tr>
<tr>
<td>Humor</td>
<td>2.70</td>
<td>1.10</td>
<td>1.61</td>
<td>0.70</td>
</tr>
<tr>
<td>Positivity</td>
<td>2.23</td>
<td>0.82</td>
<td>2.46</td>
<td>0.84</td>
</tr>
</tbody>
</table>

Rating scale for affective perceptions ranged from 1-4. All p < .001.

These results demonstrate satirical and non-satirical texts differed significantly for all three affective perception measures, with satirical texts perceived to be more humorous (large effect), less sincere (medium effect), and less positive (small effect) when compared to the non-satirical texts, providing an answer to the first research question (Are there significant differences in affective perceptions between satirical and non-satirical texts?). These results suggest that significant differences exist in perceptions of satirical and non-satirical texts which align strongly with theoretical definitions of satire. The largest difference was in perceptions of humor, with satirical texts being perceived as significantly more humorous when compared to non-satirical texts. Although not a form of humor itself, recognition and appreciation of a satirical message commonly results in humor (Simpson, 2003).

At the same time, the satirical texts were also perceived to be significantly less sincere than the non-satirical texts, with a medium effect size. This makes intuitive sense, as none of the satirical news stories in this study presented the findings of actual scientific studies and were thus, at least on the surface level, not presenting sincere and accurate information to the reader. This also reflects the very nature of satire as a figurative genre, in that the intended meaning of satire is constructed from the very recognition that the surface level of the satire is not to be
taken at face value. Finally, the satirical texts were perceived to be significantly less positive than the non-satirical texts, which aligns well with some descriptions of satire as a critical and polemic style of discourse (A. Nilsen & Nilsen, 2008; Pfaff & Gibbs, 1997; Simpson, 2003), and also reflects linguistic results that demonstrated satirical writing contained significantly more negative and significantly fewer positive words when compared to non-satirical writing reviews (Skalicky & Crossley, 2015). The small effect also recognizes that some types of satire are not necessarily critical, and instead are more overt in their attempt to be humorous (Boukes et al., 2015). Taken together, the differences in perceptions of humor and sincerity suggest that, on average, participants in this study were capable of detecting differences among the two text types that are thought to be fundamental to the definition of satire.

4.8.2 Reading times.

An LME model with reading time (in seconds) as the dependent variable (using logarithmic correction with a constant of 1) for the satirical (n = 286) and non-satirical (n = 286) texts and affective perceptions, participant individual differences and demographic information, text type (satirical vs. non-satirical), linguistic features, and order of the texts as fixed effects, with participants and items entered as crossed random effects and a random slope of text type on subjects included three main effects and three significant interactions. Specifically, perceptions of humor were a significant negative predictor of reading times, suggesting that as perceptions of humor increased, reading times decreased. In addition, males read the texts significantly quicker than females. Finally, text order was significant, suggesting that participants read texts quicker as they proceeded through the experiment. There were no other significant main effects or interactions. This model reported a marginal $R^2$ of .372 and a conditional $R^2$ of .763. Table 4.6 reports the coefficients, standard error, $t$ and $p$ values for this model.
Despite locating several significant predictors of text reading times for both text types, these results suggest that neither individual differences, affective perceptions, linguistic features, nor levels of genre familiarity attenuated or amplified any differences in reading times between the two text types. Instead, the reported main effects suggest that participants read the texts
quicker as they proceeded through the experiment, that males read texts quicker than females, and that higher perceptions of humor resulted in significantly quicker reading times for all texts. Because there were no significant interactions between any of these features and text type, these significant main effects are not specific to satire or non-satire when measured in total reading time and may reflect general influences on reading times.

4.8.3 Satirical comprehension.

A generalized linear mixed effects model with satirical comprehension (yes vs. no, based on participants answer to the prompt what was the author’s intended meaning for each satirical text) as the dependent variables for the satirical (n = 286) texts and participant individual differences, affective perceptions, demographic features, genre familiarity, linguistic features, time spent writing the comprehension question answer, and reading times as fixed effects with participants and items included as crossed random effects included five significant main effects. First, English age of onset was a significant negative predictor of satirical comprehension, with the odds ratio reporting that each increase in English age of onset resulted in a 72% lower chance to produce an answer reflective of a satirical interpretation. Second, the affective perceptions of sincerity, humor, and positivity were all significant predictors. Both perceptions of sincerity and positivity were significant, negative predictors, with the odds ratios reporting that each increase in perceptions of sincerity resulted in a 53% lower chance to produce an answer reflective of a satirical interpretation, and that each increase in perceptions of positivity resulted in a 47% lower chance of producing an answer reflective of a satirical interpretation. Conversely, perceptions of humor were a significant positive predictor, with the odds ratio reporting that each increase in perceptions of humor resulted in 2.9 times higher likelihood to produce an answer reflective of a

5 Because the numerical predictors were z-scored, each one-unit increase corresponds to an increase in standard deviation as opposed to each unit on the scale.
satirical interpretation. Finally, each standardized increase in average content word frequency (COCA newspaper subsection) resulted in a 49% lower chance to produce an answer reflective of a satirical interpretation. There were no other significant main effects or interactions. This model reported a marginal $R^2$ of .591 and a conditional $R^2$ of .749. Table 4.7 reports the coefficients, standard error, $z$ and $p$ values, as well as the odds ratio and 95% confidence intervals for this model.

Table 4.7 Logistic linear mixed effects model predicting satirical comprehension using individual differences and reading times.

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>SE</th>
<th>z</th>
<th>p</th>
<th>OR</th>
<th>5% CI</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-1.773</td>
<td>0.395</td>
<td>-4.493</td>
<td>&lt;.001</td>
<td>0.170</td>
<td>0.089</td>
<td>0.325</td>
</tr>
<tr>
<td>Reading Time</td>
<td>0.068</td>
<td>0.344</td>
<td>0.198</td>
<td>0.843</td>
<td>1.070</td>
<td>0.608</td>
<td>1.886</td>
</tr>
<tr>
<td>Comprehension Time</td>
<td>0.366</td>
<td>0.277</td>
<td>1.323</td>
<td>0.186</td>
<td>1.442</td>
<td>0.915</td>
<td>2.272</td>
</tr>
<tr>
<td>Cognitive Measures</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Abstract Thinking</td>
<td>0.284</td>
<td>0.376</td>
<td>0.756</td>
<td>0.450</td>
<td>1.329</td>
<td>0.716</td>
<td>2.467</td>
</tr>
<tr>
<td>Need for Cognition</td>
<td>-0.387</td>
<td>0.361</td>
<td>-1.073</td>
<td>0.283</td>
<td>0.679</td>
<td>0.375</td>
<td>1.229</td>
</tr>
<tr>
<td>Crystallized Intelligence</td>
<td>-0.033</td>
<td>0.484</td>
<td>-0.068</td>
<td>0.946</td>
<td>0.968</td>
<td>0.437</td>
<td>2.144</td>
</tr>
<tr>
<td>Fluid Intelligence</td>
<td>0.664</td>
<td>0.375</td>
<td>1.772</td>
<td>0.076</td>
<td>1.943</td>
<td>1.049</td>
<td>3.600</td>
</tr>
<tr>
<td>Broad Retrieval</td>
<td>0.251</td>
<td>0.344</td>
<td>0.729</td>
<td>0.466</td>
<td>1.285</td>
<td>0.730</td>
<td>2.264</td>
</tr>
<tr>
<td>Working Memory Capacity</td>
<td>0.308</td>
<td>0.354</td>
<td>0.871</td>
<td>0.384</td>
<td>1.361</td>
<td>0.760</td>
<td>2.435</td>
</tr>
<tr>
<td>Language Aptitude</td>
<td>-0.074</td>
<td>0.368</td>
<td>-0.200</td>
<td>0.842</td>
<td>0.929</td>
<td>0.507</td>
<td>1.703</td>
</tr>
<tr>
<td>Demographic Measures</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>0.035</td>
<td>0.361</td>
<td>0.098</td>
<td>0.922</td>
<td>1.036</td>
<td>0.572</td>
<td>1.877</td>
</tr>
<tr>
<td>Sex: Male</td>
<td>0.247</td>
<td>0.767</td>
<td>0.322</td>
<td>0.747</td>
<td>1.281</td>
<td>0.363</td>
<td>4.521</td>
</tr>
<tr>
<td>English Age of Onset*</td>
<td>-1.268</td>
<td>0.537</td>
<td>-2.360</td>
<td>0.018</td>
<td>0.281</td>
<td>0.116</td>
<td>0.681</td>
</tr>
<tr>
<td>Affective Perceptions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Humor*</td>
<td>1.082</td>
<td>0.281</td>
<td>3.846</td>
<td>&lt;.001</td>
<td>2.949</td>
<td>1.857</td>
<td>4.684</td>
</tr>
<tr>
<td>Positivity*</td>
<td>-0.634</td>
<td>0.246</td>
<td>-2.581</td>
<td>0.010</td>
<td>0.531</td>
<td>0.354</td>
<td>0.795</td>
</tr>
<tr>
<td>Sincerity*</td>
<td>-0.748</td>
<td>0.280</td>
<td>-2.675</td>
<td>0.007</td>
<td>0.473</td>
<td>0.299</td>
<td>0.750</td>
</tr>
<tr>
<td>Genre Familiarity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>News amount (direct online)</td>
<td>0.332</td>
<td>0.355</td>
<td>0.933</td>
<td>0.351</td>
<td>1.393</td>
<td>0.777</td>
<td>2.500</td>
</tr>
<tr>
<td>News amount (social media)</td>
<td>0.004</td>
<td>0.308</td>
<td>0.013</td>
<td>0.989</td>
<td>1.004</td>
<td>0.605</td>
<td>1.666</td>
</tr>
<tr>
<td>Satire Familiarity (written)</td>
<td>0.464</td>
<td>0.408</td>
<td>1.138</td>
<td>0.255</td>
<td>1.591</td>
<td>0.813</td>
<td>3.112</td>
</tr>
<tr>
<td>Satire Familiarity (TV)</td>
<td>0.047</td>
<td>0.442</td>
<td>0.106</td>
<td>0.916</td>
<td>1.048</td>
<td>0.507</td>
<td>2.166</td>
</tr>
<tr>
<td>Linguistic Measures</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age of Acquisition</td>
<td>0.123</td>
<td>0.230</td>
<td>0.533</td>
<td>0.594</td>
<td>1.131</td>
<td>0.774</td>
<td>1.651</td>
</tr>
<tr>
<td>Word Frequency*</td>
<td>-0.672</td>
<td>0.288</td>
<td>-2.336</td>
<td>0.020</td>
<td>0.511</td>
<td>0.318</td>
<td>0.820</td>
</tr>
<tr>
<td>Paragraph CW Overlap</td>
<td>-0.373</td>
<td>0.229</td>
<td>-1.629</td>
<td>0.103</td>
<td>0.689</td>
<td>0.473</td>
<td>1.004</td>
</tr>
<tr>
<td>Word Imageability</td>
<td>-0.077</td>
<td>0.283</td>
<td>-0.273</td>
<td>0.785</td>
<td>0.926</td>
<td>0.581</td>
<td>1.474</td>
</tr>
</tbody>
</table>

* = Significant predictor. SE = Standard Error, OR = Odds Ratio. Baseline for Sex = Female. All numerical predictors were z-scored before being input into the model.
Results from this model suggest that affective perceptions are significant predictors of whether or not the participants produced an answer reflective of a satirical interpretation, further highlighting the important links between humor, authorial sincerity, and criticism typically associated with satire (Pfaff & Gibbs, 1997; Simpson, 2003; Skalicky & Crossley, in press). Moreover, these effects were in the expected directions, with increased perceptions of humor leading to a higher chance of satirical meaning comprehension, and with increased perceptions of sincerity and positivity leading to lower chances of satirical meaning comprehension. Although perceptions of humor are not completely necessary for satirical meaning comprehension (A. Johnson et al., 2010), these results suggest that satire comprehension is strongly tied to perceptions of humor, providing further empirical support for theoretical links between satire and humor (Simpson, 2003). Moreover, these results provide further support for the role that perceptions of sincerity play in satirical comprehension, which to date has only been reported in a study using interviews (Pfaff & Gibbs, 1997), and provides initial evidence that emotional sentiment (i.e., perceptions of positivity) are also important in leading to a satirical meaning.

There was also a significant effect for word frequency, where increased word frequency resulted in significantly lower chances for participants to produce an answer reflective of a satirical interpretation. This means that *Onion* texts with more frequent content words on average (as found in the newspaper subsection of COCA) were also those which were more difficult for participants to understand the satirical message. This may be due to *Onion* texts with higher frequency words from the COCA newspaper subsection more closely resembling legitimate, non-fictional news articles through their inclusion of content words more likely to appear in legitimate news articles.
For example, the *Onion* text with the lowest average word frequency (379) was an article titled *Study: Average Person’s Enjoyment Of Vacation Drops 36% For Each Additional Family Member Present* which satirized the phenomenon of large family gatherings devolving into enmity by tallying the percentage that each additional family member contributed to a lowering a vacation’s overall enjoyment level. Conversely, the article with the highest average word frequency (876) was titled *Study: More Couples Delaying Divorce Until Kids Old Enough To Remember Every Painful Detail*, satirizing the relatively high rate of divorce in the United States and elsewhere. It may be the case that the first story reported on a topic relatively infrequently covered in the news, especially when compared to divorce rates, which may be a more familiar and newsworthy topic in the United States. Therefore, *Onion* articles with more frequent content words when measured via the newspaper subsection of COCA may have better aligned to the participants’ linguistic expectations of a news genre and thus masked the satirical tone of the piece to a greater degree. While these two articles had similar numbers of satirical interpretations (13 and 10, respectively), the *Onion* article with the second-highest average word frequency also had the lowest number of satirical interpretations (four), whereas the *Onion* article with the second-lowest average word frequency (496) had the highest number of satirical interpretations (16). Therefore, there may be a general effect of word frequency on satirical comprehension in the context of these articles, but is not the only influence on the likelihood of satirical comprehension.

