A Psychometric Evaluation of Emotion Detection Lexicons: Construct Validity and Measurement Differences

Tara Lucile Valladares

Purcellville, Virginia

MA, University of Virginia, 2019

BA, University of Virginia, 2016

A Dissertation Presented to

the Graduate Faculty of the University of Virginia

in Candidacy for the Degree of Doctor of Philosophy

Department of Psychology

University of Virginia

December, 2022

Committee Members:

Karen M. Schmidt, PhD (Chair)

Steven M. Boker, PhD

Gregory J. Gerling, PhD

Hudson F. Golino, PhD

Abstract

Emotion detection (ED) encompasses a wide variety of tools and techniques to automatically extract emotion content from text. ED has become increasingly popular in psychology, linguistics, the data sciences, and many other fields, however the construct validity of ED methods has received minimal attention. General purpose emotion lexicons are one common ED tool that contain predetermined word-emotion associations. Though ED lexicons measure psychological constructs, many are justified with scant psychological theory. Further, different methods of constructing lexicons may also lead to differences in their word-emotion associations. The measurement similarities of different lexicons is currently unclear, which presents issues for researchers who are concerned with construct validity or who wish to compare results across multiple studies. This dissertation used a novel application of item response models directly on emotion lexicons to understand their similarities and differences. A dual confirmatory and exploratory psychometric approach was taken to compare how lexicons are typically used and how their categories actually inter- and intra-relate. In the confirmatory approach, a strict hypothetical structure was imposed on the lexicons where same-named discrete emotion variables were forced onto the same factors. The final confirmatory model fit poorly. In the exploratory approach, lexicon variables were free to associate. The final exploratory model indicated that while emotion lexicons generally had similar word-emotion associations for the same discrete emotions, there were significant distinctions. Limitations and future directions are discussed.

Contents

Acknowledgements

I am eternally grateful to everyone who came alongside me throughout my graduate career. First, praise be to the Lord, for his steadfast grace and mercy and the blessings he provided in finishing this dissertation. I also could not have done this without my husband, Dagoberto, and the unending support he provided. Our daughter Lucy was born in the midst of writing this dissertation and quickly became the light of our lives. Though little, Lucy taught me that inspiration cannot strike at two in the morning if you sleep through it. I would also like to thank my adviser Karen, for her patience, teaching, and guidance, as well as all my colleagues for their wisdom and friendship. And finally, thank you to my sister Carolyn and all the friends and family who cheered me on along the way. It truly takes a village.

Author Note

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. 1842490. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

A Psychometric Evaluation of Emotion Detection Lexicons: Construct Validity and Measurement Differences

Emotion is so fundamental to human communication that how we express it is quite literally coded into our genes (Ekman, 1992a; Jackson et al., 2019). In conversation, emotion is simultaneously displayed through word choice, facial expression, body language, and tone of voice. Though emotion is also critical to written communication, it is much more difficult to measure as there are fewer cues available to rely on. This presents a challenge to the field of psychology as accurate measurement is a cornerstone of quality research.

The focus of this dissertation is the validity of measuring emotion in text using tools from emotion detection. Emotion detection (ED) refers to the collection of theories, methods, and techniques that are used to measure the presence and degree of emotion expression within text. Despite measuring psychological constructs, ED processes have largely been developed outside of psychology by researchers in data science, engineering, and related fields (Acheampong et al., 2020). These external methods have in turn been adopted by psychological researchers, but there are still many unanswered questions about the validity of the information these methods produce.

One of the most popular ED tools used in psychological research is the emotion lexicon. Emotion lexicons are collections of words and their associations with discrete emotion constructs like *Anger* and *Joy* or dimensions like valence and dominance. Table 1.1 shows an excerpt from the DepecheMood++ lexicon for illustration; shown are three words and their associations with eight discrete emotions. Emotion lexicons can be used in both simple, low-computation analyses and in larger, more complex pipelines. Because emotion lexicons are often the only tool used by a study to assign emotion to text, the quality of the lexicon is fundamental to the quality of the analysis.

The "quality" of an emotion lexicon ultimately boils down to its construct validity. Construct validity refers to the degree that a tool accurately measures the underlying

Table 1.1
An Excerpt from the DepecheMood++ Lexicon

	Afraid	Amused	Angry	Annoyed	Don't Care	Happy	Inspired	Sad
lawn	0.06	0.33	0.10	0.07	0.05	0.10	0.18	0.11
rotunda	0.33	0.06	0.04	0.19	0.15	0.12	0.08	0.03
virginia	0.15	0.17	0.11	0.12	0.14	0.07	0.10	0.13

construct it claims to measure (Cronbach & Meehl, 1955; Embretson, 1983). Construct validity is difficult to ascertain because psychological constructs are often latent and not directly measurable. For emotion lexicons, construct validity is the degree to which the lexicon's evaluation of a text accurately reflects the text's emotional content. That is, if a lexicon labels a text as "Angry", is the text truly full of *Anger*? Those familiar with text or emotion research will be quick to ask what the meaning of *Anger* is, what does it mean to have lots of *Anger*, and if a text is *Angry* when the writer was conveying *Anger* or if the reader reacts with *Anger*. Each question here relates to construct validity, which can rarely be ascertained based on the results of one study and is not a single monolith. A more tangible question will be explored in this dissertation: Do different general purpose emotion lexicons measure the constructs they claim to measure in the same way? For if not, the construct validity of each lexicon is called into question.

For ED to be most useful for psychology, ED techniques must to be evaluated through the field's theoretical and statistical perspective, including an evaluation of the psychological constructs ED claims to measure. In this dissertation, I use psychometric techniques to model and describe the measurement qualities of four different popular emotion lexicons. This dissertation will begin with a discussion of psychological theories of emotion, a brief background on emotion detection and emotion lexicons, and how the two fields integrate. Following, I will describe how item response models were used to compare and contrast the emotion lexicons and conclude with a discussion on their uses and limitations.

Prior to this, I will lay out how homonymous terms will be distinguished from each

other for clarity. A word like "joy", for example, can represent three separate entities in this dissertation: words, concepts, or variables within a lexicon. If a word represents the literal word itself, it will be denoted with with quotation marks (e.g., the word "joy"). Conceptual emotions or constructs will be denoted with capitalized italics (e.g., *Joy* from Ekman's theory of emotions). When the word represents a labeled measurement of emotion found within a lexicon, it will be indicated with capitalized styling (e.g., the variable Joy from the EmoSenticNet lexicon).

2 Emotion

For the layperson, emotion hardly requires a definition; learning to label emotions is so natural that it is part of typical childhood development (Wellman et al., 1995). However, the perceived intuitiveness of emotion makes it difficult to scientifically define and divide into its components (Fiske, 2020), and contributes to the continuing conflict in psychology to define a clear framework (Barrett, 2006a). The combination of ease of intuition and the lack of a clear scientific consensus leaves interpretive room for ED researchers. Theories of emotion dictate how emotion is operationalized and divided into measurable parts. In this section, I will summarize psychological theories of emotion relevant to ED, as well as how emotion is expressed in text. The issue of construct validity underlies many of the concerns herein as how emotion is defined dictates the validity of its measurement.

2.1 Theories of Emotion

There is an extensive body of research on human emotion describing how it evolved, how it exists bodily and is physically expressed, and how its function impacts everyday life. It is generally accepted that some emotions are simple and others are complex combinations or derivatives (Ekman, 2016). Emotions can be described dimensionally by valence and arousal, among others (Jackson et al., 2019; Warriner et al., 2013), and some commonalities can be found across cultures based on the similarities of facial expressions (Ekman, 1992b; Jack et al., 2016). However, there is not a scientific consensus on how many emotions there are, if any are universal, or how they should be structured and divided scientifically (Ekman, 2016; Moors, 2017). It is not unusual for studies to identify more than twenty discrete emotions, and large crowd-sourced datasets can lead to even more fine-grained divisions (Cordaro et al., 2018; Cowen & Keltner, 2017). But what one study would call an emotion, another would not - for example, see sexual desire as an emotion in Cowen & Keltner

(2017) versus a state in Royzman & Sabini (2001). Further, other lines of research would reject theories of discrete, individual emotions entirely, and replace them with structural dimensions (Posner et al., 2005). Psychological theories of emotion typically fall into either the discrete or dimensional camp. This dissertation focuses on discrete emotion lexicons, however dimensional frameworks will be also be discussed because they provide important context and background for interpretation.