Finally, English age of onset was also a significant predictor of satirical comprehension, in that participants with a higher English age of onset had significantly lower odds of producing an answer reflective of a satirical interpretation. Understanding satire is complex, even for native speakers of a language, because a satirical message requires more than just linguistic knowledge,
it also requires cultural and pragmatic knowledge that is typically attained through lived experience and cannot be easily taught explicitly. Therefore, while English age of onset may be a partial proxy for overall English fluency attainment, it may be just as likely that English age of onset served as a measure of exposure to United States culture through an English medium. Accordingly, participants with a higher English age of onset may have less familiarity with American culture and therefore lack some of the cultural background knowledge required to understand some of the satirical meanings in the *Onion* articles. Because none of the other variables related to English proficiency (e.g., participants’ self ratings of English proficiency) were significant predictors of satirical comprehension, it may be more likely that this tapped into cultural understanding rather than English proficiency. Regardless, these findings thus highlight the importance of considering a participants’ language and cultural background when performing research into satire processing and comprehension, and call for future research more tightly controlling and measuring the participants’ language proficiency. Further research using different populations of participants with more tightly controlled English proficiencies is needed in order to better understand the role that English proficiency plays on satire comprehension.

None of the other variables were significant predictors of satirical comprehension, suggesting that a wide number of cognitive, linguistic, and demographic differences did not significantly influence the comprehension of the satirical messages. The model testing comprehension also reported that text reading time was not a significant predictor of satirical comprehension, suggesting that any variation in reading times for the satirical texts did not increase or decrease the chances that a participant would arrive at a satirical interpretation, which may further suggest that processing is not a strong predictor of satirical meaning comprehension.
That there was no significant effect of processing (or significant differences in processing between the satirical and non-satirical texts) may be the result of *good enough* processing, or the construction of a situation model that is coherent enough with a reader’s understanding of the world to not merit any additional reading or processing time (Gibbs & Colston, 2012; ÒBrien & Cook, 2016; van den Broek & Helder, 2017). Moreover, as has been demonstrated in prior satire research, participants can fail to activate a satirical message but still claim to understand the text because they only interpret the literal, non-satirical meaning of a text (A. Johnson et al., 2010; LaMarre et al., 2009; Pfaff & Gibbs, 1997; Simpson, 2003). Because this was a laboratory experiment, participants’ standards of coherence (i.e., the threshold for switching between passive and reader-initiated reading processes) may have been lowered, because they had little motivation to read the texts beyond the confines of completing the experiment, and thus may have constructed comprehension responses that were sufficient for the purposes of the experiment, but not necessarily satirical uptake, a phenomena previously reported in a study of satirical headlines (Skalicky & Crossley, in press). Additional research employing think aloud protocols or stimulated recall interviews would provide a better understanding into how standards of coherence influence satire processing, as would experiments that test the time between reading satire and humorous uptake.

4.9 Conclusion

This study has revealed significant differences between satirical and non-satirical texts related to affective perceptions, individual differences, and linguistic features. The satirical and non-satirical texts differed significantly when comparing participant affective ratings of humor, sincerity, and positivity. This finding lends empirical justification to theoretical descriptions of satire as a form of discourse that is strongly related to humor, relies on subtle insincerity, and
typically contains a critical or polemic tone (Simpson, 2003), and further highlights the importance of considering the role that emotional or affective responses may play in the interpretation of satire. Although satirical discourse evokes a figurative meaning (i.e., a meaning other than the surface meaning of the words) in a similar manner as other forms of figurative language, the ability to detect and comprehend a satirical message may be tied to a participants’ ability to recognize and react appropriately to a satirical tone.

These findings also suggest that the processing of satirical and non-satirical texts was not significantly different, and that the individual differences and linguistic features measured in this study did not influence one text type in a significantly different manner than the other. This suggests that measuring the processing of satire may not tap into the ability to understand satire, which is further supported by the findings demonstrating that affective perceptions, English age of onset, and word frequency significantly predicted satirical comprehension.

As studies such as this one and others (Skalicky & Crossley, in press) continue to demonstrate empirical links between satire comprehension and humorous responses, it is important to consider factors that influence humor comprehension, such as differences in attending to humorous form versus humorous content (Ruch, 1983, 1992). Although satire most likely relies heavily on content-based humor, some examples of satirical headlines exploit linguistic form to create a satirical and humorous response through the employment of extremely infrequent collocations (Skalicky, in press). Therefore, future studies of satire comprehension may benefit from examining potential differences in responses to satirical meaning based on structure versus content. Moreover, one of the more common forms of satire is in the form of political critique, and previous studies have demonstrated that one’s political ideology influences perception of political satire (LaMarre et al., 2009). Political ideology has also been shown to
interact with humor comprehension. Specifically, hearers who are politically conservative tend to appreciate humor where an incongruity is completely resolved, whereas they tend to not appreciate nonsense humor where an incongruity may still remain (Ruch & Hehl, 1986a, 1986b). Resolving the incongruity of a satirical message may be more difficult than other forms of humor, and therefore participants who are more conservative may not appreciate all forms of satire, regardless of if it includes a political target or not. Accordingly, it may be important to measure political ideologies of participants as well in order to identify another potential influence of satire comprehension.

This study adds to the larger body of figurative language research by investigating a relatively understudied form of figurative language. However, the satirical texts in this study were of a specific genre (i.e., scientific news reports), and more investigation is needed using a larger number of texts from a wider number of genres, including televised satire, as it has been made popular through television shows such as The Daily Show and The Colbert Report. Regardless, the results here support calls for a more dynamic perspective in figurative language processing (i.e., one which assumed the influence of predictors is always in flux depending on other variables, rather than representing static cause and effect relationships), because all of the significant predictors used in this study (e.g., affective perceptions) are unique to individuals and cannot be fully captured through a binary comparison of a literal and non-literal meaning (Gibbs & Colston, 2012). The role of linguistic features in satirical text also calls for further investigation, as different satirical genres may interact with linguistic features in different ways. Additionally, investigations using more fine-grained reading measures such as eye-tracking are warranted. In conclusion, these results shed further light on the nature of satirical text while also identifying further avenues of study.
5 INFLUENCE OF CREATIVE LANGUAGE ON PERCEPTIONS OF CREATIVITY

5.1 Introduction

The purpose of this study is to investigate potential influences of figurative language (e.g., verbal irony) and linguistic creativity (e.g., wordplay) on perceptions of creativity. At a broader, psychological perspective, creativity is defined as one’s capacity to produce novel and valuable ideas in an appropriate context and is closely related to cognitive measures of intelligence (Acar, Burnett, & Cabra, 2017; Nusbaum & Silvia, 2011; Plucker, Beghetto, & Dow, 2004; Runco & Jaeger, 2012; Sternberg, 2006). From a linguistic perspective, examples of figurative language and wordplay are commonly described as creative forms of language because they deviate from expected linguistic conventions through playful manipulation of meaning, form, or both (Cook, 2000; Gerrig & Gibbs, 1988; Partington, 2009; Veale, 2012b). While researchers in psychology and linguistics continue to separately investigate notions of cognitive and linguistic creativity, some initial links have been made between creative language and creative ability.

For example, research in psychology has demonstrated that producing or being exposed to humorous language can increase one’s creativity, with general agreement that this effect is due to the positive emotional enhancement of humor (Martin, 2007). Other research has identified links between sarcasm and creativity, in that exposure to sarcasm through use or recall boosts levels of abstract thinking, which in turn increases performance on creative problem solving tasks (Huang et al., 2015). Moreover, metaphors have been used as prompts in creative tasks, based on the assumption that the ability to produce higher quality metaphors is a result of greater creative ability (Beaty & Silvia, 2013; Silvia & Beaty, 2012). However, none of these studies compared whether the simple presence or absence of figurative language or linguistic creativity
influenced perceptions of creativity. Although figurative language, humorous language, and wordplay are considered to be creative instances of language use, some examples of creative language are highly conventionalized and ubiquitous in everyday conversation (Gerrig & Gibbs, 1988; Gibbs, 2000; Veale, 2012b), suggesting they may be more routine than creative. Therefore, it is unclear creative language is actually perceived to be more creative than non-creative language (e.g., figurative language vs. non-figurative language). In order to further identify potential links between psychological and linguistic notions of creativity, it is important to investigate whether creative forms of language are actually perceived to be more or less creative when compared to non-creative language. The purpose of this study is to explore links between linguistic and cognitive definitions of creativity by testing whether the presence or absence of creative language (specifically, figurative language) influences perceptions of creativity during a creative production task.

5.2 Research Question

The following research question guides Study 3:

1. What effect (if any) does the presence or absence of creative language have on perceptions of creativity?

5.3 Method

In this study, participants were recruited online and asked to complete a creative production experiment where they provided one-sentence reactions for two news stories and also provided self-ratings of creativity for their news reactions when compared to pre-constructed news reactions containing different levels of figurative language. In addition, the participant news reactions were coded for the presence or absence of linguistic creativity. Statistical analyses were then carried out in order to determine if there was a statistically significant relation
between participant ratings of creativity and differing amounts and types of figurative language in the pre-constructed news reactions as well as the presence or absence of linguistic creativity in the participant news reactions.

5.4 Participants

For this study, a total of 423 participants were recruited from Amazon Mechanical Turk (AMT), an online crowdsourcing website where anonymous workers complete online tasks for financial compensation. Because of the large number of possible participants, recruiting AMT workers is an ideal method for obtaining a relatively large sample size in a short amount of time and for a reasonable cost. Research studies investigating the demographics of the AMT worker population have demonstrated that AMT workers are more demographically diverse when compared to traditional undergraduate research subjects, (Buhrmester, Kwang, & Gosling, 2011), and that data obtained from AMT workers for research purposes is just as reliable as that obtained in a laboratory, as long as certain precautions regarding data quality are maintained, such as including attention check questions, recruiting only participants with high reputation scores, and using third-party software sites to ensure the proper, randomized dissemination of experiments (Buhrmester et al., 2011; Litman, Robinson, & Abberbock, 2016; Munro et al., 2010; Peer, Vosgerau, & Acquisti, 2014).

The 423 recruited participants for this study represented an approximate 50/50 split among males and females (females = 216, males = 209). The average age of participants was 34 (SD = 11.14), with a highest age of 71 and a lowest age of 20. Seven of the participants indicated their first language was not English, with their first languages including Spanish (3), Russian (2), Tagalog (1), and Indonesian (1). All but one of the participants completed high school, with 231 (55%) of the participants completing a four-year college degree or higher. When asked to self-
rate their socioeconomic status, 399 (94%) of the participants indicated they were in the middle class or lower, whereas the remaining 26 participants indicated they were near or at the top of the socioeconomic ladder.

5.5 Materials

5.5.1 News story prompts.

Two separate news stories were selected as prompts to elicit the creative responses. These news stories were selected from the American Voices section of the satirical news website The Onion. The American Voices section presents a headline and brief summary of a real news story alongside fictional, satirical responses to the news story purported to be from regular readers of the website. For example, The Onion published the following headline and story:

**Pentagon Has U.F.O. Hunting Program**

*The Defense Department has been investing $22 million per year into investigating unidentified flying objects, a New York Times report found, contradicting government statements that the program was shut down in 2012. What do you think?*

The fictional reactions provided by The Onion included the following quotes along with the person’s name and their occupation:

1. “I knew it! The Pentagon exists!” - Fred Dalton, Wishing Well Technician
2. “It’s like a real-life X-Files where all the Mulders must’ve been extremely disappointed all the time.” - Harriet Samson, Pistachio Sheller
3. “Surely there are more productive wastes of taxpayer money.” - Nick Broderick, Confetti Organizer
As can be inferred from the occupations listed for the quoted readers (as well as the very nature of being printed on *The Onion*), these reactions are fictional and were designed to provide satirical and humorous reactions to the legitimate news story presented above. Based on the example set by *The Onion* of providing fictional and creative reactions to real news stories, two different news stories from the website were selected. The stories were purposefully selected to contain no political or religious content in order to avoid affective responses that might bias responses from the participants (LaMarre et al., 2009). For each of the two news stories chosen, four fictional responses were pre-constructed by the researcher. These responses differed in that they either contained no figurative language, contained sarcasm only, contained metaphor only, or contained metaphor and sarcasm. The two news stories and responses are shown below:

**Story 1: Woman Sues Starbucks Over Amount Of Ice In Drinks**

*A Chicago woman has filed a $5 million lawsuit against the Starbucks Corporation claiming false advertising of 24-ounce iced drinks that allegedly only contain 14 ounces of liquid. What do you think?*

1. **No figurative language**: “This lawsuit is a waste of time.”
2. **Sarcasm only**: “I’m so glad our legal system is spending time on such important issues.”
3. **Metaphor only**: “This woman is a parasite wasting the time of our judicial system.”
4. **Sarcasm and Metaphor**: “I’m so glad our legal system is spending time on parasites like this.”
In this news story, the first statement in the no figurative language condition directly expresses the speaker’s concern that the lawsuit is all a waste of the court’s time. In the second statement, sarcasm is signaled through the speaker’s exaggeration of the importance of the lawsuit, which is frivolous in nature yet has a high award amount associated with it (i.e., 5 million dollars), and expresses the opposite meaning of the no figurative language condition on the surface level, with the same meaning expressed at the figurative level. In the third and fourth statements, the woman named in the news story is construed as a parasite, a common type of conceptual metaphor that describes objectionable human behavior as animal behavior (Kövecses, 2002). Sarcasm is signaled in the fourth statement in a similar manner to the second, with the speaker expressing apparent joy about the news of the lawsuit, which clashes with the description of the woman as a parasite in the same sentence.

**Story 2: Octopus Escapes New Zealand Aquarium**

_Inky the Octopus is believed to have escaped the national aquarium of New Zealand by opening the lid of his tank, slithering across the floor, then squeezing through a 5-inch-wide drainage pipe and out to sea. What do you think?_

1. No figurative language: “The sea will be much better than his tank.”
2. Sarcasm only: “I’m sure the sea will be much worse compared to his cramped tank.”
3. Metaphor only: “Good, the sea is a much better place than that prison cell.”
4. Sarcasm and Metaphor: “I’m sure the sea will be much worse compared to his prison cell.”

---

6 Although the phrases waste of time and spending time index a conceptual metaphor describing time as a currency, these phrases are well entrenched and conventionalized when compared to the other examples of figurative language and do not directly refer to the main entity in the story (i.e., do not create a comparison of the woman or Starbucks).
In the second news story, the *no figurative language* response directly expresses the opinion that Inky the Octopus is better in the sea than in the aquarium. In the second statement, the same opinion is provided sarcastically, with the speaker feigning favor for the negative, cramped conditions over the natural home of the octopus (i.e., the sea). In the *metaphor only* condition, the aquarium tank is described as a prison cell, activating a conceptual metaphor between the target domain of prison and the source domain of aquariums and zoos. The fourth statement combines conditions three and four to produce a sarcastic effect employing the same conceptual metaphor (i.e., an aquarium is a prison).

For both of the news stories, all four of the responses contain similar semantic meaning (i.e., all Story 1 reactions imply the court case is a waste of time, and all Story 2 reactions imply that the sea is a better home for the octopus). Additionally, the responses all present similar pragmatic meaning, demonstrating the opinion of the speaker (i.e., all Story 1 reactions show disapproval towards the woman, and all Story 2 reactions show a preference for the octopus to be back in the sea). Accounting for both of these meaning types (semantic and pragmatic) are important and often overlooked components of figurative meaning (Gibbs & Colston, 2012). Therefore, the primary difference among the four conditions rests in how language is used in terms of the differing amounts of figurative language.

### 5.6 Procedure

The two news stories and four pre-constructed news responses were entered into the online survey software Qualtrics. A total of 16 versions of the survey were created in order to randomize and counterbalance the news stories and four possible pre-constructed responses. For example, the first version displayed the *no figurative language* response for both news stories, the second version displayed the *no figurative language* response for the Starbucks story and the
metaphor only response for the Inky the Octopus story, and so on. Moreover, the order of the two tasks was randomized within each of the 16 experiments, so that there was an equally as likely chance to read either the Starbucks story or the Inky the Octopus story first or second. The experiment was then posted to AMT using the third-party website TurkPrime (Litman et al., 2016), which prevents workers from completing the experiment more than once by recording their AMT identification number, and also recruits workers in smaller batches spread out over time, which further helps randomize the demographics of workers (e.g., avoids recruiting only people who are awake at 10am). Only workers who lived inside the United States were eligible for participation, and workers were required to have a minimum approval rating of 99% for at least 10,000 completed jobs to participate in the study. Workers who met this requirement first read a brief description of the experiment, which informed them that they would write one-sentence creative reactions for two news stories and also complete a demographic survey in exchange for $1.00 USD. Workers who agreed to participate then followed a link to the Qualtrics survey and were forwarded to one of the 16 different survey versions. Qualtrics was programmed so that each version of the survey received a minimum of 25 participants and assigned participants into different survey versions in a random yet counterbalanced order.

Once participants began the survey, they first read a page containing the informed consent, which described the nature and purpose of the research, ensured the workers their data would remain confidential, and reiterated that they would receive $1.00 USD for their participation (see Appendix L). Participants who agreed then completed a short demographic survey asking them for their age, gender, first language, number of fluent languages known, level of education, and socioeconomic status. After completing the survey, participants read the first instruction screen, which informed them that their task was to develop creative news reactions
for two real news stories. Participants were given a definition of creativity as *a novel yet effective solution to a problem*. They were then told the basic order of the task was to read a story, provide their creative reaction to the story, and then compare their response to another response for the same story, deciding whether their response was more, less, or equally as creative as the other response. Participants were not informed that the responses to which they compared their own responses were pre-constructed by the researcher.

Participants were then briefed on the nature of the task in more detail by being informed that they were to pretend to be a 24-hour news channel host whose job was to develop interesting and creative reactions to different news headlines. Participants were instructed to pretend that the news channel’s ratings depended on their ability to hold a fictional audience’s attention, and therefore their job relied on their ability to be creative. They were asked to be as creative as possible, and also told that their reactions should reflect something they would actually say during live television approximately equivalent to one written sentence in length. The instructions further explained that participants should type exactly what they would say, as they would say it, without including any additional detail about their thoughts or reactions to the story. After reading these instructions, participants were presented with the first news story, which presented the headline and brief summary of the story above a reminder of the task and a blank text box to enter their answer. After they entered their creative news response, participants then saw their answer displayed above one of the four pre-constructed answers designed by the researcher (i.e., the *no figurative language*, *metaphor only*, *sarcasm only*, or *sarcasm and metaphor* conditions), and were asked to choose whether their answer was more, less, or equally as creative when compared to the pre-constructed answer. After making this decision, participants then proceeded to repeat the task for the second news story. Once completing the
task for the second news story, participants were informed the task was complete and were paid $1.00 USD for their participation. All of the participant data was manually checked to ensure participants completed the task seriously. Participants whose data demonstrated a lack of attention to the task (i.e., provided nonsense or blank answers) were replaced until at least 25 participants provided sincere answers for each of the 16 survey versions, (eight participants were replaced in this manner), resulting in the final sample size of 423. This procedure is summarized in Figure 5.1.

![Figure 5.1 Procedure flow](image-url)
5.7 Data Coding

At the end of the experiment, the participant response data was downloaded from Qualtrics and coded for the presence or absence of linguistic creativity. To do so, the researcher read each response provided by the participants and marked whether it did or did not include instances of figurative or creative language (e.g., sarcasm, metaphor, wordplay). Although the pre-constructed answers provided to participants as comparisons contained either no figurative language, metaphor, sarcasm, or sarcasm and metaphor, the codes for the participants’ answers were not limited to these categories. Instead, any use of any figurative or creative language in the participants’ answer was coded as a 1, whereas the absence of any figurative or creative language was coded as a 0. The entire data set was first coded by the researcher, and then a random sample of 20% of the data was coded by a second researcher, with agreement between the two reaching 93%. Results from the coding identified the presence of creative language in a total of 185 (21.8%) of the participants’ answers. The Starbucks task included a higher number of instances of linguistic creativity when compared to the Inky the Octopus task (101 [23.8%] vs. 84 [19.8%]), although for both of the tasks, the majority of answers included no linguistic creativity.

5.8 Statistical Analysis

A single linear mixed effects (LME) model was constructed in order to answer the research question for this study. The dependent variable for the LME model was the participant comparison ratings for their answers (i.e., more creative, equally creative, less creative). Because the dependent variable was categorical with three levels, an ordinal regression procedure was used. An ordinal approach is similar to the logistic LME model employed in Chapter 4, but differ in that this model predicts the probability of three outcomes, rather than two. Therefore, the interpretation of probabilities and odds ratios for ordinal regression is slightly different, in that
the chances of predicting a higher level category is compared against the probability of all lower
levels combined (Field, 2013; Levshina, 2015). In the context of this data, positive coefficients
and odds ratios greater than one would reflect that there was a higher chance of participants
indicating their news response was more creative when compared to the combined probability of
that the participants choosing equally or less creative. Negative coefficients and odds ratios
lower than one would reflect the opposite interpretation, reporting a lower chance of being rated
as more creative when compared to the combined probability of the other two options.

The model was built in R (R Core Team, 2017) using the clmm function from the ordinal
package (Christensen, 2015). The model contained participant ratings (more, less, or equal) as
the dependent variable and participants’ age, gender, and socioeconomic status\(^7\) as fixed effects,
as well as fixed effects of task (Starbucks vs. Inky the Octopus), task order (first vs. second),
presence of linguistic creativity in participants’ responses (yes vs. no), and comparison condition
(no figurative language, sarcasm, metaphor, or sarcasm and metaphor). Participants and items
were entered as crossed random effects, and an interaction was tested between presence of
linguistic creativity and figurative language condition. Odds ratios were calculated through
exponentiation of the coefficients, providing a measure of effect size for each predictor variable
(Levshina, 2015), with 95% confidence intervals for odds ratios computed using the confint
function from the base stats package in R. All numerical predictors were z-scored before being
entered into the model in order to center and standardize the numerical scales. Because the model
was an ordinal LME models, it was not possible to obtain general measures of model effect sizes
(i.e., variance explained through \(R^2\)).

---

\(^7\) Socioeconomic status was converted to an ordinal numerical variable before being entered into the model. These
variables were included in order to control for any possible differences demographic differences that may affect
perceptions of creativity. This study contains no specific predictions about their effects, if any.
5.9 Participant Creativity Ratings

Figure 5.2 displays the percentages and frequencies of participant self-ratings of creativity for each of the four different comparison conditions. As can be seen, the vast majority of ratings indicated that participants reported that their news reactions were more creative than the pre-constructed news reactions across all four conditions (*no figurative language*, *sarcasm only*, *metaphor only*, and *sarcasm and metaphor*). Indeed, of the 846 total ratings, 550 (65%) of the ratings indicated that participants felt their news response was more creative when compared to the pre-constructed responses. However, when comparing the frequency of *more creative* ratings among different pre-constructed response conditions, the *no figurative language* condition resulted in the highest percentage of participants rating their news reactions as more creative, as well as the lowest percentage of participants rating their news reactions as less creative. Among the three figurative language comparison conditions, the *sarcasm only* condition contained the highest percentage of participants rating their news reactions as more creative (67.5%), with the *metaphor only* and *sarcasm & metaphor* conditions having lower percentages of participants rating their news reactions as more creative.
Figure 5.2 Participant creativity self-ratings
Percentages of participant self-ratings for whether they felt their news reactions were more, less, or equally as creative when compared to the researcher-prepared news reactions. Raw frequency counts are included under each percentage.

5.10 Linguistic Creativity in Participant Responses

Figure 5.3 displays the instances of linguistic creativity in participants’ answers and their self-ratings of creativity when compared to the pre-constructed news reactions. As Figure 5.3 displays, the percentage for selecting a rating of more creative is lower if the participant’s answer did not include linguistic creativity when compared to answers that did contain linguistic creativity (~59% vs. ~87%). The relatively high percentage of more creative ratings for answers containing linguistic creativity means that in almost every instance, if a participants’ answer contained linguistic creativity, they also felt that their new reaction was more creative than the researcher-provided news reaction. While there was still a high percentage of participant
responses that did not contain linguistic creativity self-rated as more creative, the participant answers with no linguistic creativity also contained higher instances of equally creative and less creative when compared to the answers that did contain linguistic creativity.

![Figure 5.3 Interaction between self-ratings and incidence of linguistic creativity](image)

Participant self-ratings of news stories compared between answers that contained and did not contain linguistic creativity. Raw frequency counts are included under each percentage.

5.11 Ordinal Linear Mixed Effects Model

An ordinal linear mixed effect model with participant ratings of creativity (more, less, or equally creative) as the dependent variable and participant age, sex (male vs. female), socioeconomic status, pre-constructed comparison condition (no figurative language, sarcasm only, metaphor only, and sarcasm and metaphor), task (Starbucks vs. Inky the Octopus), task
order (first vs. second), and presence of linguistic creativity (yes vs. no), and an interaction between pre-constructed comparison condition and presence of linguistic creativity included significant main effects for pre-constructed comparison condition, age, and presence of linguistic creativity. The interaction between pre-constructed comparison condition and presence of linguistic creativity was not significant and thus discarded from the model in order to properly interpret the main effects. Table 5.1 displays the estimates, standard error, z and p values for the predictors, along with the odds ratios and 95% confidence intervals.

<table>
<thead>
<tr>
<th>Table 5.1 Ordinal linear mixed model predicting participant comparison ratings of creativity</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Variable</strong></td>
</tr>
<tr>
<td>Age*</td>
</tr>
<tr>
<td>Sex: Male</td>
</tr>
<tr>
<td>Socioeconomic Status</td>
</tr>
<tr>
<td>Task: Starbucks</td>
</tr>
<tr>
<td>Order: Second</td>
</tr>
<tr>
<td>Sarcasm</td>
</tr>
<tr>
<td>Metaphor*</td>
</tr>
<tr>
<td>Sarcasm &amp; Metaphor*</td>
</tr>
<tr>
<td>Linguistic Creativity: Yes*</td>
</tr>
</tbody>
</table>

* Significant effect, OR = Odds Ratio Baselines for categorical variables: Sex = Female, Task = Inky the Octopus, Figurative Language Condition: No Figurative Language, Linguistic Creativity: No

Odds ratios for the significant main suggest that each increase in one standard deviation of age led to a 1.3 times greater likelihood the participant would rate their answer as *more creative* when compared to the likelihood of rating is as *equally creative* or *less creative*. The odds ratio for presence of linguistic creativity reports that if an answer contained linguistic creativity, there was a 6.1 times greater likelihood of that answer being rated as *more creative* when compared to the likelihood of being rated as *equally or less creative*. For the significant effect of pre-constructed comparison condition, odds ratios reported that the *metaphor only* and *sarcasm and metaphor* conditions resulted in a significantly less likely chance for the
participants to rate their answers as more creative when compared to the no figurative language condition (i.e., the baseline level in the model). Specifically, the metaphor only condition was 55% less likely, and the sarcasm and metaphor condition was 59% less likely to elicit more creative ratings when compared to the no figurative language condition, indicating that the no figurative language condition resulted in significantly higher incidences of more creative ratings (suggesting participants found their answers significantly more creative than the no figurative language condition when compared to two of the three comparison conditions that did contain figurative language). Follow up pairwise comparisons reported no significant differences among the sarcasm only, metaphor only, or sarcasm and metaphor conditions.

5.12 Discussion

The purpose of this study was to further investigate connections between creativity and creative language by examining whether perceptions of creativity interacted with the presence or absence of creative language (e.g., metaphor or sarcasm). To do so, 423 online participants were asked to produce creative reactions to two short news stories and then also provide self-ratings of creativity when compared to sample news reactions prepared in advance by the researcher, each of which contained differing levels of figurative language. By doing so, this study empirically tests whether the simple presence or absence of creative language influences perceptions of creativity, which may in turn provide a better understanding into the relation between two different conceptualizations of creativity (i.e., psychological and linguistic).

Results demonstrated that a majority of ratings (65%) indicated that participants felt their new reactions were more creative than the pre-constructed news reactions. One possible explanation for the relatively higher frequency of more creative ratings may be related to conceptualizations of egoistic versus moralistic influences on self perception (Paulhus & John,
Research on biases in self-reporting has identified two basic motivations that can influence answers on surveys, job applications, and other measures of self-reporting. Specifically, participants may be egoistic, in that they have very positive self-perceptions of intelligence, creativity, and other skills (Paulhus & John, 1998). Conversely, participants may be moralistically biased with increased self-perceptions of agreeableness and sense of duty. Because the prompt for this study implored the participants to be as creative as possible, and also because the participants’ fictional job as a news caster relied on their ability to be creative and interesting, it is possible that participants fed into an egoistic bias which was job related (Honkaniami & Feldt, 2008), in that they intended to increase their perceived usefulness for the task and help ensure successful completion and payment. However, more research is needed to examine this tendency to self-report as more creative than the pre-constructed news reactions.