Discrete vs. Dimensional Frameworks

Discrete emotions are labeled entities of emotion such as *Anger*, *Sadness*, or *Joy*. They may also referred to as emotion labels, emotion categories, vernacular emotions, or emotion sets/subsets (Barrett, 2006b; Moors, 2017). Emotions are generally thought about in everyday life as discrete entities; unsurprisingly then, they are a hallmark of classic theories of emotion (Barrett, 2017; Fiske, 2020) and are frequently treated as "natural kinds" by researchers (Barrett, 2006a). That is, discrete emotions are often conceptualized to exist in reality as fundamental, non-arbitrary natural groupings and therefore are not merely created by cultural convention or the innate human propensity for categorization. However, though discrete emotions are ubiquitous in everyday thought, stringent divisions between them are typically not found when they are measured somaticaly, neuronally, behaviorally, through self-report, or via facial expressions (Cacioppo et al., 2000; Jack et al., 2016). Even the term "emotion" itself originally arose from linguistic ease of use and cultural beliefs rather than through specific scientific inquiry (Fiske, 2020).

In contrast to discrete emotions are emotion dimensions. Emotion dimensions describe features of emotions that vary along a continuum. As mentioned, two of the most commonly used dimensions are valence and arousal (Fontaine et al., 2007; Jackson et al., 2019; Russell, 1980; Warriner et al., 2013). Older dimensional frameworks often perceived valence and arousal as sufficient to organize most emotional experiences (Posner et al., 2005; Russell,

1980), while other newer frameworks reject this simplicity (Fontaine et al., 2007). Additional dimensions such as dominance, power, and novelty have been proposed (Fontaine et al., 2007; Salinas et al., 2015), as well as sociality, certainty, and approach-avoidance (Jackson et al., 2019).

Discrete emotions may be described by emotion dimensions. For example, though prototypical *Anger* is seen as a high arousal, high dominance emotion (Russell & Mehrabian, 1977), different flavors of *Anger* like irritation or rage can be specifically described within a dimensional model. However, some dimensional advocates reject discrete emotions as existing as useful entities, in the sense that their categorization and definition is relative and inconsistent (Posner et al., 2005).

Basic Emotions

Basic emotions theories are the most frequently cited framework in ED research. Basic emotions are discrete emotions that are proposed to have an evolutionarily derived, biological basis and are thus universally expressed and experienced in the same way in all human societies across the globe[1] (Ekman, 1992b). They are the most fundamental and simple of emotions, the most natural of "natural kinds". The existence of basic emotions and their association with specific facial expressions was first proposed by Charles Darwin and still frequently appears in psychological theory (Darwin, 1872; Ekman, 2016; Jack et al., 2016). The two theories most frequently referenced in ED are those of Ekman and Plutchik Ekman et al. (1969).

However, which and whether emotions are basic is not fully settled. There have been persistent questions on which, if any, emotions are basic or universal since their proposal

[1]Not all psychological research uses "basic" to denote universality. Instead, "basic" may simply be used in contrast to complex or higher-order emotions. However, Ekman and Plutchik are the most common theorists cited in ED and both of their theories purport to be universal. There is also near perfect overlap between sets of evolutionary "basic" emotion categories and "basic" emotions under the more relaxed definition. I will therefore be using the universal/biological definition of "basic".

(Ortony & Turner, 1990). Many of the arguments against emotions as "natural kinds" arose through criticisms on basic emotions (Barrett, 2006a). In 2016, Ekman surveyed 248 published emotion researchers on their beliefs about the universality of emotions across cultures[2]. Those surveyed responded that there were five emotions that are empirically indicated as universal: *Anger* (endorsed by 91%), *Fear* (90%), *Disgust* (86%), *Sadness* (80%), and *Happiness* (76%). *Surprise*, *Shame* and *Embarrassment* were only endorsed by 40-50% of those surveyed. All other emotions were endorsed by 37% or fewer researchers. Interestingly enough, only 32% of researchers believed that there was enough research to support *Love* as a universal basic emotion.

As is evident from Ekman (2016), even for emotions with the most consensus, at least 10% of those surveyed dissented. Recent papers have pointed out that accuracy rates of cross-cultural facial expression recognition range from 100% to only 30% (Jack, 2013). Many studies of basic emotions use forced choice paradigms which may artificially inflate agreement on the relationship between facial expressions and identified emotions (DiGirolamo & Russell, 2017; Gendron et al., 2018; Jack et al., 2014). And finally, theories of basic emotions generally go hand in hand with discrete theories of emotion. As mentioned, there are significant empirical issues with the discrete organization of emotion (Barrett, 2017; Cacioppo et al., 2000).

However, emotion detection bypasses many of these issues; ED's primary concern is the technical measurement of emotion, not its conceptual investigation. To measure emotion, though, some level of operationalization is required. ED research typically uses one of three common theories as a basis of measurement: Ekman's, Plutchik's, or the Valence, Arousal, and Dominance model. By no coincidence, these are also some of the most well-known models in psychology. All three of these theories provide straightforward and intuitive discrete emotions or dimensions to measure. Because ED research largely cites Ekman's

[2]As also mentioned by Ekman, this is clearly not a completely representative sample of all the beliefs held by emotion researchers. However, there is much to be gained by understanding how researchers think about their research, even if done by one of the lead champions of discrete, basic emotions.

and Plutchik's *original* theories at the time of publication, that is what will be focused on. The following descriptions are by no means exhaustive of the various frameworks of emotion in psychology or the numerous researchers who have contributed to them.

Ekman's Basic Emotions

Beginning in the 1960's, Paul Ekman proposed that humans possessed a set of six universal basic emotions: *Anger*, *Fear*, *Sadness*, *Disgust*, *Joy*, and *Surprise*. He hypothesized that if unconnected human societies all recognized the same emotions from the same facial expressions, then these emotions existed innately within the human species and were not entirely socially constructed. Though this theory had previously been discussed in part by Darwin (1872) and Tompkins (1962), Ekman's cross-cultural work was ground-breaking, persuasive, and has been extremely influential in emotion research.

Ekman supported his theory with a series of cross-cultural studies where participants were asked to match facial expressions with emotion labels or emotion-eliciting scenarios (Ekman et al., 1969, 1987; Ekman, 1992b). His research consistently found that participants could recognize certain emotions regardless of their cultural background or exposure to other people groups. Later, Ekman separated *Contempt* from *Disgust* as a seventh basic emotion (Ekman et al., 1987; Ekman & Heider, 1988). These seven basic emotion labels were later broadened into emotion categories or "families" of emotion (Ekman, 1992a). However, Ekman (1992a) explicitly stated that these emotion families were not fuzzy sets, although each family contained variations around a core theme. In his theory, core emotions were fundamentally separate from one another.

Though Ekman has formally integrated contempt into his theory of emotions (Paul Ekman Group, 2021), it has not necessarily permeated popular or scientific consciousness. The phrase "Ekman's six" returns over twice as many hits on Google as "Ekman's seven" (3.2 million vs. 1.4 million). Psychological research citing Ekman's categories often does not

use *Contempt* (e.g., An et al., 2017).

Plutchik's Wheel of Emotions

Robert Plutchik sought to create one single comprehensive theory of emotion that would be suitable across all areas of psychology. His theory was inspired by the mixing of colors: just as colors are mixed to create various hues and possess complements and opposites, so too could emotions (Plutchik, 1962). Plutchik proposed that there existed a finite number of pure primary emotions, of which all other emotions are created through their combination and variations in intensity. These primary emotions differed from each other in physiology and behavior, but at their core they were idealized, hypothetical constructs. Primary emotions were derived from evolution and thus were shared outside of the human species to an extent.

Plutchik's theory was based upon categorical labels that existed within a psuedo-dimensional framework where pure emotions were organized in a wheel or cone. He named eight pure emotions: *Anger*, *Fear*, *Sadness*, *Disgust*, *Joy*, *Surprise*, *Trust*, and *Anticipation*. Plutchik's emotions overlapped with Ekman's six alongside the addition of *Trust* and *Anticipation*, but Plutchik's placed contempt (Ekman's seventh) as a derivative emotion rather than a core emotion. Though the labels are similar to Ekman's, Plutchik's emotions have opposite relations to each other and are fuzzy sets rather than strict divisions due to their hypothetical nature.