Results from the LME model also suggest that the different amounts and types of figurative language in the pre-constructed news reactions significantly affected participant self-ratings of creativity. First of all, participants were significantly more likely to report their news reaction as more creative if they were comparing their news reaction against a pre-constructed news reaction that contained no figurative language when compared to the metaphor only and the sarcasm and metaphor pre-constructed comparison conditions. This suggests that the presence of figurative language in the pre-constructed news responses significantly increased their perceived creativity, as evidenced by participants having a lower likelihood to rate their answers as more creative in comparison, but only in the conditions that contained metaphors.

While there were no significant differences in participant ratings among the three pre-constructed conditions containing figurative language, the different effect sizes and levels of significance for these conditions (i.e., the likelihood of participants rating their news reactions as
more creative in comparison) when compared to the no figurative language condition suggests that metaphor was perceived as more creative than sarcasm. This is because the sarcasm only condition was not identified as eliciting significantly different self-ratings of creativity, as it was only 44% less likely to elicit more creative ratings when compared to the no figurative language condition (see Table 5.1). Conversely, the metaphor only and metaphor and sarcasm conditions were significantly different and were 55% and 59% less likely to elicit more creative ratings when compared to the no figurative language condition. Metaphor is typically more linguistically creative than sarcasm, as it relies on apt, conceptual comparisons between two entities as a means to impart a specific description of one entity by assigning it qualities of a seemingly unrelated entity (Gibbs, 1994, 2011; Glucksberg, 2001). Accordingly, metaphor uses combinations of words to apply new interpretations or understandings of a particular entity (e.g., all lawyers are sharks), and the metaphorical meaning imparted is figurative and metaphorical regardless of the context within which it is used. Sarcasm, on the other hand, is a form of verbal irony that relies heavily on contextual incongruity to highlight the implied meaning of an utterance that would otherwise be considered literal or non-figurative (Colston, 2017), and does not require the same level of linguistic creativity as metaphors. For example, depending on the context, phatic utterances such as thank you can be perceived to be sarcastic or non-sarcastic (e.g., after receiving good or terrible customer service). Therefore, differences in the likelihood of participants rating their answers as more creative between the pre-constructed answers containing figurative language may reflect a tendency for metaphor to be perceived as more creative than sarcasm. Even though these conditions did not significantly differ, more research directly testing this hypothesis is warranted.
The LME model also included a significant main effect for participant answers containing linguistic creativity. Specifically, participants who included some form of linguistic creativity in their news reaction were 6.1 times more likely to rate their news reaction as more creative, regardless of the type of pre-constructed news reaction they were comparing their answer to. Upon closer inspection, a majority of the linguistic creativity that occurred in the participant news reactions was an example of wordplay or a pun, in that the participants played with the multiple meanings of words. Indeed, many of the answers containing wordplay for the Starbucks task involved puns or playful uses of idiomatic expressions related to coffee, temperature, or ice, focusing on the notion that the woman suing Starbucks was upset about receiving ice in her iced coffee. For example, the following response “That's a latte money to be asking for. She seems quite steamed” plays with the sound of the word latte (i.e., the reduced form lotta from a lot of) as well as describing the woman as steamed, which takes on a double meaning of being upset as well as the connotations related to preparing espresso and other coffee drinks. Other participants focused on the importance of ice in the story by rhyming ice with nice (e.g., Too much ice isn't nice?). Slightly different wordplay strategies were employed for the Inky the Octopus story, with most participants purposefully manipulating the form and sounds of words in order to develop puns related to the word Inky or Octopus. For instance, the response “Sounds like some zookeepers are going to be octopied finding it” manipulates the form of the word occupied by forcing a morphological blend between the words occupied and octopus. Another participant combined the word octopus with kidnapped to make a new blend (Are they sure he wasn't Octopus-napped?).

Puns and wordplay are commonly associated with genres related to news and broadcasting. In a study of 2183 headlines for print ads in newspapers and magazines, Leigh
reported that 74.3% of the headlines contained some form of figurative language, and that 19% of those headlines contained puns. Chovanec (2005) analyzed news headlines in British sports reporting of the 2004 European Football Championship, which featured a final competition between the Czech and Greek national football teams. Chovanec reported a wide number of examples of puns and wordplay in headlines based on the two teams, such as describing a game loss for the Greek team as a tragedy (evoking classic Greek culture) or playing with the sound of Czech (e.g., Cashing in on the Czechs). Ritchie’s (2004) book-length account of humor draws heavily from examples in news media when explaining puns. Partington (2009) investigated a corpus of British newspapers and found that headline authors relied on a diverse and robust set of wordplay and punning strategies. These studies attest to a known relation between wordplay/puns and print advertisements and news, which may have prompted participants in this study to use similar attention-grabbing devices when developing their creative news reactions. While puns are generally perceived to involve manipulation of both form and meaning, classifying the previous example using octopied an imperfect pun (Hempelmann, 2004; Ritchie, 2004), these examples are all overt attempts on the part of the participants to be linguistically creative. As such, these results suggest that participants who consciously manipulated the form or meaning of language felt their answers were more creative, providing more evidence of a relation between linguistic creativity and perceptions of creative ability.

In addition to the presence of creative language, the final significant effect included in the first LME model was age, in that increases in age led to a higher likelihood of participants rating their news reaction as more creative. This suggests that older participants had higher self-perceptions of creativity in regards to their news reactions. Surveys of aging artists demonstrate that they view aging as a positive influence on creativity, because new experiences that come
with greater age lead to new ideas and approaches to artistic styles (Lindauer, Orwoll, & Kelley, 1997). Participants with a higher age (64-89) also outperformed younger participants (22-35) on a conceptual comparisons task, where participants were asked to interpret meanings from atypical noun-noun combinations such as *fog scarves* and *honey road*, suggesting that a higher age allowed for a greater range of conceptual concepts from which to draw novel and creative interpretation (Marshal & Coblentz, 2014). The findings here further suggest a positive relation between age and creativity, although this time in regards to self-perceptions of creativity.

### 5.13 Conclusion

The purpose of this study was to empirically test whether the perceptions of creativity are influenced by the presence or absence of creative language. The results demonstrated that figurative language was perceived to be more creative than non-figurative language, but only for answers containing metaphors. The results also suggested that the employment of linguistic creativity on the part of the participants significantly increased their self-perceptions of creativity. Although this study only examined answers for one creative task using two separate prompts, it does provide evidence of a relation between creative language and perceptions of creativity. Future work in this area should examine this phenomenon in a wider range of contexts, including different tests of creativity and a wider range of creative language types.

### 6 Conclusion

The main motivation for this dissertation was to analyze influences beyond the literal-figurative binary definition of meaning that is commonly investigated in studies of figurative language use (e.g., comparing reading times between literal and figurative sentences). Although the conceptualization of figurative and literal meaning is inherent to the definition of figurative language and is most likely what draws attention to figurative language in the first place, the
focus on differences in figurative and literal meaning may cause other important factors in figurative language use to be overlooked. Accordingly, the purpose of this dissertation was to conduct three original experiments that cohered to provide a better understanding into various factors that may influence figurative language perception and use. Results from those three experiments highlighted influences attributable to individual differences, linguistic features, and perceptions of creativity in metaphor, sarcasm, satire, and wordplay. Overall, these results demonstrate the importance of adopting a dynamic approach towards understanding figurative language use, because the results attested to the difficulty of isolating a single influence on figurative use. In totality, the results provided answers to the main research questions for this dissertation, which were:

1. What individual differences (including language background), if any, influence the production and processing of figurative language?

2. Do linguistic features related to lexical sophistication and cohesion contribute to a better understanding of figurative language?

The simple answer to both of these questions is that many individual differences and linguistic features influence the processing, production, and perception of figurative language, but that these effects are modulated by a number of other variables, such as the type of figurative language, the language background of the participants, affective perceptions, the operationalization of figurative language quality, and the experimental prompts designed to elicit figurative language production. Therefore, while it is possible to list all of the individual significant effects as they appeared in each of the three studies, the purpose of this chapter is not to summarize specific influences of figurative language use that can be generalized, but rather to further emphasize the dynamic and contextual processes that occur during figurative language
use. At the same time, just because some variables in these studies did not emerge as significant predictors does not mean they should be ruled out as potential influences on figurative language use. Rather, they were not significant in the specific context of these studies.

The first two studies in this dissertation were based on data collected from the same group of participants recruited from among the undergraduate and graduate student population at Georgia State University. These participants completed a wide battery of individual differences tests (e.g., intelligence, working memory capacity), demographic surveys, a metaphor and sarcasm production experiment, and a satire processing experiment. In the first study (Chapter 3), the participants’ metaphors and sarcastic responses were analyzed using human ratings and automatic text analysis tools in order to determine if any of the individual differences or linguistic features reported by the surveys and text analysis tools were predictive of the human ratings of metaphor and sarcasm. Although not the initial purpose of the study, results from the human ratings demonstrated that for both the metaphors and sarcastic responses, ratings related to the creativity of the figurative language (i.e., combined novelty and mirth ratings) were separate constructs when compared to theoretical definitions of the figurative language (i.e., conceptual distance for metaphor and incongruity for sarcasm). Moreover, statistical models reported stronger effect sizes when predicting the creativity (i.e. novelty and mirth) of metaphors and sarcastic responses when compared to the theoretical constructs of metaphors and sarcastic responses (i.e., conceptual distance and incongruity), meaning that many of the features measured in the first study (individual differences and linguistic features) were more strongly related to perceived creativity in figurative language, rather than the ability to adhere to theoretical definitions of figurative language. Therefore, these findings suggest that the human raters may have relied on generalized definitions of creativity when rating the metaphors and
sarcastic responses, which in turn suggests potential links between creative ability and figurative language use that should be investigated in future studies.

Furthermore, many of the significant predictors of both metaphor and sarcasm production quality interacted with other experimental features, such as metaphor type (novel vs. conventional), the sarcasm prompts (black and white, desert island, and three-panel comic), as well as with other predictor variables (e.g., English age of onset interacting with Broad Retrieval for sarcasm incongruity scores), demonstrating that many of the individual differences predictive of metaphor and sarcastic response quality depended on the precise type of metaphor or sarcastic context. For example, it is not surprising that participants with greater background knowledge (i.e., Crystallized Intelligence) produced metaphors with greater conceptual distance scores for conventional as compared to novel metaphors, and that participants with greater working memory capacity produced metaphors with greater conceptual distance scores for novel as compared to conventional metaphors, because those different metaphor types tap into different cognitive abilities. Thus, while it would be simple to state that increased cognitive ability as measured through intelligence and executive control leads to better metaphor production ability (via the ability to connect more remote concepts), it is more appropriate to mediate the answer based on the specific type of metaphor being discussed. The same could also be said for production time or English age of onset; while these variables seemed to exert consistent effects on figurative language production ability (i.e., higher English age of onset led to lower ratings, and higher production time led to higher ratings), these features also interacted with prompt type and other individual differences such that their effects always depended on other effects and contextual variables.
Similar conclusions can be drawn regarding the role of linguistic features of metaphor and sarcasm production, although there was also more consistency in the findings when compared to the individual differences analysis. Specifically, the role of lexical sophistication was more straightforward in regards to metaphor production, with higher incidences of linguistic features representative of lexical sophistication generally, but not always, leading to higher ratings of conceptual distance or creativity among the metaphors, regardless of whether the metaphors were novel or conventional. In comparison, lexical sophistication did not strongly influence sarcasm incongruity scores, and differing levels of lexical sophistication in the sarcastic responses interacted with the sarcasm prompt, such that increased levels of lexical sophistication led to higher novelty/mirth scores for sarcasm prompts with more contextual information (i.e., three panel comics) but lower novelty/mirth scores for prompts with lower amounts of contextual information (i.e., desert island comics). The end result is that linguistic features were able to predict a greater amount of variance in human ratings for the metaphors when compared to the sarcastic responses, which is most likely a result of metaphorical meaning being more self-contained in the linguistic context of a metaphor, whereas sarcastic meaning is reconciled through a consideration of the surrounding situational context and thus less reliant on the linguistic features of sarcastic responses alone.

In short, the results from Study 1 emphasize the dynamic role of different individual differences and linguistic features on metaphor and sarcasm production ability, and suggest that it may be difficult to ever truly control for these features or find generalizable influences on figurative language production ability. That being said, the results do demonstrate that the cognitive and demographic backgrounds of participants are important to take into account when measuring figurative language use, and any experiment that does not consider these differences
may overlook significant influences or improperly attribute differences to a limited range of independent variables. Moreover, although the linguistic features accounted for a relatively small amount of variance in human perceptions of figurative language quality, these results still attest to the effect that specific linguistic features can have on human perceptions of figurative language and creativity.

Study 2 (Chapter 4) was similar to Study 1, except that it investigated the role of individual differences, affective perceptions, and linguistic features in regards to the processing and comprehension of satirical texts. The first finding of this study reported that the satirical and non-satirical texts were significantly different when measured for affective perceptions of humor, sincerity, and positivity, serving to support theoretical and intuitive definitions of satire that describe it as a negative, subtle form of ironic language that can result in humor. Despite these differences, there were no significant differences in base reading times between satirical and non-satirical texts, and there were also no significant interactions among any of the other predictor variables (i.e., individual differences, linguistic features, affective perceptions) and text type. This suggests that satirical and non-satirical meaning are computed or processed at the same rate, and in turn support some theoretical models of figurative language processing over others. However, results from the final statistical analysis in this study suggests that satirical processing times may not be a valid measure of satirical comprehension, because the processing times for the satirical texts were not predictive of comprehension.

Specifically, the model testing satirical comprehension reported several other variables that were significant predictors of satirical comprehension, including affective perceptions of positivity, sincerity, and humor, as well as English age of onset and average lexical frequency of content words. That affective perceptions were strong predictors of satirical comprehension
further emphasizes the importance of collecting multiple types of evidence when measuring figurative language comprehension and also serve to reinforce the relationship between figurative meaning and pragmatic, contextual interpretations. Indeed, participants who were able to recognize the pragmatic insincerity of the author, found the text to be more negative, and those that also found the text to be humorous were better able to produce a paraphrase of the text that was reflective of a satirical interpretation. In other words, participants took multiple sources of information into account when engaging with the satirical texts, and these influences could not be measured through processing time alone. At the same time, the lexical frequency of the satirical texts influenced comprehension, in that satirical texts with higher word frequency (i.e., lower lexical sophistication). This finding further highlights the potential role that linguistic features of stimuli may have on figurative language comprehension. Finally, a greater English age of onset was a significant, negative predictor of satirical comprehension, with participants who reported that they started learning English later in life significantly less likely to produce a paraphrase reflective of satirical comprehension. As mentioned in Chapter 4, it is most likely that this variable captured differences in cultural familiarity and exposure, which would in turn have a large influence on the ability to understand American satire.

Study 3 differed from the first two studies in that it more directly investigated potential connections between perceptions of creativity and figurative language. The results from Study 3 strongly suggest that figurative language does indeed influence perceptions of creativity, in that participants were significantly less likely to rate their own answers to the provided news stories as more creative than pre-constructed answers containing metaphor, but not significantly so for pre-constructed answers containing sarcasm. An even stronger effect was found for participants who included some form of purposeful linguistic creativity in their own responses, with those
participants being significantly more likely to rate their own answers as more creative when compared to any of the pre-constructed news responses. Study 3 thus provides empirical links between the presence of linguistic creativity (either as figurative language or wordplay) and perceptions of creativity, which serves to validate intuitive claims that examples of creative language (e.g., metaphor) are also associated with creative ability, but also suggests some types of figurative language are perceived to be more creative than others (i.e., metaphor is more creative than sarcasm).

6.1 Future Directions

The results from all three studies in this dissertation outline broad influences on figurative language use. Some logical future directions are to consider the depth of some of these influences. In regards to figurative language production, more work should be undertaken examining contextual and pragmatic influences on sarcasm production, because the different sarcasm prompts significantly interacted with several of the individual differences and linguistic features variables. One possible way to examine the role of contextual information on figurative language production is to purposefully test prompts with varying degrees of contextual information, such as including a desert island prompt with or without information about the people depicted in the cartoon. Investigations of figurative language production could also benefit from the use of more naturalistic data, which could be captured via recording of spontaneous conversation or through analyses of previously existing corpora.

Although there were not significant differences located between satirical and non-satirical text processing, the processing of satire should also continue to be investigated using more granular measures of processing, including eye-tracking methods and measures of embodied cognition, such as tracking mouse movements as participants answer questions related to the
comprehension of figurative language. At the same time, explicit tests of satire processing using discourse processing frameworks (e.g., situation models) may help to explain figurative language processing using processing and comprehension models that already exist and may be adaptable to figurative language. This would obviate the need to continue developing separate processing models and mechanisms specifically tailored to figurative language.

The role of language knowledge during figurative language use should continue to be investigated. Results from the second study investigated satire processing and comprehension indicated that a higher English age of onset resulted in a significantly less likely chance to understand the satirical message of *The Onion* articles, but it is unclear whether this effect is related to cultural knowledge or other variables related to language proficiency. This finding is also difficult to interpret because despite gathering detailed profiles of the participants’ language backgrounds, the disparate nature of the participants’ differing first languages made it difficult to group the participants into homogenous language user groups. Indeed, many of the participants in the first and second study spoke more than one language (including the native English speakers), and the proficiency data in this dissertation was only based on participants’ self reporting. Future studies investigating figurative language use and language background should consider employing more standardized tests of English proficiency appropriate for specific tasks, such as reading tests or c-tests that can tap into the lexical knowledge or reading ability of the participants. Additionally, recruiting participants with relatively homogenous L1 backgrounds (e.g., recruiting only L1 Romanian L2 English users) could allow for more detailed analyses comparing linguistic features from participants’ first and second languages, and how these features may influence figurative language use.
The relation between lexical sophistication and figurative language use provides a final avenue for future research. The results from the studies in this dissertation suggest that linguistic features representative of lexical sophistication may play more of a role in the production and interpretation of figurative language because certain types of figurative language are more situated in their immediate linguistic context (e.g., metaphor). Yet, linguistic features related to lexical sophistication played less of a role in both sarcasm and satire (Chapter 3 and 4), suggesting that the surface level linguistic forms of these two types of figurative language were less associated with human raters’ perceptions of figurative language quality or with how satire was processed (aside from word frequency, which may have been related to the news genre itself). Regardless, future research specifically testing different versions of the same figurative utterance or text (e.g., satirical news stories) which contain differing levels of lexical sophistication (e.g., word frequency) may provide a better understanding into connections between lexical sophistication and figurative language.

Finally, figurative language production and comprehension interacts with a number of social, pragmatic, and contextual variables (Colston & Katz, 2005), and these variables can be difficult to quantify or are reduced into homogenous groups that may blur individual variation (e.g., broad comparisons between males and females). And when these variables produce a significant effect (e.g., males receiving higher novelty/mirth scores than females in Study 2), it can be difficult to interpret why that effect exists. Therefore, future research involving these variables may benefit from employing qualitative research methods such as interviews, think aloud protocols, or stimulated recalls that may in turn complement the quantitative findings.
6.2 Conclusion

There are a wide number of factors that influence figurative language use, and this dissertation has identified several of them by conducting three original empirical studies into the production, processing, and perception of figurative language. Nevertheless, there still remain a large number of potential areas to continue investigating figurative language from a dynamic, contextualized perspective. Whether one is studying metaphor, sarcasm, satire, or some other form of figurative language, it is important to carefully consider how features beyond literal and figurative meaning, such as participant individual differences, linguistic features, and multiple comprehension measures may influence investigations of figurative language use.
REFERENCES

Creativity Research Journal, 29(2), 133–144.
https://doi.org/10.1080/10400419.2017.1302776


https://doi.org/10.1002/hbm.22242

https://doi.org/10.1037/cep0000011


https://doi.org/10.1080/0163853X.2012.687863


https://doi.org/10.1016/j.jml.2006.11.006


https://doi.org/10.1080/15248370802247939

https://doi.org/10.1080/01638539709544980


https://doi.org/10.1177/0146167209351886


https://doi.org/10.1027/1614-0001/a000119


https://doi.org/10.1515/humor-2014-0094


https://doi.org/10.1007/978-3-319-17464-8_6


https://doi.org/10.1080/0163853X.2014.954951


https://doi.org/10.1080/0163853X.2013.870024


Version = 1.7.5.


APPENDICES

Appendix A. Demographic and Language Background Survey

Instructions. Thank you for agreeing to participate in this research. Before you arrive in the lab for your scheduled time, please complete this online survey. The purpose of this online survey is to collect initial data about your background that will be compared to the data you provide in the lab. Please do not complete this experiment using a mobile device or tablet. It is important you use a laptop or desktop computer with a keyboard and mouse. On the next screen you will see the informed consent for these experiments. Please read through it and indicate whether you would still like to continue with the research. On the day of the research in the lab, you will be given two informed consent forms to sign. You will receive $50 worth of Amazon.com gift cards for participating in this research. You will receive the gift card(s) on the day of your scheduled laboratory experiment.

Please enter the ID number assigned to you by the researcher in the box below.

Informed Consent

Georgia State University
Department of Applied Linguists/ESL
Informed Consent
Title of Study:
Individual differences in figurative language production and processing
Principal Investigator(s):
Scott Crossley, Georgia State University, P.I.
Stephen Skalicky, Georgia State University, Student P.I.

I. Purpose
You are invited to participate in a research study. The purpose of this study is to investigate how people produce and process figurative language, such as metaphor and sarcasm. A total of 100 participants will be recruited for this study. Participation will require approximately 90 minutes in the lab, and approximately 60 minutes completing this online survey.

II. Procedures:
The tasks for this experiment are computer mediated. If you decide to participate, you will be given instructions on how to complete the tasks. You will first take surveys measuring background and language information as well as tests of cognitive style using online surveys.

Afterwards, you will then meet with the researcher in the lab to complete the figurative language production and processing experiments and two additional surveys. The figurative language production experiment involves four separate figurative language production tasks in a random order. You will be asked to create examples of metaphors based on scenarios given to you, create sarcastic responses for situations provided for you, and also to recall a time when you used sarcasm. You will complete these tasks on a desktop computer using the mouse and keyboard. During the processing experiment, you will read 10 texts on a computer monitor that is equipped with a device that will track your eye-movements without physically touching you. For each text,
you will read it and then answer a set of comprehension questions about the text. Once you complete this for all 10 texts, the experiment will end. You will complete these tasks using a desktop computer with a mouse and keyboard.

III. Risks:
In this study, you will not be exposed to higher risks than you face during a normal day of life.

IV. Benefits:
Participants are not expected to benefit directly. The study will provide valuable research into figurative language.

V. Compensation
You will receive Amazon.com gift cards totaling $50 in value for your participant in this study. Even if you do not complete the entire experiment you will still receive the gift cards.

VI. Voluntary Participation and Withdrawal:
Participation in this research is voluntary. You do not have to be in this study. If you decide to be in the study and change your mind, you have the right to drop out at any time. You may stop participating at any time. Whatever you decide, you will not lose any benefits to which you are otherwise entitled.

VII. Confidentiality:
No identifying information will be collected. We will keep your records private to the extent allowed by law. The P.I.’s—Scott Crossley and Stephen Skalicky—will have access to the information you provide. Information may also be shared with those who make sure the study is done correctly (GSU Institutional Review Board, the Office for Human Research Protection). We will use random subject numbers rather than your name on study records. The information you provide will be stored in a locked computer with firewall protection. Your signed consent form will be stored in a locked filing cabinet in the Psycholinguistics Lab at Georgia State University for 3 years, after which it will be shredded. The findings from this study will be summarized and reported in group form. You will not be identified personally.

VIII. Contact Persons:
Stephen Skalicky (sskalicky1@gsu.edu), or Scott Crossley (scrossley@gsu.edu) at 404-413-5179 if you have questions, concerns, or complaints about this study. You can also call if think you have been harmed by the study. Call Susan Vogtner in the Georgia State University Office of Research Integrity at 404-413-3513 or svogtner1@gsu.edu if you want to talk to someone who is not part of the study team. You can talk about questions, concerns, or suggestions about the study. You can also call Susan Vogtner if you have questions or concerns about your rights in this study.

You will be presented with a print version of this informed consent to sign when you arrive for the laboratory experiments. Please indicate whether or not you would still like to participate in this research.

○ I DO NOT AGREE and no longer wish to participate.
○ I AGREE and would like to participate.
You will now complete several surveys. Some surveys are timed, so please be sure to read all the instructions thoroughly before starting. When you are ready to begin, please click continue. This entire survey should take approximately one hour.

**Demographic and Language Information**

In this survey you will be asked to provide information about yourself, including specific information about the languages that you can use. Please answer each question to the best of your ability.

- Please enter your current age in years.
- Please choose your sex.
  - Male
  - Female

Are you right-handed or left-handed?
  - I am right-handed
  - I am left-handed
  - I am ambidextrous (both right- and left-handed)

What is your current college year?
  - Freshman (1st year)
  - Sophomore (2nd year)
  - Junior (3rd year)
  - Senior (4th year)
  - Senior+ (5th year)
  - Master's Student
  - PhD Student

What country were you born in?
Imagine that this ladder represents where people stand in your society in terms of wealth and status. People with less income and status are on the bottom, whereas people with more wealth and status are closer to the top.
Based on your immediate family's wealth and status (i.e., you and your parents), what part of the ladder would you place yourself on?

2. On the bottom
3. Near the bottom
4. In the middle
5. Near the top
6. On the top

Is English your first language?

1. Yes, English is my first language.
2. No, English is not my first language.

Please list all of the languages that you know. Only include languages that you can use to conduct a meaningful conversation.

3. First Language
4. Language 1
5. Language 2
6. Language 3
7. Language 4
8. Language 5
9. Language 6
10. Language 7
11. Language 8
Please rate your proficiency for all the languages that you know

<table>
<thead>
<tr>
<th></th>
<th>Low Medium</th>
<th>Low-Medium</th>
<th>Medium</th>
<th>Medium-High</th>
<th>High</th>
<th>Native</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Language</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Language 1</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Language 2</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Language 3</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Language 4</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Language 5</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Language 6</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Language 7</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Language 8</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Language 9</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
</tbody>
</table>

(Questions marked with * were only displayed if the participant answered that their first language was not English)

*In what situations do you mainly use your first language? (Choose all that apply)

- ○ Work
- ○ School
- ○ With Friends
- ○ With Family
- ○ Alone (i.e., reading, playing video games).

*In what situations do you mainly use English? (Choose all that apply)

- ○ Work
- ○ School
- ○ With Friends
- ○ With Family
- ○ Alone (i.e., reading, playing video games).
*During a typical day in your life, do you use your first language more, less, or equally as much as you use English?
  o More
  o Less
  o Equal

*How old were you when you started learning English?
*What is the primary way you learned English?
  o Mostly through school
  o Mostly through private tutoring
  o Mostly through self-study
  o Mostly through natural exposure as a child

*In months, how long have you spent in an English-speaking country?
*Please rate your ability to use the following skills in English using a scale of 1 (low) to 5 (high)

<table>
<thead>
<tr>
<th>Skill</th>
<th>Low</th>
<th>Low-Medium</th>
<th>Medium</th>
<th>Medium-High</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Understand spoken English</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Understand written English</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Speak using English</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Write using English</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
</tbody>
</table>

Appendix B: Fluid Intelligence Task

Letter Sets: Instructions. Each problem in this test has five sets of letters with four letters in each set. Four of the sets of letters are alike in some way. You are to find the rule that makes these four sets alike. The fifth letter set is different from them and will not fit this rule. Choose the set of letters that is different.

Note: The rules will not be based on the sounds of sets of letters, the shape of letters, or whether letter combinations form words or parts of words.

Examples:
1. NOPQ     [DEFL]  ABCD  HIJK  UVWX
2. NLIK     PLIK  QLIK  [THIK]  VLIK

In the first example, DEFL is the correct choice because it does not follow alphabetical order, whereas the other four sets do. In the second example, THIK is the correct choice because it does not contain the letter “L”, whereas the other four sets do. You will encounter these rules and more during the test.

(Correct items are marked with an *)
Patterns Answer Key
1. VCVC
2. VCCV
3. CVVC
4. CCVV
5. Contains the letter “V”
6. Contains the letter “S”
7. Contains the letter “X”
8. Contains the letter “B”
9–12. Forwards alphabetical
13–16. Contains only one vowel

Appendix C. Crystallized Intelligence Test

*World Knowledge Test: Instructions.* In this test, you will answer multiple choice questions that measure your knowledge on a variety of topics. There are a total of 30 questions. Question topics will be world history, literature, and science. Please answer the questions as accurately as you possibly can. If you don't know the exact answer, make your best informed guess. Each question only has one right answer.