The Valence, Arousal, Dominance Model

The Valence, Arousal, and Dominance (VAD) model was described by Albert Mehrabian and James A. Russell in the 1970's (Russell & Mehrabian, 1977). Also known as the Pleasure, Arousal, and Dominance model, VAD posits that emotions can be understood through their location on three dimensions: positive-negative, degree of arousal, and submission-dominance. In contrast to either Ekman or Plutchik, emotions states fell within

this three-dimensional space. Prototypical *Sadness*, for example, was a low valence, low arousal, and low dominance emotion, while *Anger* was a low valence, high arousal, and high dominance emotion. The three dimensions were sufficient to describe all emotional states, and were theorized to be independent, though in practice they are often related linguistically (Jack et al., 2016; Mohammad, 2018; Russell & Mehrabian, 1977; Warriner et al., 2013). Though VAD dimensions were popularized by Russell and his colleagues, Wilhelm Wundt had proposed a similar theory in the early 1900s (Rosensohn, 1963).

2.2 Emotions in Language

Classic scientific concepts of emotion focused on communication through facial expressions. However, how emotion is expressed in language may not directly map onto these divisions. If emotion lexicons use theories of emotions that are not relevant to the written word, measurement and validity may be poor. There are multiple conflicts between facial expressions and text-based emotion expression; what will be discussed here are cultural differences in discrete emotions and how language blurs the lines between distinct emotion states.

Classification across Cultures

Language expresses cultural ideas and is constructed by cultural ideas. Accordingly, linguistic labels of discrete emotions do not necessarily reflect universal, stable meanings. Though all humans experience *Anger*, what is labeled *Anger* varies across people and ethnic groups. Not all languages have the same number of equivalent emotion labels with the same connotations (Fiske, 2020), nor is the lived experience of *Anger* equivalent across the globe (Mesquita et al., 2016). The divisions of emotions may be akin to color perception: while light wavelengths are continuous, discrete color categories are created and experienced through cultural beliefs (Barrett, 2006b). In this way, emotion are often divided following

their cultural conceptions (Jack et al., 2016; Jackson et al., 2019), rather than physiological divisions between different states.

Differences in cultural beliefs about emotions influence the organization of emotion words. Closely related emotion concepts have words that are likely to co-occur in similar texts, may have similar linguistic roots, and are believed by the speakers to express related ideas. The separation between linguistic emotion expression and facial expression was displayed in Jack et al. (2016). In the first part of the study, British English and Chinese speaking participants rated the semantic similarity of various emotion words, and in the second part another group rated similarities between facial expressions and emotion words. Semantically similar English words separated out into eight clusters (*Anger*, *Happy*, *Disgust*, *Sad*, *Surprise*, *Fear*, *Pride*, *Shame*), but Chinese speakers had twelve slightly different groupings (*Anger*, *Happy*, *Disgust*, *Sad*, two clusters of *Surprise*, two of *Fear*, as well as *Embarrassment*, *Shame*, *Pride*, and *Despise*). When examining facial expressions, both groups were aligned only on facial expressions associated with *Happy*, *Surprise*, *Anger*, and *Sad*. Though Chinese and English speakers clearly both experience *Fear*, how *Fear*-related words were clustered and how they were physically expressed did not completely overlap. The seven emotion categories found within English words did not fully align with either Plutchik or Ekman's theories, and they significantly outnumbered the four "universal" facial expressions that were found.

In a different exploration of word-emotion association, Jackson et al. (2019) used network modeling and colexification to describe how emotion concepts showed semantic variability across cultures. Colexification is when individual words express multiple meanings because they are conceptually related in that culture. For example, while the languages in Jackson et al. (2019) each had a word for *Anger*, the emotional connotations of *Anger* words were not the same. For Austroasiatic languages, *Hate* words were closely related to *Anger* and *Envy*. In contrast, Indo-European languages had *Hate* in its own separate cluster with far less relation to *Anger* and *Envy*. There were six word communities found in the Indo-European

languages: *Joy*, *Liking*, *Surprise*, *Hate*, *Anger/Fear*, and *Proud/Shame/Sad*[3]. Importantly, which equivalent words fell under basic emotion categories such as *Anger* and *Sadness* was not consistent across cultures.

It is evident that while several "universal" emotions may plausibly exist, how these emotions are expressed, defined, and understood varies in text and cross-culturally. Both Ekman and Plutchik drew from their American cultural backgrounds when defining their theories of emotion. Though they acknowledged that the situations that elicited different emotions varied cross-culturally (Ekman & Friesen, 1971), the emotions themselves were not thought to. It is ambiguous how well these original theories of discrete emotions are universally transferable to text-based expression or to writers from other cultures.

Linguistic Separation of Emotion

Discrete emotions are not necessarily cleanly separated through language expression. For example, linguistically *Envy* and *Jealousy* refer to similar but distinct emotions. *Envy* is defined as coveting what someone else has, while *Jealously* is the fear that a rival will take what you have. These two emotions, which have distinct scientific meanings, are often conflated in practice. In a study by Haslam & Bornstein (1996), participants were prompted to recount a time where they felt jealous or envious. The prompts were worded ambiguously enough as to refer to either jealous or envious situations without explicitly using either term. The participants then responded to separate scales on jealousy and envy, and to single items that asked specifically how jealous and how envious they felt at the time. While a factor analysis clearly differentiated the two constructs on separate factors, the participants' self-reported single-item jealousy and envy ratings did not. Thus, while envy and jealousy were statistically distinct, participant's internal concepts and actual usage of the labels were blurred. The participant's lexical labeling did not distinguish between the two different emotion reactions.

[3]These are my labels for the communities; Jackson et al. (2019) did not provide labels.

Similarly, *Disgust* is generally defined academically as a reaction to being in the proximity of or offense taken towards noxious stimuli (Plutchik, 1962; Rozin et al., 2000). However, the lay use of the term "disgust" often differs from the academic definition. In Nabi (2002), when participants were asked to retell a time they felt "disgust" or "disgusted", 75% retold stories mainly featuring *Anger* (e.g., when they had been lied to, gossiped about, etc.). In contrast, when stories were specifically elicited using the term "grossed out", over 90% of stories fell into the academic definition of *Disgust* and featured topics like blood, vomit, vermin, or decay. Similar linguistic overlap between *Anger* and *Disgust* was found by Roseman et al. (1994) and Jack et al. (2016). Many of the same words were used to describe experiences of the two different emotions. This linguistic overlap does not necessarily imply their visceral, experienced co-existence. Rather, they likely share metaphorical and linguistic patterns of expression more than shared physical experiences (Royzman & Sabini, 2001).

Similarly, there is the question of how moral *Disgust* (relating to morally offensive transgressions such as lying) relates to core *Disgust* (relating to biologically offensive stimuli like feces). There is active debate in psychology on how and if these emotion categories truly biologically overlap, versus only metaphorically and/or linguistically, and, again, their connection to *Anger*[4] (Cameron et al., 2015; Oaten et al., 2018; Royzman et al., 2014; Royzman & Kurzban, 2011; Schnall et al., 2015; Vicario et al., 2020). For example, morally wrong behavior (e.g., lying to get a promotion) may be described as "vile" or "disgusting". If moral *Disgust* is truly a subset of *Disgust*, then a linguistic overlap poses no issues for emotion detection. If moral *Disgust* is really *Anger* and is only metaphorically linked to core *Disgust*, then most emotion lexicons are likely conflating these separate but linguistically similar entities as one single emotion.

The messiness surrounding *Disgust* is of particular concern as *Disgust* is a core emotion

[4]In fact, some argue that *Disgust* should not be even considered a basic emotion. See Royzman & Sabini (2001) and Jack et al. (2014) for different discussions.

in both Plutchik and Ekman models of emotions and is thus used broadly in ED research. If the creation of emotion lexicons, labeled datasets, or other emotion-stimuli pairs are based upon crowd-sourcing, there is not a guarantee that *Disgust* is properly separated from *Anger* and other negatively valenced categories, or is well defined as a distinct entity.