(Correct answers are marked with an *)
1. The edible part of the sweet potato is the
   o stem tissue.
   o root tissue.*
   o fruit.
   o seed.
2. The ancient Romans’ most significant contribution to Europe has been in the area of
   o economics.
   o poetry.
   o drama.
3. Which represents a chemical change in matter?
   1. a metal beginning to rust*
   2. water dissolving salt to form a solution
   3. water undergoing evaporation
   4. carbon dioxide undergoing sublimation

4. The poisons produced by some bacteria are called
   - antibiotics.
   - toxins.*
   - pathogens.
   - oncogenes.

5. Who is the author of the mystery fiction “Sherlock Holmes?”
   - Agatha Christie
   - Arthur Conan Doyle*
   - Edgar Allan Poe
   - James Joyce

6. Which of the following is the setting used in “The Great Gatsby?”
   - New York*
   - Boston
   - New Orleans
   - Paris

7. The writers and philosophers of the Enlightenment believed the government decisions should be based on
   - fundamental religious beliefs.
   - the concept of divine right of kings.
   - laws of nature and reason.*
   - traditional values.

8. An acidic solution could have a pH of
   - 7.
   - 10.
   - 3.*
   - 14.

9. In the Soviet Union, a negative aspect of the Cold War Era was the
   - attempt to preserve democratic ideals.
   - development of peaceful uses for modern technology.
   - development of effective means of international cooperation.
   - high cost of maintaining the arms race.*

10. Which of the following tissues produces voluntary body movements?
    - skeletal muscle*
    - cardiac muscle
    - smooth muscle
    - fibrous connective tissue

11. Which statement best describes a characteristic of the Renaissance in Europe?
    - The social structure became very rigid.
    - Creativity in the arts was encouraged.*
    - The political structure was similar to that of the Roman Empire.
Humanism decreased in importance.

12. Who is the author of “A Street Car Named Desire?”
   - William Faulkner
   - Tennessee Williams*
   - Marlon Brando
   - Arthur Miller

13. Which of these causes ocean tides on Earth?
   - the gravitational pull of the moon*
   - the revolution of the Earth around the sun
   - differences in wind speed around the Earth
   - the tilt of the Earth’s axis

14. Which of these is a compound?
   - oxygen in the air
   - liquid nitrogen
   - neon in lights
   - carbon dioxide gas*

15. Which of the following most closely captures the central theme of “Animal Farm?”
   - environmental problems
   - struggle in the animal kingdom
   - vulnerability of the socialist system to corruption*
   - problems in capitalist society

16. Which theme is most prominent in the book “Catcher in the Rye?”
   - teenagers’ identity crisis*
   - teenagers’ curiosity for adult life
   - teenagers’ desire for success
   - teenagers’ drug problem

17. The American Revolution was primarily motivated by
   - land disputes between France and England.
   - taxation without representation.*
   - the confrontation at the Alamo.
   - a decline in the price of cotton.

18. A painter who was also knowledgeable about mathematics, geology, music, and engineering was
   - Michelangelo.
   - Cellini.
   - Titian.
   - da Vinci.*

19. Blood is supplied to the heart wall by the
   - hepatic portal vein.
   - coronary arteries.*
   - auricular artery.
   - coronary veins.

20. One important result of the French Revolution was that
   - France enjoyed a lengthy period of peace and prosperity.
   - the church was restored to its former role and power in the French government.
   - political power shifted to the bourgeoisie.*
21. Which of the following best captures the theme of “Lord of the Flies?”
- the dark side of human nature without civilization and order*
- a utopia in an isolated tropical island
- the horror of a nuclear war
- love relations between poor children

22. From the following lists of states, which state was the last state to join the union?
- Tennessee
- Florida*
- New York
- Missouri

23. According to the protoplanet hypothesis, the solar system began as which of the following?
- a star
- a vacuum
- a huge cloud of dust and gas*
- a group of comets

24. The first successful colonial settlement in the United States was located in
- Salem, Massachusetts.
- Roanoke, North Carolina.
- Plymouth, Massachusetts.
- Jamestown, Virginia.*

25. Who is the author of the poem “The Raven?”
- Edgar Allan Poe*
- Walt Whitman
- Allen Ginsberg
- Arthur Rimbaud

26. Which of these has a positive charge and is found in the nucleus of an atom?
- neutrons
- protons*
- electrons
- elements

27. In “Of Mice and Men,” what is the job of the main character?
- a car salesman
- a ranch owner
- part time ranch helper*
- police officer

28. In the 1920’s and 1930’s, the rise of totalitarian governments in Germany, Italy, and Spain was largely the result of
- the success of the Communists in establishing a command economy in the Soviet Union.
- severe economic and social problems that arose in Europe after World War I.*
- the active support of the United States.
- movements demanding the return of the old monarchies.

29. Who is the author of “Great Expectations?”
- Hemmingway
- Shakespeare
- Elliot
30. What is the setting of “Les Miserables” by Victor Hugo?
   - 19th Century France*
   - 12th Century Italy
   - 12th Century France
   - 19th Century Italy
Appendix D. Abstract Thinking Test

Behavior Description Task: Instructions. Any behavior can be described in many ways. For example, one person might describe a behavior as "writing a paper," while another person might describe the same behavior as "pushing keys on the keyboard." Yet another person might describe it as "expressing thoughts." This survey focuses on your personal preferences for how a number of different behaviors should be described. Below you will find several behaviors listed. After each behavior will be two different ways in which the behavior might be identified.

For example:
Attending Class
1. sitting in a chair
2. looking at a teacher

For each of the following behaviors, your task is to choose the option that best describes the behavior for you.

<table>
<thead>
<tr>
<th>Behavior</th>
<th>Option 1</th>
<th>Option 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Making A List</td>
<td>Getting Organized</td>
<td>Writing Things Down</td>
</tr>
<tr>
<td>Reading</td>
<td>Following Lines of Print</td>
<td>Gaining Knowledge</td>
</tr>
<tr>
<td>Joining The Army</td>
<td>Helping the Nation's Defenses</td>
<td>Signing Up</td>
</tr>
<tr>
<td>Washing Clothes</td>
<td>Removing Odors From Clothes</td>
<td>Putting Clothes Into The Machine</td>
</tr>
<tr>
<td>Picking An Apple</td>
<td>Getting Something To Eat</td>
<td>Pulling An Apple Off A Branch</td>
</tr>
<tr>
<td>Chopping Down A Tree</td>
<td>Wielding An Axe</td>
<td>Getting Firewood</td>
</tr>
<tr>
<td>Measuring A Room For Carpeting</td>
<td>Getting Ready To Remodel</td>
<td>Using Measuring Tape</td>
</tr>
</tbody>
</table>
Cleaning The House
  o Showing Your Cleanliness
  o Vacuuming The Floor

Painting A Room
  o Applying Brush Strokes
  o Making The Room Look Fresh

Paying The Rent
  o Maintaining A Place To Live
  o Writing A Check

Caring For Houseplants
  o Watering Plants
  o Making The Room Look Nice

Locking A Door
  o Putting A Key In The Lock
  o Securing The House

Voting
  o Influencing The Election
  o Marking A Ballot

Climbing A Tree
  o Getting A Good View
  o Holding On To The Branches

Filling Out A Personality Test
  o Answering Questions
  o Revealing What You Are Like

Brushing Your Teeth
  o Preventing Tooth Decay
  o Moving A Brush Around In Your Mouth

Taking A Test
  o Answering Questions
  o Showing Your Knowledge

Greeting Someone
  o Saying "Hello"
  o Being Friendly
Resisting Temptation
  o Saying "no"
  o Showing Moral Courage

Eating
  o Getting Nutrition
  o Chewing And Swallowing

Growing A Garden
  o Planting Seeds
  o Getting Fresh Vegetables

Traveling By Car
  o Following A Map
  o Seeing The Countryside

Having A Cavity Filled
  o Protecting Your Teeth
  o Going To The Dentist

Talking To A Child
  o Teaching A Child Something
  o Using Simple Words

Pushing A Doorbell
  o Moving Your Finger
  o Checking If Someone Is Home

Appendix E. Need For Cognition Survey

Preferences Task: Instructions. People have different preferences for different things. For this survey, please list how much you agree or disagree with each statement using the scale provided.
  o Very Strongly Agree
  o Strongly Agree
  o Moderately Agree
  o Slightly Agree
  o Neither Agree Nor Disagree
  o Slightly Disagree
  o Moderately Disagree
  o Strongly Disagree
  o Very Strongly Disagree

1. I would prefer complex to simple problems.
2. I like to have the responsibility of handling a situation that requires a lot of thinking.
Thinking is not my idea of fun.
I would rather do something that requires little thought than something that is sure to challenge my thinking abilities.
I try to anticipate and avoid situations where there is likely a chance I will have to think in depth about something.
I find satisfaction in deliberating hard and long for hours.
I only think as hard as I have to.
I prefer to think about small, daily projects compared to long-term ones.
I like tasks that require little thought once I've learned them.
The idea of relying on thought to make my way to the top appeals to me.
I really enjoy a task that involves coming up with new solutions to problems.
Learning new ways to think doesn't excite me very much.
I prefer my life to be filled with puzzles that I must solve.
The notion of thinking abstractly is appealing to me.
I would prefer a task that is intellectual, difficult, and important to one that is somewhat important but does not require much thought.
I feel relief rather than satisfaction after completing a task that required a lot of mental effort.
It's enough for me that something gets the job done; I don't care how or why it works.
I usually end up deliberating about issues even when they do not affect me personally.

Appendix F. Language Aptitude Test

Language Analysis: Instructions. The list in the box below contains words/phrases from an imaginary language along with their English translation. Following this, there will be 14 short English sentences, each with four possible translations into the imaginary language. Based on the examples given in the box, we would like to ask you to try and work out which of the four options is the correct translation of each sentence.

<table>
<thead>
<tr>
<th>English</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>kau</td>
<td>dog</td>
</tr>
<tr>
<td>meu</td>
<td>cat</td>
</tr>
<tr>
<td>pasau meud bo</td>
<td>Our dog is chasing the cat</td>
</tr>
<tr>
<td>pa meud bo</td>
<td>The dog is chasing the cat</td>
</tr>
<tr>
<td>pa</td>
<td>we, us</td>
</tr>
<tr>
<td>xa</td>
<td>you</td>
</tr>
<tr>
<td>meud bi</td>
<td>The dog was chasing the cat</td>
</tr>
<tr>
<td>so</td>
<td>watch</td>
</tr>
<tr>
<td>ciu</td>
<td>mouse</td>
</tr>
<tr>
<td>meud bo</td>
<td>The cat was chasing the cat</td>
</tr>
<tr>
<td>pa</td>
<td>We are chasing you</td>
</tr>
<tr>
<td>bor</td>
<td>We aren’t chasing the cat</td>
</tr>
</tbody>
</table>

*Items marked with * indicate the correct answer.*

1. The dog is watching the cat.
   A. kau meud so*   B. kau meud si
   C. meu kaud so    D. meu kaud si

2. The cat was watching the mouse.
   A. meud ciu so    B. meu ciud so
   C. meud ciu si    D. meu ciud si*

3. You are watching us.
<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>4. You were chasing the dog.</td>
<td>A. ( \text{xa kaud bo} )</td>
<td>B. ( \text{pa kaud bo} )</td>
<td>C. ( \text{pa kaud bi} )</td>
</tr>
<tr>
<td>5. We were watching you.</td>
<td>A. ( \text{xapsi} )</td>
<td>B. ( \text{paxso} )</td>
<td>C. ( \text{paxsi}^* )</td>
</tr>
<tr>
<td>6. You are not watching the cat.</td>
<td>A. ( \text{xa meud bor} )</td>
<td>B. ( \text{xa meud sor}^* )</td>
<td>C. ( \text{xa meud sir} )</td>
</tr>
<tr>
<td>7. You are not chasing us.</td>
<td>A. ( \text{paxbor} )</td>
<td>B. ( \text{xapbo} )</td>
<td>C. ( \text{xapabor} )</td>
</tr>
<tr>
<td>8. We were not watching the dog.</td>
<td>A. ( \text{pa kaud sir} )</td>
<td>B. ( \text{pa kau sir} )</td>
<td>C. ( \text{pa kaud sor}^* )</td>
</tr>
<tr>
<td>9. We were not chasing you.</td>
<td>A. ( \text{xapbir} )</td>
<td>B. ( \text{paxbir}^* )</td>
<td>C. ( \text{paxbor} )</td>
</tr>
<tr>
<td>10. Your cat is chasing the mouse.</td>
<td>A. ( \text{xacu meud bo} )</td>
<td>B. ( \text{xaseu ciud bo}^* )</td>
<td>C. ( \text{meuxa ciud bo} )</td>
</tr>
<tr>
<td>11. You are not watching our dog.</td>
<td>A. ( \text{xa paseud bor} )</td>
<td>B. ( \text{xa pasaud sor}^* )</td>
<td>C. ( \text{xa pasaud so} )</td>
</tr>
<tr>
<td>12. Our mouse was not chasing the dog.</td>
<td>A. ( \text{oasiu kaud bi} )</td>
<td>B. ( \text{xasiu kaud sir} )</td>
<td>C. ( \text{xasiu kaud bi} )</td>
</tr>
<tr>
<td>13. Your mouse is chasing us.</td>
<td>A. ( \text{xa ciu pabo} )</td>
<td>B. ( \text{xasiu pbo}^* )</td>
<td>C. ( \text{xaciu pa bo} )</td>
</tr>
<tr>
<td>14. Our cat was not chasing your dog.</td>
<td>A. ( \text{pseu xasaud bir} )</td>
<td>B. ( \text{pseu xsaud bir} )</td>
<td>C. ( \text{paseu xasaud bir}^* )</td>
</tr>
</tbody>
</table>
Appendix G. Metaphor Production Task Instructions and Items

Instructions. A metaphor is a comparison between two things in order to help describe something.

For example, “her hair was like cotton”
This is a metaphorical statement that compares someone’s hair to cotton. Because cotton is something that is soft, this metaphor is implying that the person’s hair is also soft. Your job for this task is to create metaphors using the information provided. For each question, use a metaphor to describe the Topic based on information provided in the Description:. It is important to not repeat words from the Description: in your metaphor!

For example, if you see:
Topic: Her hair
Description: Soft and gentle

You would create a metaphor that describes “her hair” as soft and gentle using a metaphor, being sure to not repeat any words from the Description:. Therefore, a metaphorical answer could be:

“Her hair was like cotton” or “Her hair is like touching a cloud”

Both cotton and clouds are thought to be things that are soft. Therefore, by comparing a person’s hair to either of these things, the metaphor is implying that a person’s hair is soft.

Here is another example:
Topic: Her voice
Description: Loud and annoying

Here, you are asked to describe someone’s voice as something that is loud and annoying by using a metaphor.