Explicit and Implicit Communication

Emotion is expressed through language both explicitly and implicitly (Clore & Ortony, 1988; Mohammad, 2021). Explicit statements such as "I am mad" state the writer's emotional state. Implicit expression describes contexts that are associated with the underlying emotion, such as "My dog is sick" or "I am crying". Most written text consists of implicit emotional expression. While explicit emotion words do not require context to disambiguate, implicit words do. Some implicit context-word pairings are near-universal (e.g., "death"), while others are cultural (e.g., the connotation of "vaccine"), and others are specifically contextual (e.g., "proposal" for graduate students or fiancés). Predicting the emotional context from a word can be precarious. Ophir et al. (2020) trained a deep neural network to understand which individuals were at risk of suicide. When specifically looking at high risk individuals (n = 361), there was only a single instance where a post with the words "suicide", "kill", or "die" (n= 72) may have been related to death: "Cramps so bad, I want to die".

Positive and Negative Valence

One pattern that can be seen in the results of both Jackson et al. (2019) and Jack et al. (2016) is that there are more negatively valenced emotion clusters than positively valenced emotion clusters. Similarly, both Ekman and Plutchik's sets also contain far more negative emotions than positive emotions. Psychologically, negative events are generally more potent, varied, and salient (Rozin & Royzman, 2001). Positive information is also

processed faster than negative information, which is hypothesized to be due to the greater similarly in positive information versus negative information (Alves et al., 2017). This may help explain, and result in, greater differentiation among negative emotions than positive emotions. Accordingly, there are more individual negative words than positive words, and this trend is not isolated to the English language (Jack et al., 2016; Rozin et al., 2010; Schrauf & Sanchez, 2004). If positive events are more similar to each other and less salient, then less differentiation is required to communicate positive experiences. Perhaps there is be some truth in the classic first sentence of Anna Karenina if applied to linguistics - "Happy [words] are all alike; every unhappy [word] is unhappy in its own way."

2.3 The Impact of Emotion Theories on Emotion Detection

ED relies on psychological research to justify that emotions are well-defined, divisible, and measurable. This reliance is necessary because ED is primarily concerned with the technical hurdles of automatically extracting emotion from text. However, these justifications are largely based upon the simple and highly palatable theories from the 1970s and 1980s. While these theories have not been wholly discredited, relying on research that is four decades old as the basis for cutting-edge applications of artificial intelligence and machine learning is precarious. It simply cannot be assumed that intuitive definitions of emotions based upon facial expression differences are best suited to measuring written language.

Without a demonstrable connection with psychological theory, the validity of ED results is called into question. Just because an algorithm can sort words into different discrete emotions does not ensure that those categories exist or that the algorithm relies on meaningful, generalizable features. That such errors and biases can occur in machine learning is extremely well-documented (Mehrabi et al., 2021). Thus, for ED to be most useful, the construct validity of the emotion measures used must be investigated, especially among emotion lexicons. In the next section, the methods and techniques involved in ED

and emotion lexicons will be described.

3 Emotion Detection in Text

In psychology, language has been used to understand cognitive processes since the field's inception. Freud hypothesized a direct link between spoken words and inner thoughts in his early works (Freud, 1901). Though his ideas have largely been discredited, the analysis of language has lived on. Through the mid-century, researchers used content analysis to measure attitudes and desires through the occurrence of sets of words (e.g., Lasswell et al., 1952). Today, analyzing emotion in text has been used to help understand everything from suicide to the spread of information online (Brady et al., 2017; Glenn et al., 2020; Pestian et al., 2012).

Prior to the invention and widespread use of computers, all transcriptions and analyses were performed by hand. This manual burden has greatly been reduced through advances in computer science and machine learning. Today, the development and use of text analysis techniques largely occurs in computational fields outside of psychology.

3.1 Emotion Detection and Natural Language Processing

In natural language processing (NLP), computers are used to analyze, parse, and create language (Chowdhury, 2003). NLP includes both emotion detection and sentiment analysis. In sentiment analysis, the focus is the polarity of text. Documents are scored or sorted, and the outcome can consist of a single measure of polarity, or multiple measures each for positivity, negativity, and occasionally neutrality.

Emotion detection is an offshoot of sentiment analysis. Emotion detection (ED, also referred to as emotion analysis or emotion recognition) is the process through which emotional content is analyzed using a set of algorithms or other automatic scoring processes (Acheampong et al., 2020). In contrast to sentiment analysis, ED measures fine grained facets of emotion. This is typically a set of discrete emotions such as *Anger*, *Fear*, and *Joy*,

but dimensions like valence, arousal, and dominance may also be used. In this dissertation, I will focus on the use of discrete emotions rather than dimensional attributes, as discrete representations are more popular (Canales & Martinez-Barco, 2014). Similar to sentiment analysis, ED can classifying or score a text on one or many emotions.

ED can be performed using a wide variety of algorithms and techniques. Some ED methods are simple and straightforward. For example, when using a bag-of-words method the number of words associated with emotions are counted and summed. ED also includes applications of machine learning and artificial intelligence, such as support vector machines, naive Bayes classifiers, neural networks, and deep learning (e.g., Abdullah et al., 2018; Muljono et al., 2016; Polignano et al., 2019). Low computation methods like bag-of-words can be used on their own, but they are frequently used to create features (variables) for more complex methods. ED methods are available for free with substantial documentation in popular R packages such as `tidytext` and `syuzhet` (Jockers, 2015; Silge & Robinson, 2016). Because of this, basic methods of ED are highly accessible to applied researchers.

Fundamentally, ED in text is a challenging endeavor. A brief review conducted by Acheampong et al. (2020) of research articles indexed by IEEE Xplore and Scopus in the last decade suggests that multimodal research on ED (e.g., using speech, body language, facial expressions, etc.) is substantially more common than pure text-based research. As naturally occurring emotion expression is multimodal, text-based emotion expression has significant limitations. Even humans find interpreting emotion and tone in text difficult at times. And by definition, ED requires the measurement of multiple emotion features that may or may not be mutually exclusive. While sentiment analysis is limited to positive, negative, or neutral valence, ED research typically uses between three and eight different emotion outcomes.

Importantly, an ED researcher must pick how many discrete emotions to measure. Too few emotions may leave many documents unsorted, while using too many emotions may be

noisy. In evidence of this, when emotion categories are combined by valence, accuracy in labeling dramatically increases (De Choudhury et al., 2012; Mohammad & Turney, 2013; Poria et al., 2013a). In the best performing lexicon examined by Kušen et al. (2017), the sensitivity for *Anger* was 68%, 34% for *Fear*, 27% for *Sadness*, but negative valence combined was 86%. The choice of an emotion set adds researcher degrees of freedom and increases variation in ED.

Yet, ED methods are critical to develop because polarity does not convey enough information in all circumstances. While *Anxiety* and *Sadness* are both negatively valenced, how they influence behavior can be quite different (Raghunathan & Pham, 1999). When the goal of an analysis is not merely prediction but understanding how the independent variables are related to the outcome, fine-grained emotion divisions are necessary. Thus, it is critical to ensure that lexicons reflect the specific constructs they claim to measure.

3.2 Emotion Lexicons

An emotion lexicon describes how various text features are associated with different emotion categories. Emotion lexicons are widely used in NLP as a source of emotion information because the majority of NLP algorithms cannot label emotions without an prior source of information. Lexicons can serve both as the sole method of analysis and as a step in a much larger pipeline. Texts can be scored by matching their words to those in the lexicon and then summing the associations, as in the bag-of-words method. While blunt, such methods are simple and easily interpreted and can be used on their own or as variables in another analysis.

Fundamentally, a lexicon consists of an entry, some number of target classes, and the associations between them. For reference, Table 3.1 shows an excerpt of the DepecheMood++ lexicon. The lexicon entries can be broadly described as *tokens*. Tokens are units of text that are broken down to the analysis level. Tokens can consist of

Table 3.1
An excerpt from the DepecheMood++ lexicon.

	Afraid	Amused	Angry	Annoyed	Don't Care	Happy	Inspired	Sad
lawn	0.06	0.33	0.10	0.07	0.05	0.10	0.18	0.11
rotunda	0.33	0.06	0.04	0.19	0.15	0.12	0.08	0.03
virginia	0.15	0.17	0.11	0.12	0.14	0.07	0.10	0.13

Note:
The numeric association between words and discrete emotions are probability weights. Higher weights indicate a higher probability that the word reflects the emotion.

words, phrases, punctuation, emoticons/emojis and any other divisible text feature that a researcher wishes to analyze. Tokens do not have to be single words; tokens can be described as n-grams, which are n number of tokens together.

Some lexicons may contain lemmas or stems rather than raw tokens, or may have part of speech tagging. *Lemmas* are the root or 'dictionary' form of a word (e.g., run, ran, and running all belong to the lemma "run"), while word *stems* are created when words have their endings removed [e.g., hike → hik-, hikes → hik-, hiking → hik-) (Schütze et al., 2008). Matching tokens to stems is easier than matching to lemmas because stemming can be performed using broad grammatical rules. Whether lemma or token-based approaches perform best is context and language dependent (e.g., Araque et al., 2018).

Entries or words are associated to target classes. Classes are any measurable dimension or category that can be found in or represented by text. In NLP, classes include everything from linguistic features to psychological constructs. In ED, the lexicon classes are generally a set of discrete emotions or dimensions. The associations between entry and class can be expressed as probability weights, binary assignments, intensity scores, etc. Some lexicons allow for entries to be related to multiple classes; others restrict entries to a single association. Dimensional lexicons tend to only use intensity scoring systems, while the scoring systems of discrete emotions vary.

The associations (also called annotations) between tokens and classes are discovered or

assigned through a variety of processes. In ED, the association between token and class almost always begins with a set of human labeled data. How large this set is, and how much of the final set this comprises, varies drastically. A small set of labeled data can be used to seed a much larger set by analyzing word co-occurrence using techniques such as topic models or word-embedding (Laville et al., 2020; Yang et al., 2014). Associations may be automatically generated by analyzing the occurrence of words inside pre-labeled datasets, like hash tagged social media posts (Hasan et al., 2019; Mohammad, 2012). Such web-scraped or seeded lexicons require relatively few resources to create, can be exceedingly large, and can detect subtle associations between token and class. However, their quality is heavily dependent on the methods used, an issue that will be discussed in Section 3.4.

Crowd-sourced lexicons ask hundreds of participants to rate the associations between words and classes. While this method has the advantage of asking humans directly about emotion content, it does have drawbacks besides cost-efficiency. For example, instructions given to raters can heavily influence associations. In Mohammad & Turney (2013), annotators agreed more often on the emotion-word association when they were asked if the emotion was *associated* with the word versus *evoked* by the word. Crowd-sourced ratings are also limited by human knowledge. While the emotional content of a word like "death" may be readily apparent, not all word-emotion associations are obvious. Similar to how pronouns can indicate mental states (O'dea et al., 2017), there are invisible associations between tokens and classes that cannot be directly opined on.

3.3 Lexicons of Interest

Below I will describe four popular, general purpose lexicons used in ED research that I will examine in this dissertation.

DepecheMood++

The original DepecheMood lexicon was created by Staiano & Guerini (2014) and updated in 2018 to DepecheMood++ (DM, Araque et al., 2018). The English DM lexicon was crowdsourced from Rappler.com, an English-language Filipino news website. Rappler contained a "mood meter" widget on every article that allowed readers to select one of eight emotional reactions: Afraid, Amused, Angry, Annoyed, Don't Care, Happy, Inspired, or Sad. By scraping text off Rappler, Staiano & Guerini (2014) were able to calculate the likelihood that a given word would be associated with the articles' overall scores on the eight emotion reactions. DM is available in three forms: token (N = 187,942), lemma (N = 175,592), and lemma with part of speech tagging (lemmaPOS; N = 284,597). The token form will be used in this dissertation.

In DM, the associations between words and emotions are expressed through probability weights, or the likelihood that a given word will be associated with an emotion. The probability weights of each word (row) sums to one. As will be discussed in the Methods Section, this produces compositional data which causes unique issues in multivariate data analysis. There are no rows in DM that do not sum to one; that is, all words have emotion associations.

Though DM is exceedingly large, not all entries are useful. Non-words such as "aaa" and "zzjjw" are present in all three forms of the lexicon. Araque et al. (2018) suggested excluding tokens that appeared less than 10 times in their corpus for optimal prediction accuracy. This screening reduces the lexicon to 26%-20% of its total size, depending on the form of the lexicon.

In regards to construct validity, DM largely relies on its crowd-sourcing premise, though some basic investigations were performed in Araque et al. (2018). The Pearson correlation between a set of annotated headlines and their associated summed DM scores was 0.33. However Pearson correlations are not always meaningful for compositional data, so the

interpretation here is ambiguous (Aitchison, 1982). Araque et al. (2018) also compared DM against several other lexicons and found that DM results were equivalent or better in both prediction and classification problems.

NRC Word-Emotion Association Lexicon

The NRC Word-Emotion Association Lexicon (EmoLex) was created by Saif M. Mohammad and Peter D. Turney from the National Research Council Canada (Mohammad & Turney, 2013). EmoLex is one of several sentiment and emotion lexicons that have been published by Mohammad. Each of the NRC lexicons are large, varied, and well-known. EmoLex contains 14,182 terms scored on Plutchik's eight emotions and two valence categories. Associations are binary, and terms can be associated with multiple emotions. Of the total terms in EmoLex,0.69 are not associated with any emotion.

To create EmoLex, over 2,000 participants were recruited from Amazon's Mechanical Turk service to manually rate the degree of association between terms and discrete emotions. Each term was rated by five participants. Data collection was fairly rigorous and involved question piloting, attention checks, and the removal of poor quality responses. Participants rated terms on a 1 to 4 scale, however the final lexicon scores were assigned 0 if the average rating was less than two and 1 if the rating was higher than two.

Mohammad & Turney (2013) compared the associations within EmoLex to two other lexicons (the General Inquirer and WordNet Affect). Agreement between the valence scores of EmoLex and the emotion categories of the General Inquirer and WordNet Affect ranged from 80% to 97%. When examining individual categories, agreement varied. For example, 66% of Surprise words overlapped between EmoLex and WordNet Affect, compared to 94% of Disgust words. Internally, Fleiss's κ for valence was 0.62 for negative (substantial agreement) and 0.45 for positive (moderate agreement).

EmoSenticNet

EmoSenticNet (ESN) is the combination of a small, popular emotion lexicon (WordNet Affect) and a sentiment lexicon (SenticNet) to create a larger, more useful emotion lexicon (Baccianella et al., 2010; Poria et al., 2012, 2013b; Strapparava & Valitutti, 2004). ESN uses Ekman's six emotions: Anger, Disgust, Joy, Sad, Surprise, and Fear. The associations within ESN are binary, and most entries are only associated with one category. Each entry has at least one emotion association. There are 13,189 terms in ESN, of which, 5,478 terms are single words and 7,711 are phrases (e.g., "blank space").

To create ESN, a Support Vector Machine classifier was trained on WordNet Affect labels and extended to words in SenticNet using features from the International Survey of Emotion Antecedents and Reactions (ISEAR) dataset (Scherer & Wallbott, 1994). The ISEAR dataset contains writings from participants on different emotion experiences, as well as specific questions about how the participant felt during the experience, and behavioral and physiological data collected during the retelling of the experience. Words in SenticNet were labeled with emotions using features from ISEAR and distance-based similarity measures. Accuracy of the labeling was measured with tenfold cross-validation against the original labels/polarity assignments in ISEAR, WordNet Affect, and SenticNet.

Linguistic Inquiry and Word Count

The Linguistic Inquiry and Word Count 2015 (LIWC; pronounced "Luke") is a proprietary text analysis program originally developed by the social psychologist James W. Pennebaker and colleagues (Pennebaker et al., 2015). LIWC uses a dictionary (lexicon) and an algorithm to count the number of words that belong to various categories. The creation and refinement of LIWC reflects its psychological origins. In contrast to crowdsourced lexicons, each entry was evaluated for inclusion based on expert opinion, experimental studies, psychometric analyses, and reliability measures such as Cronbach's

alpha. Non-experimental sources of words included blog posts, social media, novels, and the New York Times. LIWC is used extensively in psychological text research, but less frequently in data science fields (e.g., Guerini & Staiano, 2015; Hasan et al., 2019).