Therefore, a possible answer could be:
“Her voice sounds like a train horn.”

Because train horns are things that can be loud and noisy, a metaphorical comparison to a train horn implies that this person’s voice is loud and noisy. When you are ready, you will start making metaphors for a series of different examples. Remember that you should not use any of the words in the Description: when writing your metaphor. Finally, there are no right or wrong answers, so be as creative as you can! If you still have questions, please ask the researcher now. Otherwise, please press the spacebar to begin the task.

Conventional Metaphor Production Stimuli

<table>
<thead>
<tr>
<th></th>
<th>Topic</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Some Jobs</td>
<td>Confining</td>
</tr>
<tr>
<td>2</td>
<td>Some College Lectures</td>
<td>Boring</td>
</tr>
<tr>
<td>3</td>
<td>Love</td>
<td>Difficult and Unpredictable</td>
</tr>
<tr>
<td>4</td>
<td>Certain Memories</td>
<td>Very Detailed</td>
</tr>
<tr>
<td>5</td>
<td>Her Family</td>
<td>Stability and Safety</td>
</tr>
</tbody>
</table>
Situation A
Think of the most boring high school or college class that you have ever had. What was it like to be in that class? If you need help, you might start your metaphor with:

*Being in that class was like*...

Situation B
Think of the most disgusting thing you ever ate or drank. What was it like to eat or drink it? If you need help, you might start your metaphor with:

*Eating/drinking _____ was like*...

Appendix H. Sarcasm Production Task Instructions and Items

*Instructions*. Sarcasm is a form of indirect language. When someone is being sarcastic, they mean something different than what they literally said.

For example, if someone causes a car accident, the other people involved in the accident might use the sarcastic phrase “you’re such a good driver!” to the person who caused the accident. A comment like this would be sarcastic because the other people in the accident would believe the person who caused the accident was not a good driver. Here is another example. Pretend that one friend tells another friend that they failed an important test in school. A sarcastic response could be “good job!” The response is sarcastic because the friend doesn’t actually mean that they did a good job. Rather, they did a bad job.

When you are ready, your task is to provide possible sarcastic responses for 12 different situations. For each of the 12 pictures you are about to see, take a moment to examine the situation and if anything has already been said. Then, place yourself in the role of the person with the blank speech bubble. Think of a sarcastic remark that you would say in that situation.
You are free to come up with whatever remark you like, but try to be as sarcastic as possible!

*Black and White Pictures*

1. I'm very sorry we splashed your clothing just now though we tried hard to avoid the puddle.

2. How awful! That was my mother's favorite vase.
You can't see a thing!

It's a shame my car had to break down and make you miss your train.
Desert Island Pictures
Three-Panel Comics

1. Let's meet at 9:00 to finish our report. It's due today at 4:00.
2. OK - I promise I will be there at 3:00!
3. She's still not here, and I finished the report by myself.
4. 3:45 PM
5. 3:58 PM
6. I'm here to help with the report!
Appendix I. Sarcasm and Metaphor Rubric

<table>
<thead>
<tr>
<th>Score</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Does not meet the criterion in any way</td>
</tr>
<tr>
<td>2</td>
<td>Does not meet the criterion</td>
</tr>
<tr>
<td>3</td>
<td>Almost meets the criterion but not quite</td>
</tr>
<tr>
<td>4</td>
<td>Meets the criterion but only just</td>
</tr>
<tr>
<td>5</td>
<td>Meets the criterion</td>
</tr>
<tr>
<td>6</td>
<td>Meets the criterion in every way</td>
</tr>
</tbody>
</table>

**Metaphor**

*Conceptual Distance*

The comment used a comparison between two things that are

| 1  | 2  | 3  | 4  | 5  | 6  |
conceptually distant from one another.

**Novelty**
The comment was original compared to other comments and common examples of metaphor.

**Mirth**
The author’s comment was interesting, humorous, or clever.

### Sarcasm

**Incongruity**
Based on the context, the literal meaning of the speaker’s comment clashes in some way with the actual, intended meaning.

**Novelty**
The comment was original compared to other comments and common sarcastic phrases.

**Mirth**
The author’s comment was interesting, humorous, or clever.

<table>
<thead>
<tr>
<th>Benchmark Examples (after adjudication and dimension reduction)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Metaphor Conceptual Distance</strong></td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
</tbody>
</table>
| 6 | N/A | N/A | N/A | }

N/A = Raters did not making any ratings at this level for this category. Typographical errors have been corrected for the purposes of clear presentation.
Appendix J. Satirical and Non-Satirical Texts

Instructions

The purpose of this study is to investigate the reading comprehension of online news stories. You will read ten different news stories in a random order. After you read each news story, you will be asked to summarize the author's intended meaning. You will then rate the texts for three features:

1. Sincerity (how honest was the story?)
2. Humor (how funny was the story?)
3. Positivity (how positive was the story?)

Once you are done with all ten news stories, you will complete a short survey about your reading habits. If you have any questions, please ask the researcher now. Otherwise, please click continue.

Satirical Texts

1. Consumer Spending
Report: Most Effective Marketing Technique Still Giving Out Little Versions Of Product

CAMBRIDGE, MA—Calling it far and away the best initiative businesses can undertake to boost sales and brand awareness, a report released this week by Harvard Business School has found the most effective marketing technique remains handing out little versions of products.

“Based on our research, there is simply no better way to attract and retain customers than giving away a product that is exactly like the one you’re trying to sell, only littler,” said lead researcher Neil Mukherjee, stressing that the method proved highly effective at promoting future product sales regardless of whether the complimentary items were little dish soaps, little paint swatches, or little chocolates.

“People will use up the little version quickly, because it’s so small. Then they’ll want the big version, and they’ll have to buy it, because the big version costs money. It’s pretty simple, really.”

The report also cautioned businesses against allowing any one person to take too many of the small versions of their products, as that was essentially the same as just giving away a big one for free.

2. Health Problems
Report: 750,000 Americans Die Each Year During First Attempt To Get Back In Shape

WASHINGTON—According to an alarming report published Monday in the Journal Of Applied Physiology, three quarters of a million Americans die annually during their first attempt to get back in shape.

“We found that, each year, about 225,000 out-of-shape Americans collapse and perish within the first three minutes of attempting to start jogging again, with most typically not making it to the end of their own block,” said study co-author Kathy Lyons, adding that a further 60,000 Americans who decide to improve their health by swimming laps succumb to overexertion every year and sink to the bottom of the pool and drown after just a few strokes.

“Perhaps most troubling, our data show that 60 percent of Americans who commit to regaining their former level of physical fitness by hiring a personal trainer at a gym drop dead
almost instantly during their first abdominal crunch. And another 30 percent die while stretching before they even begin their workout.”

The study further noted that, of the Americans who managed to survive their first attempt at exercise in years, nine in 10 suffered from debilitating pain for the rest of their lives.

3. Stress
Study: Average Person’s Enjoyment Of Vacation Drops 36% For Each Additional Family Member Present
COLLEGE PARK, MD—Finding consistent results across all types and durations of vacation, from multi-week cruises to brief weekends spent camping, a report released Monday by the University of Maryland revealed that the average person’s enjoyment of their time away from work or school drops 36 percent for each additional family member present.

“We studied more than 3,000 vacationing subjects and found that their overall levels of relaxation and satisfaction dropped by more than a third for every parent, sibling, child, aunt, or uncle accompanying them,” said lead researcher Yvonne Ryan, warning that individuals who spent their vacation with three or more family members effectively neutralized any potential enjoyment they could have expected from such a break.

“While 36 percent represents the average amount that a relative reduces one’s enjoyment of a vacation, we found that certain family members have far greater adverse impacts on one’s pleasure, with elderly relatives and most in-laws reducing enjoyment levels by 50 percent or more.”

Ryan added, however, that such negative effects could largely be counteracted by a threefold increase in one’s normal alcohol consumption for each additional family member present.

4. Divorce
Study: More Couples Delaying Divorce Until Kids Old Enough To Remember Every Painful Detail
CHICAGO—In a new study published this week in The American Journal Of Sociology, researchers reported that parents throughout the United States are increasingly opting to delay divorce until their children are old enough to remember each and every traumatizing detail.

“What we found is that more and more couples are deliberately holding off on dissolving their unhappy marriages until their children are 9 or 10, the approximate age at which they’re cognitively capable of retaining every unbearably painful moment,” said study co-author Anna Dasgupta, adding that children at that stage of maturation will generally have the ability to recall for the rest of their lives the moment their dad told them he was moving out.

“And by not rushing the announcement, parents ensure that their children have accumulated at least some memories of happier times, such as Christmases and birthday parties when the whole family was together, which they will use as sources of self-torment in the broken homes of their adolescence.”

The study also noted that by postponing their divorce, parents helped ensure their children had sufficiently developed their sense of agency enough to blame themselves for everything.

5. American GDP/Manufacturing
Report: America Still World Leader In Manufacturing Excuses
NEW BRUNSWICK, NJ—Revealing that Americans still excel in assembling all types of justifications, a report released Monday by researchers at Rutgers University confirmed that the United States remains the world’s unrivaled leader in manufacturing excuses.

“Our data shows that the American people are able to churn out millions of excuses every day, a rate five times faster than that of our closest international competitor,” said author Tom McCullough, who noted that the country continues to dominate in every excuse-manufacturing subsector, with Americans mass-producing rationalizations for everything from why they didn’t finish college to which of their colleagues should be blamed for them not receiving a promotion.

“Whether U.S. citizens are trying to get out of work, an upcoming party, or even a relationship, American excuses are still the envy of the industrialized world. Few countries can produce the same quality or quantity of pretexts, alibis, or half-assed explanations.”

McCullough added that the report probably could have been far better had his team been given more time and resources.

6. Gender Inequality
Study Finds Girls Outperforming Future Employers In School
NASHVILLE, TN—The results of a comprehensive multiyear study charting the educational achievement of children throughout the United States were released Friday, revealing that the nation’s girls are increasingly outperforming their future employers in all subjects.

“We looked at test scores from all 50 states and found that, across every demographic group, girls are consistently outscoring those who will someday have the power to hire and fire them,” said the study’s lead author, Jennifer Malone, of Vanderbilt University, who noted that the gap between female students and those who will hold 86 percent of top executive positions at the companies where they work emerges early in elementary school and continues to grow throughout high school, college, and graduate school.

“For years, girls have performed better than their future bosses in areas like writing and reading comprehension, but more recently, they’ve started to surpass them in STEM subjects as well. At the same time, those individuals who will one day pay their female classmates a fraction of the industry standard have fallen further behind.”

Malone added that if current trends in education continue, women will soon outnumber nearly all future tech workers, financial analysts, and government leaders at the nation’s universities by a two-to-one margin.

7. Vacation Habits
Going To Tops Of Things Still Favored By Nation's Tourists
NEW YORK—According to a report released Monday by the American Tourism Society, going to the tops of things is still the preferred activity among the nation's tourists.

"Although driving past things and swimming in things have both grown in popularity over the last decade, going to the tops of things still surpasses both by nearly 30 percent," said ATS president Kimberly Davis, who was careful to point out that the photographing of things was not included in the report, since the near constant occurrence of this activity makes its frequency impossible to calculate.

"In 2008, tourists remained committed to standing in long lines at the bottoms of things, paying upwards of $20 to gain access to the tops of those things, and then staring at other smaller, more distant things for a few minutes before descending, often to have funny pictures of themselves drawn incorporating the things in the background."
Davis added that, perhaps as a consequence of the declining economy, the purchasing of miniature representations of the things that tourists enjoy going to the tops of has dropped by 14 percent.

8. Civic Engagement
Report: Majority Of Nation’s Civic Engagement Centered Around Oppressing Other People
WASHINGTON—Providing insight into how American citizens participate in the political process, a report released Tuesday by the Pew Research Center revealed that the vast majority of civic engagement in the U.S. centers around oppressing other people.

“We found that Americans are fairly active in making their voices heard and engaging in the public sphere, typically in an effort to restrict the rights of those who look or behave differently from them,” said lead researcher Dana McNeil, who added that well over two-thirds of the nation’s rallies, petitions, and letter-writing campaigns were aimed at making life worse for a portion of the population of which the participants were not a part.

“Taking away the rights of other people was by far the most common subject of calls placed to congressional representatives, as well as the single greatest factor in predicting voter turnout. Indeed, when Americans join a citizens group or volunteer their time for a cause, more often than not, it’s to shout angrily at a specific group of people and call for their persecution.”

McNeil added that, at the very least, it was encouraging to see so many Americans using the democratic process to make lasting change.

9. Global Habitat
Alarming Report Finds Only 6% Of Earth’s Surface Indoors
LAWRENCE, KS—Drawing attention to the distressing prevalence of outside areas on the planet, researchers at the University of Kansas released an alarming report Monday revealing that a mere 6 percent of the Earth’s surface is actually indoors.

“Our team found that a shocking 185 million square miles across the globe are fully out-of-doors in the open air,” said lead researcher Priya Chatterjee, adding that on all seven continents, there exist vast stretches of forests, mountain ranges, and bodies of water that are completely exposed to the elements, without so much as a single room where an individual might go inside and relax.

“It is quite disconcerting to realize that nearly all of the Earth’s surface is outside, unprotected from meteorological conditions. The reality is that the vast majority of the planet is at risk of getting cold, windblown, or totally soaked.”

The report also concluded that more research was necessary to determine whether certain as-yet-unclassified locations, such as gazebos, courtyards, and breezeways, should be categorized as indoors or outdoors.

10. Political Habits
Report: More Americans Turning To Louder Sources For Their News
PHILADELPHIA—Saying that the trend signals a major shift in the media landscape, a report issued Thursday by the University of Pennsylvania revealed that a growing number of Americans are turning to louder sources for their news.

“Over the past 10 years, media-consumption habits have changed markedly as more people eschew traditional news outlets in favor of sources that report the latest stories at a far higher volume,” said the study’s lead author Emily Harding, noting that audiences for television
and radio programs that deliver news in a reasonable, non-damaging tone of voice have fallen by nearly 40 percent since last decade.

“Respondents were shown one well-reported clip of environmentally sustainable practices on a large dairy farm, followed by a second, poorly-reported but eardrum-shattering clip about how factory farms have ripped open the ozone and will lead to our impending doom. Given the thunderous din of the latter, it wasn’t surprising which one they trusted more. It seems, for a steadily growing segment of the population, that heavily blaring, borderline deafening media outlets are their only source of news.”