In essence, LIWC is a stream-lined bag-of-words application containing almost 6,400 words, stems, and emoticons (Pennebaker et al., 2015). LIWC matches are based either on whole words or stems, so terms like "run" and "running" may receive the same score. Word associations are binary, and a limited number of words are associated with multiple categories. LIWC contains the emotion classes of Positive Emotion, Anger, Anxiety, and Sadness. Other variables that can be measured include psychological constructs such as Affiliation and Power, as well as linguistic and grammar categories such as pronouns, adverbs, and negation. Not all terms in LIWC's dictionary are associated with an emotion class; because LIWC's dictionary is proprietary an exact count is not possible. However, of all the words found in the other lexicons that also are in LIWC's dictionary, approximately 20% belong to at least one emotion class.

Lexicon Construction Summary

I will briefly summarize differences between the four lexicons. LIWC is heavily theory based. Many dictionary words were hand-picked and assigned by emotion researchers, though the occurrence of words within texts was also examined (Pennebaker et al., 2015). ESN and DM were automatically generated through statistical modeling. DM can be considered "naively crowdsourced" as ratings were generated by participants, though the participants were not rating articles for the purpose of lexicon development. EmoLex was purely crowd-sourced from online participants. LIWC and EmoLex had the most internal validity and reliability checks during their construction, while ESN and DM mainly examined prediction accuracy.

3.4 Lexicon-Text Match and Interpretation

Emotion lexicons contain words that either explicitly or implicitly denote emotion (Clore & Ortony, 1988; Mohammad, 2021). Depending on the application, this distinction may be critical or trivial. However, it is important to understand that very large lexicons are constructed almost entirely of implicit and contextual emotions. As discussed in the last section, there are simply far fewer explicit synonyms for emotions than there are implicit representations of them.

Emotions are reactions to stimuli, and which stimuli provoke which emotions is contextually determined on both large scales (cultures) and small scales (contexts). Because emotion is contextual, lexicons contain some degree of contextual associations; by definition contextual associations are not universal or stable. For example, the word "unpredictable" may be a positive appraisal for a movie, but not for a car. A completely universal context-free lexicon would necessarily be a very small lexicon. Therefore, the match between the lexicon and the analyzed text is critically important to reduce unintended associations as much as possible and to maximize the amount of scored text.

How a lexicon is constructed determines the emotion-word associations it contains. For example, crowd-sourced lexicons are heavily influenced by their participant sample. Crowd-sourced EmoLex contains scores for some seemingly neutral words (e.g.,"tree" with Disgust and Anger) and scores that reveal cultural opinions (e.g., "lesbian" with Sadness and Disgust) (Zad et al., 2021). Automatically generated lexicons suffer the same pitfalls. The DM lexicon contains emotion weights for the names of politicians such as "Obama" and "Clinton" because it was based upon reader reactions to news articles (Araque et al., 2018). Some amount of unintended associations is likely unavoidable unless every word-emotion association is checked. Caliskan et al. (2017) showed that common human biases surrounding race and gender are easily recovered from general purpose text corpora when using machine learning models. These biases are invisibly threaded through text and then are reflected in

the lexicon associations. However, while sometimes these biases and associations are errors or artifacts, other times they may actually reflect meaningful differences in measurement.

Let us imagine a study of tweets about two groups who are writing about individual freedoms. The analysis uses DM and finds that Group A uses more Angry words. Prior to concluding that Group A tweets with more *Anger*, it is important to determine what is driving the Angry scores. Say Group A focused on the right to own a gun, while the Group B did not. DM was based upon reactions to the news website Rappler. If the readers of Rappler were largely anti-gun *or* the articles with gun vocabulary involve violence rather than marksmanship competitions, then any gun-related tweet would be scored negatively regardless of its true emotional content. Group A could tweet about how collecting and shooting guns at ranges is a positive social experience, but DM scores would not reflect this. Similarly, it is true to say that some groups are pro-LGBTQ and others are not. Using a lexicon that has negative LGBTQ word-emotion associations may be fundamentally useless when scoring tweets written by LGBTQ activists, yet accurately reflect the emotions of religious, conservative writers. Because lexicons are not context-free, the results of their application will be influenced by their construction. Researchers must be prudent in matching lexicons to texts to ensure their interpretations are an accurate reflection of reality.

For the purposes of this dissertation, the issue at hand is that general purpose lexicons from different sources may not truly generalize, and therefore may measure different emotion constructs. Some degree of difference is to be expected. However, if these differences are widespread then the constructs that each lexicon claims to measure may drift apart. In the next section, the issue of construct validity across emotion lexicons will be more thoroughly described.

4 Emotion Theories used in Emotion Detection

In this section, I will described how psychological theories of emotion influence the sets of discrete emotions that are used in ED research. Because emotion must be operationalized to be measured, some set of discrete or dimensional emotions must be chosen as features or outcomes (dependent or independent variables). Which set of emotions are chosen begins a cascade of influence on measurement, and if a lexicon is used then those will be the emotions a study relies upon. Sets of emotions can be divided into theory-based or non-theory based sets. *Theoretical sets* are those that are justified by psychological research. *Non-theoretical sets* are not based upon research; they are often created by the researcher or arise intrinsically from a data-source.

4.1 Theoretical Sets

It is logical that ED researchers would turn to psychology to justify picking the discrete emotions to extract from text. The most common emotion categories used in ED are those proposed by Ekman and Plutchik (Ekman, 1992b; Plutchik, 1962). Truly, the seminal works of least one of these authors are cited in nearly every ED paper that I have read, regardless of what emotion sets are ultimately used.

Ekman's emotion categories are pervasive in ED. This may be due, in part, to its small number (6) and that it contains less abstract emotions than Plutchik (i.e., there is no *Trust* or *Anticipation*). When Ekman is referenced, it is almost exclusively as "Ekman's Six" with *Contempt* ignored (e.g., Abdullah et al., 2018; Acheampong et al., 2020; S.-Y. Chen et al., 2018; Mohammad & Turney, 2013; Sykora et al., 2013). This exclusion is also done by EmoSenticNet (Poria et al., 2013a; Strapparava & Valitutti, 2004).

The dataset from EmotionX 2019, the shared task of SocialNLP 2019 Workshop, used Ekman's six plus a neutral category (S.-Y. Chen et al., 2018). However, the challenge itself

only included *Joy*, *Sadness*, *Anger*, and *Neutral* due to low coverage of the other three labels in the dataset. Such exclusion is a common issue in ED and natural language processing. Competitions like EmotionX 2019 are very popular and their datasets tend to be large, clean, and publicly available. In ED, these datasets are often referred to as "gold-standards", or equivalent to the "ground-truth". They are easy, well-known benchmarks to test new methods and lexicons against. Thus, shared task workshops and competitions can have disproportionate influence on the field in regards to construct measurement and theory as their emotion sets are replicated every time the dataset is used.

Plutchik's work is also frequently used in ED. His model is largely treated as consisting of discrete labels only, even though his theory involves dimensions and his discrete labels were hypothetical constructs, not real entities. Further, aspects of intensity and/or mixed emotion states outside of his core eight are not typically used in ED. This application of Plutchik's model is likely a combination of tradition and quantitative limitations rather than theoretical beliefs.

A SemEval2018 shared task used Plutchik's eight plus three additional categories: *Love*, *Optimism*, and *Pessimism* (Mohammad et al., 2018; "SemEval-2018 Task 1: Affect in Tweets," 2018). Abdullah et al. (2018) used the SemEval2018 dataset to detect *Anger*, *Joy*, *Sadness*, and *Fear* from Arabic tweets using deep learning. Polignano et al. (2019) detected *Happy*, *Sad*, *Angry*, and Other using the SemEval2018 dataset. As mentioned, EmoLex also uses Plutchik's eight emotions. Vosoughi et al. (2018) used EmoLex and its set to evaluate the spread of truth and falsehood on Twitter.

Occasionally, other psychological theories are used in ED. Bollen et al. (2011) associated the stock market with Twitter posts by creating a mood lexicon based on the Profile of Mood States questionnaire (POMS, McNair et al., 1971). The analysis used the six POMS categories of *Calm*, *Alert*, *Sure*, *Vigor*, *Kind*, and *Happy*. Sykora et al. (2013) surveyed the works of Drummond, Ekman, Izard, and Plutchik as well as their own experience with

Twitter in deciding to use Ekman's six plus *Confusion* and *Shame*. Wang et al. (2012) used the set defined by Shaver et al. (1987) (*Love*, *Joy*, *Surprise*, *Anger*, *Sadness*, and *Fear*) plus their own addition of *Thankfulness*.

Though ED researchers may seek to use psychological emotion sets in their work, entire emotion categories are often excluded due to poor inter-rater agreement or low coverage. As mentioned, this happened in EmotionX 2019. In another example, Liu et al. (2019) set out to use Plutchik's eight in building a large, pre-labeled ED dataset consisting of paragraphs, rather than sentences. Raters had difficulty labeling passages with *Trust*, so it was dropped and replaced with *Love*. The researchers also recommended against the use of their *Disgust* and *Surprise* labeled passages due to their small size and high level of noise.

Through reading many ED articles, it is clear to me that psychological theory is often cited to justify the emotions a researcher wishes to use, rather than as the basis of measurement. This is not surprising as emotions are a tool in ED research; philosophical questions about the nature of emotion are largely immaterial. Psychological theories provide small, convenient sets of emotion that both overlap with lay beliefs about emotion and are backed up by research. Small sets in particular are also computationally easier, more reliable, and for lexicons, cheaper to use (Mohammad, 2021). However, these sets are often applied outside of the theoretical framework they come from, often transforming "basic" into a synonym for "simple" or "common". The larger implications about universality or fundamentality are not considered; this is especially evident when other emotions like *Love* and *Confusion* are tacked on at the end. Thus, the use of theoretical sets does not entirely imply the construct validity of the emotions measured, though the absence of validity is also not inevitable.

4.2 Non-Theoretical Sets

Sets that are not based on psychological theories of emotion are also common in ED. Non-theoretical sets typically arise from intrinsically or pre-labeled datasets, though researchers may also simply pick their own to use. If a large text dataset has intrinsic emotion tags, then those tags will assuredly become ED features or outcomes. This is because ED typically requires large amounts of data; the more thousands of texts available, the more accurately the algorithms can perform. Manually labeled datasets require significantly more time and financial investment to create as opposed to cleverly exploiting intrinsic text features. However, the construct validity of non-theoretical emotions is less assured than theoretical emotions.

Primarily, the issue is that intrinsic sets of labels may not represent meaningful divisions of emotion. As previously mentioned, the DepecheMood++ lexicon was built with pre-labeled data, which allowed it be so large. Araque et al. (2018) scraped data off the news website Rappler.com which featured a native "Mood Meter" widget that provided emotion scores for each article. Therefore, the eight emotions in DM were chosen by Rappler and were not selected based upon any scientific theory; "Don't Care" is not an emotion recognized anywhere else but in Rappler. In another example, Zimmerman et al. (2015) analyzed emotions expressed through Facebook posts that were tagged with different emotions. When a dataset of these posts was released by Lamprinidis et al. (2021), the categories Anger, Anticipation, Fear, Joy, and Sadness were created out of the over 30 original tags. I consider this a non-theoretical set because the emotions were chosen based off of which tags were most popular, and were created by combining different hashtags that may or may not have measured the same constructs.

Non-theoretical emotions may cause poor measurement and muddy interpretations, especially if the emotion within a set are closely related. For example, if *Love* is tacked on to Ekman's set, the difference between *Love* and *Joy* is difficult to interpret. If a text is labeled

Love, does this imply there is no *Joy* present, or are all *Love* texts also *Joyful*? Another example of this is the DM lexicon which contains both Angry and Annoyed. *Annoyance* does not frequently appear as a separate emotion from *Anger* in psychological research, even in studies identifying more than twenty discrete emotions (Cordaro et al., 2018; Cowen & Keltner, 2017). Because *Anger* is likely superordinate to *Annoyance*, it is unclear how the two categories should be interpreted. Their separation implies that the Angry category measures all facets of *Anger* except *Annoyance*. Yet it is unclear if this is true, or if Annoyed contains less intense *Anger* words than Anger, or if Annoyed contains words related to specific situations that are not encompassed by Angry. This is one advantage of theoretical sets; clear differentiation between emotions is more likely.

It should be noted that non-theoretical sets may still be appropriate. Quan & Ren (2009) created a Chinese-language corpus labeled with eight emotions: Expectation, Joy, Love, Surprise, Anxiety, Sorrow, Anger, and Hate. These labels do not align with any cited set of emotions. While these set may seem messy to a Western audience, these labels may be entirely appropriate for the Chinese language. In fact, there is significant overlap between these labels and the Chinese langauge clusters found by Jack et al. (2016). It is beyond the scope of this dissertation to further speak on whether psychological theories of emotion are ill-suited to non-English ED research, however I believe this friction represents a rich area of research that has yet to be explored. While it is quite common for English lexicons to be translated into other languages (EmoLex is available in 108 languages), the validity of applying English emotion sets to non-English text has not been examined.

4.3 Comparisons between Sets

In summary, ED research uses a combination of theoretical and non-theoretical sets of discrete emotions as features and outcomes. Most emotion sets overlap at least partly with Ekman or Plutchik's theories, regardless of their generating source. The exclusion of

theoretical emotions and the inclusion of non-theoretical emotions is common, and may be more or less appropriate in different circumstances. However, inconsistent emotion categories is not just a problem for non-theoretical sets, but also when comparing non-overlapping theoretical sets. This is especially critical when comparing emotion lexicons and the studies that use them.

It is difficult to ensure that two lexicons measure the same constructs when the sets do not overlap. Researchers often use "close enough" matches to make comparisons across lexicons, or assume that same-named variables measure the same constructs, even if the sets are different (Araque et al., 2018; Kušen et al., 2017). For example, if Angry-DM does not contain *Annoyance* words, then Angry-DM may not measure *Anger* in the same way as Anger from EmoLex, ESN, or LIWC. This difference would not be obvious to an unfamiliar researcher. Similarly, neither DM or LIWC have a *Disgust* category, but there is a close linguistic connection between *Anger* and *Disgust*. Angry-DM and Anger-LIWC may contain some *Disgust* terminology that muddles their relationships with Anger and Disgust from EmoLex and ESN. Again, the same issue plays out with *Surprise*. Do DM and LIWC not measure *Surprise* at all, or are positive aspects of *Surprise* subsumed into their categories of Joy and Happy? It is these issues of measurement and overlap that I am keen on investigating in my dissertation. As natural language processing techniques continue to multiply in psychological research, an evaluation of their measurement constructs is critical.

Currently, standard secondary analyses of lexicons are limited to head-to-head comparisons of their prediction accuracy. Prediction accuracy is typically measured through precision (positive predictive value) and recall (sensitivity), and/or F-scores (a combination of precision and recall) when classifying texts into a single emotion class. While prediction accuracy is incredibly important, it does not fully describe the measurement abilities of lexicons. First, prediction comparisons are limited to emotions that overlap between all lexicons *and* the target dataset (e.g., Kušen et al., 2017; Raji & De Melo, 2020; Tabak & Evrim, 2016). Therefore, there is little information on non-overlapping emotions - which

are often the most contentious. Second, classification accuracy does not provide deep insight into the constructs being measured. If two lexicons correctly assign labels to 80% of the sample, it is unknown what aspects of the 20% is causing the mis-prediction. One lexicon's Anger may include *Disgust* terminology, while another lack words relating to *Hostility*. By only looking at prediction, these distinctions would not be revealed. And finally, prediction comparisons reveal little about the internal relationships within lexicons. Emotion categories are not orthogonal. Some correlations should be present, while other emotions should be distinct. Basic "sanity checks", like how closely *Anger* and *Joy* overlap or if *Anger* and *Disgust* are separable, are not always performed or reported.

Because emotion lexicons are often used in psychological research, it is critical to understand the constructs they measure. If two lexicons measure same-named emotions differently, then their results may not be comparable and their construct validity is called into question. In the next section, I will describe how this dissertation investigates the construct validity and shared measurement abilities of four different emotion lexicons.