Harding added that if current trends continue, the U.S. news cycle is likely to reach tornado siren levels exceeding 125 decibels by 2020.

Non-Satirical Texts

1. Consumer Spending
Study Finds People Willing To Pay More For New Biofuels
KANSAS, MO–When it comes to second generation biofuels, a study published in November in the journal Energy Economics from researchers at Washington State University shows that consumers are willing to pay approximately 11 percent more than the cost of conventional fuel.

"We were surprised the premium was that significant," said Jill McCluskey, WSU professor in the School of Economic Sciences, and adding that they wanted to study people in different regions of the country, to make sure they weren't just getting a local result, and people in all three cities studied said they would pay more for these products.

"This new biofuel doesn't exist commercially yet, so we have to do these surveys to make sure there's a potential market for it, and this shows there clearly is a market. People were concerned that the new fuel may put their car at risk, but they also saw the added benefit to the environment."

Before they were surveyed, half of the participants were given information about second generation biofuels. Those participants were more willing to pay a greater premium, which suggests that marketing the benefits of the new biofuels would improve consumers’ perceptions.

2. Health Problems
Report: Heart Attack Risk Doubled For People With Less Education
BOSTON, MA–People who leave school without a school certificate are more than twice as likely to have a heart attack as those with a university degree, according to a groundbreaking new study published in the Journal of Medicine investigating age and heart disease.

"The lower your education, the more likely you are to have a heart attack or a stroke - that's the disturbing but clear finding from our research," said lead researcher Dr. Rosemary Korda, who elaborated that a similar pattern of inequality existed between household income and cardiovascular disease events.

"What these differences in heart attack rates between more and less disadvantaged groups show us is just how much heart disease in the population can be prevented with proper research. Studies are continuing in this area to better understand what is driving these socioeconomic differences in order to help these people."

These findings demonstrated the value of the study as an important resource making it possible for researchers to investigate big questions in large numbers of people, and to get faster answers that are useful for policy makers in higher levels of government.
3. Stress
Study Suggests High Stress Levels May Delay Women Getting Pregnant

NEW HAMPSHIRE–Healthy women trying for a baby may have reduced chances of conceiving a child in any month if they are stressed, the results of a study by researchers at Oxford University and the US National Institutes of Health suggests.

“This is the first study to find that a biological measure of stress is associated with a woman's chances of becoming pregnant that month,” explained lead author Dr. Cecilia Pyper, adding that females between ages 18 to 40 with high levels of a chemical marker for stress were less likely to succeed in conceiving.

“We want to understand the things that affect the chances of conception, and many younger couples are very keen to know what they should do to improve their chances of having and raising a healthy baby.”

This research is part of a larger experiment, which has recruited 1500 young prospective mothers in the UK trying to become pregnant and aims to see whether aids such as a fertility monitor could also help them. It is also looking for the effect of other negative factors like smoking, alcohol and caffeine on chances of successful birth rates.

4. Divorce
Report: More Sex Partners Before Marriage Doesn't Necessarily Lead To Divorce

CHICAGO–Common sense says having multiple sexual companions prior to marriage leads to less happy marriages and increases the odds of divorce, but a report published this week in the Journal of Sociology suggests some of the strongest predictors of divorce in years gone by no longer matter as much as they once did.

"Overall, American women are far more likely to have had multiple premarital sex with multiple different people in recent years," lead researcher Nicholas Wolfinger said, pointing out that as premarital sex became more acceptable, its negative effects on marital stability decreased, and that Americans became more accepting of nonmarital intercourse.

"Perhaps it is not unexpected that having many partners increases the odds of divorce. The greater surprise is that this only holds true in recent years; previously, women with two partners prior to marriage had the highest divorce rates."

Also noteworthy is the decline in the proportion of women who get married having had only one previous partner. As late as the 1980s, over half of new brides were virgins or had had only one sex partner.

5. American GDP/Manufacturing
Report: Diabetes, Heart Disease, And Back Pain Dominate US Health Care Spending

TEXAS–Just 20 conditions make up more than half of all spending on health care in the United States, according to a new comprehensive financial analysis that examines spending by diseases and injuries published in the Journal of the American Medical Association.

"While it is well known that the US spends more than any other nation on health care, very little is known about what diseases drive that spending," said Dr. Joseph Dieleman, lead author of the paper, who estimates that approximately $300 billion in costs remain unaccounted for, indicating total personal health care costs in the US reached $2.4 trillion in 2013.

"IHME is trying to fill the information gap so that decision-makers in the public and private sectors can understand the spending landscape, and plan and allocate health resources more effectively."
While the majority of those costs were associated with non-communicable diseases, the top infectious disease category was respiratory infections. Public health education and advocacy initiatives, such as anti-tobacco and cancer awareness campaigns, totaled an estimated $77.9 billion in 2013, less than 3% of total health spending.

6. Gender Inequality
Report: Men More Likely To Be Seen As 'Creative Thinkers'
CALIFORNIA–People tend to associate creative skill with stereotypical masculine qualities, according to a new paper published in Psychological Science suggesting that the work and achievements of men tend to be assessed as more creative than similar work and achievements produced by women.

"Our research shows that views about what it takes to 'think creatively' overlap substantially with the unique content of male stereotypes, creating systematic bias in the way that men and women's originality is evaluated," says lead researcher Devon Proudfoot, saying that this could lead people to view creative problem solving as an ability more common among men than women.

"In suggesting that females are less likely than males to have their innovative thoughts recognized, our study points to a unique reason why women may be passed over for leadership job positions within many different industries when climbing the ladder."

Gender bias in creativity judgments may affect tangible economic outcomes for both genders in the workplace, as previous experiments have shown that people in relatively higher power positions are more likely to rely on stereotypes when forming decisions about people.

7. Vacation Habits
Report: Recession Drives Down U.S. National Park Visitation
CLEVELAND, OH–A national recession doesn't just affect Americans' wallets. It also impacts their travel to national parks. According to a new study published by the U.S. Department of Interior, the significant decrease in national park visitation is linked to the weakened United States economy.

"Economic downturns come and go, but the impact on resource conservation and park visitor management could last longer," said first author Neelam Poudyal, stating that the findings could help park managers plan ahead for revenue shortfall and a decrease in park visitation, particularly as the economic forecast for the coming years remains bleak.

"While the recession affected all types of park visitation, trips involving longer timeframes and higher costs such as overnight visits and backcountry camping are likely to be hit harder than shorter and relatively cheaper trips such as day visits and weekend camping."

Researchers have linked many factors to a decline in park visitation, including natural disasters and the growing reliance on indoors technology for recreation, but this study was the first to examine economic factors linked to park visitation during peak and non-peak tourism times.

8. Civic Engagement
Report: Civic Engagement May Strengthen Brain And Improve Memory
OLYMPIA, WA–Instead of shrinking as expected, the memory center in the brains of seniors maintained their size and, in men, grew modestly after two years in a program that involved them in civic engagement, new Johns Hopkins Bloomberg School of Public Health research suggests.
"Someone once said to me that being in this program removed the cobwebs from her brain and this study shows that is exactly what is happening," says study leader Michelle Carlson, adding that the participants were followed for two years, which in this case was long enough to see changes that wouldn't have been detected after just one year.

"By helping others, participants are helping themselves in ways beyond just feeding their souls. They are helping their brains. The brain shrinks as part of aging, but with this program we appear to have stopped that shrinkage and are reversing part of the aging process."

It's not entirely clear which elements of the program account for the improved memory function and increased brain volumes because the program increased involvement in many different kinds of activities that retired people may not have engaged in otherwise.

9. Global Habitat
Report: Global Habitat Loss Still Rampant Across Much Of Earth
FLAGSTAFF, AZ–Habitat destruction still far outstrips habitat protection across many parts of the planet, according to a new study released this week by the Journal of Ecology which revealed more than half the planet could now be classified as completely converted to human-dominated land use.

"An area of 4.5 million square kilometres, or about two thirds the size of Australia, has been converted to human-dominated land use in the past two decades alone," lead researcher James Watson said, adding that there had been considerable gains in global efforts to increase the size of protected areas, but alarming levels of habitat loss persisted.

"As a consequence of past and recent habitat loss, almost half of the world's regions now must be classified at very high risk, as 25 times more land has been converted than protected."

The researchers identified 41 regions across 45 nations that are in a 'crisis state', where humans have converted more than 10 per cent of the little remaining habitat in the past two decades.

10. Political Habits
Report: Americans Are More Politically Independent, More Polarized Than Ever
ATLANTA, GA–Stating that many older citizen’s political views contrast significantly when compared to people from previous generations, a new study published in the Journal of Political Science has found that approximately half (46 percent) of adult Americans today now identify as political independents.

"Americans, especially young people, are abandoning the two major parties to declare themselves independent, demonstrating large groups such as major parties are less popular," the study’s lead author Jean Twenge, adding that those who do claim devotion to the Democratic or Republican party in the United States are more similar in their views when compared to independents.

"It may be that the definition of what they consider conservative is changing. Overall, younger Americans may not be as reliably liberal and Democrat as many had predicted, especially as they are likely to grow more politically independent as they age."

In recent years, there has also been an rise in political conservatism among young people. High school seniors in 2010 were 38 percent more likely to identify as political conservatives than their age-matched peers in the 1970s.
Appendix K. Genre Background Survey

1. How many days per week do you read a print newspaper or magazine?
   - 0
   - 1
   - 2
   - 3
   - 4
   - 5
   - 6
   - 7

*If answer more than 0 for Question 1*

1. When you do read a print newspaper, how long do you typically spend reading it?
   - Less than 15 minutes
   - Between 15 and 30 minutes
   - Between 30 and 45 minutes
   - Between 45 and 60 minutes
   - More than 60 minutes

2. How many days per week do you read news online directly from online news sources, (i.e. not through social media, etc.), using your phone or computer?
   - 0
   - 1
   - 2
   - 3
   - 4
   - 5
   - 6
   - 7

*If answer more than 0 for Question 2*

2.5 When you read news online directly from online news sources, how long do you typically spend reading it?
   - Less than 15 minutes
   - Between 15 and 30 minutes
   - Between 30 and 45 minutes
   - Between 45 and 60 minutes
   - More than 60 minutes

3. How many days per week do you use social media (e.g., Facebook, Twitter) to learn about news from friends, family, and other social media users?
   - 0
   - 1
   - 2
   - 3
   - 4
If answer more than 0 for Question 3
3.5. When you learn about news from social media, how long do you spend reading about it?
- Less than 15 minutes
- Between 15 and 30 minutes
- Between 30 and 45 minutes
- Between 45 and 60 minutes
- More than 60 minutes

4. How many days per week do you watch the television for news?
- 0
- 1
- 2
- 3
- 4
- 5
- 6
- 7

If answer more than 0 for Question 4
4.5. When you watch the television for news, how long do you spend watching it?
- Less than 15 minutes
- Between 15 and 30 minutes
- Between 30 and 45 minutes
- Between 45 and 60 minutes
- More than 60 minutes

5. How familiar are you with satirical television news programs, such as *The Colbert Report*, *The Daily Show*, or *Weekend Update* on *Saturday Night Live*?
- Not familiar at all
- Somewhat unfamiliar
- Somewhat familiar
- Very familiar

6. How familiar are you with satirical newspapers such as *The Onion* or *The Borowitz Report*?
- Not familiar at all
- Somewhat unfamiliar
- Somewhat familiar
- Very familiar

**Appendix L. Study 3 Online Informed Consent and Task Instructions.**

**Informed Consent.** We are researchers at Georgia State University. We are interested in measuring the creativity of human responses to certain tasks. We are also interested in how people judge their levels of creativity compared to others. The data you are providing will be used for research purposes. Participation in this research is voluntary. The data collected during your participation in this task will not be saved with any of your personal information (because you are anonymous). No information collected during this study will make it possible to identify you. Your data will remain confidential.
For this research task, you will be first asked to answer a short demographic survey. You will then complete a task where you pretend to be a television news host. Your job will be to provide a creative reaction to two separate news stories. You will first read a news story, and then you will type your answer to the story. Afterwards, you will see a random answer to the same story provided by the researchers. You will rate whether you feel your answer was more creative, less creative, or equally as creative as that answer. You will repeat this process for the second news story. After completing the second news story, the task will be complete.

You will be paid $1.00 for your work. If you have questions about this research, please contact Stephen Skalicky at sskalicky1@gsu.edu. If you agree to participate, click I AGREE. You will immediately begin the task.

First instructions screen

The purpose of this task is to come up with creative reactions to two real news events. For the purposes of this task, creativity is defined as a novel yet effective solution to a problem. You will first be presented with one of the news stories and then asked to submit your response. Afterwards, you will see an answer submitted by another participant and asked to rate whether you believe your answer was more creative, less creative, or equally creative compared to that answer. You will then repeat this process for the second news story. When you are ready, please click continue.

Second instructions screen

Your new job as a 24-hour news channel host has just started! Being a member of a 24-hour cable news network means you’ll need to come up with an interesting and creative reactions to different news headlines as they occur throughout the day. In order to get a feel for what this is like, we’ll have you react to some news stories right now, and then compare your reactions to those of other people. Since your job ratings will depend on how many viewers you can keep watching the channel, it is imperative to be as creative as possible when reacting to the headlines! Your reactions should be something you would say during live television, and be about as long as a single sentence. In other words, please only type the language you would say, exactly as you would say it.

Please do not type what you would do or think – only type what you would say.

Prompts as participants saw them

Octopus Escapes New Zealand Aquarium
Inky the Octopus is believed to have escaped the national aquarium of New Zealand by opening the lid of his tank, slithering across the floor, then squeezing through a 5-inch-wide drainage pipe and out to sea. What do you think?

Remember, type a response exactly as you would say it live on television. Your purpose is to be as entertaining and creative as possible in approximately a single sentence - keep those people wanting to watch you! Please type your response below.

Woman Sues Starbucks Over Amount Of Ice In Drinks
A Chicago woman has filed a $5 million lawsuit against the Starbucks Corporation claiming false advertising of 24-ounce iced drinks that allegedly only contain 14 ounces of liquid. What do you think?

Remember, type a response exactly as you would say it live on television. Your purpose is to be as interesting and creative as possible in approximately a single sentence - keep those people wanting to watch you! Please type your response below.

Example of rating screen

Now compare your answer to this answer. Do you think your answer was more creative, less creative, or equally as creative?

Your Answer:

"Example for dissertation"

Their Answer:

"The sea will be much better than his tank."

- My answer was more creative
- My answer was equally as creative
- My answer was less creative