5 Study Overview and General Methods

In this dissertation, I applied psychometric techniques to evaluate four popular general purpose lexicons used in emotion detection. The purpose of this study was to understand how each lexicon's discrete emotions were associated both internally and externally from a construct validity perspective (Cronbach & Meehl, 1955; Embretson, 1983). Item response models were used as the statistical framework for this investigation. If variables from different emotion lexicons truly measure the same constructs, then these same-named emotion should share substantial variance that can be modeled through latent factors. A core assumption here is that what is shared across lexicons best reflects the latent emotion construct. That is, the best measurement of *Anger* is what the *Anger* variables from all four lexicons agree upon. Therefore, lexicon variables that have greater association with the conceptual factors also show stronger evidence for construct validity.

A two-pronged confirmatory and exploratory approach was used to test both a hypothesized factor structure and to understand the native relationships among the lexicons. I believe that the combined confirmatory and exploratory approach is particularly advantageous in understanding lexicon measurement. The confirmatory analysis speaks to how lexicons are currently used by researchers, while the exploratory method allows for a deeper investigation of the construct validity of the lexicons.

In the confirmatory approach, I fit an item response model (IRM) with a factor structure that takes emotion category labels at face value. The confirmatory method sought to understand the measurement capacity of the lexicon variables as they were intended to be used and are currently applied in the literature. The structure of the model was theory driven. It assumed that same-named variables measure the same latent emotion construct and that there are distinct divisions between these emotions. For example, Angry-DM, Anger-EmoLex, Anger-ESN, Anger-LIWC all loaded onto the *Anger* factor,

and no other factor. There were no cross-loading of items between factors; there was simple structure. While this structure was strict, it is how cross-lexicon comparisons are typically conceptualized. The hypothesized factor structure is described in more detail alongside results in Section 6: Confirmatory Approach.

In the exploratory approach, I sought to build the best-fitting and most parsimonious model from the ground up regardless of variable labeling. It is currently unknown how lexicon variables relate to each other beyond comparisons of prediction and classification ability. In this approach, lexicon variables were free to associate with each other regardless of how they were named. Unlike in the confirmatory section, same-named variables were not forced to load together and ill-fitting variables were more freely removed from the model to improve fit. More details and results are described in Section 7: Exploratory Approach.

In this section, I will start by describing the structure of the data/lexicons and then discuss the modeling framework that is pertinent to both the confirmatory and exploratory approaches. But first I will reiterate how homonymous words, categories, and factors are distinguished from each other. Individual words are denoted with quotation marks ("joy"), lexicon emotion variables with uppercase (Joy), and conceptual emotions like factors are in capitalized italics (*Joy*). Lexicon emotion variables/categories are variables found within a lexicon (Joy, from ESN), while conceptual emotions are the hypothesized constructs that these categories seek to measure. For example, Happy from DM and Joy from EmoLex both seek to measure the same *Joy* construct and are associated with the word "happy".

5.1 Comprehensive Lexicon

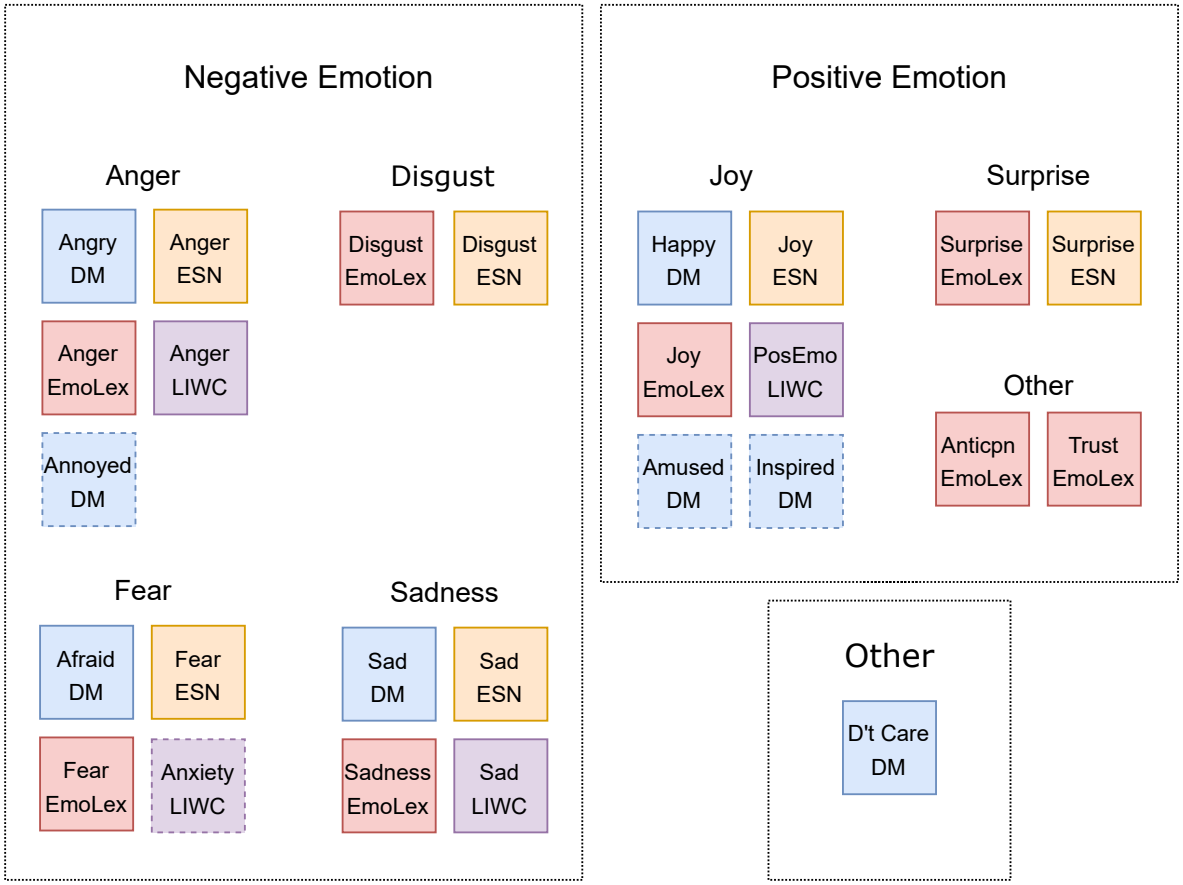
The following lexicons were analyzed: DepecheMood++ (DM), the NRC Word-Emotion Association Lexicon (EmoLex), EmoSenticNet (ESN), and the Linguistic Inquiry and Word Count 2015 (LIWC). These four lexicons were combined into one overarching, comprehensive lexicon referred to as CompLex. DM contributed eight "Mood Meter" variables: Afraid,

Amused, Angry, Annoyed, Don't Care, Happy, Inspired, and Sad. EmoLex contributed eight Plutchik emotions: Anger, Anticipation, Disgust, Fear, Joy, Sadness, Surprise, and Trust. ESN contributed six Ekman emotions: Anger, Disgust, Fear, Joy, Sad, and Surprise. LIWC contributed four emotion-related categories: Positive Emotion (PosEmo), Anxiety, Anger, and Sadness. No discrete emotions were excluded from any lexicon.

Figure 5.1 shows how the emotion variables overlap with each other. This proposed organization is based upon shared labeling across the lexicons as well hypotheses about their structure (e.g., Annoyed-DM as a derivative of *Anger*). Because emotions do not fully overlap between each lexicon, using four lexicons helped ensure that major latent emotion factors had at least three constituent variables.

A sample of CompLex can be seen in Table 5.1. The columns (N = 26) contain the individual emotion variables from each lexicon. Each row contained a single word that was present in at least 3 of the 4 lexicons. Not all words from all lexicons could be included because not all words were found in all lexicons. Some lexicons had more unique words, while others had more substantial overlap (Figure 5.2). This overlap was of significant concern from an analysis perspective as it was functionally equivalent to missing data. Smaller lexicons necessarily have more "missingness" than larger lexicons. When lexicons of different sizes are compared, smaller lexicons end up with more "missing" than "observed" words. For example, DM and EmoLex are the largest two lexicons in this study and share 13,000 words. However, there are over 174,000 DM entries that are not present in EmoLex. Missingness on this scale is not appropriate for statistical analysis and would heavily penalize the smaller lexicons. And specifically for DM, many of its 174,000 tokens have low frequencies and are likely of low quality.

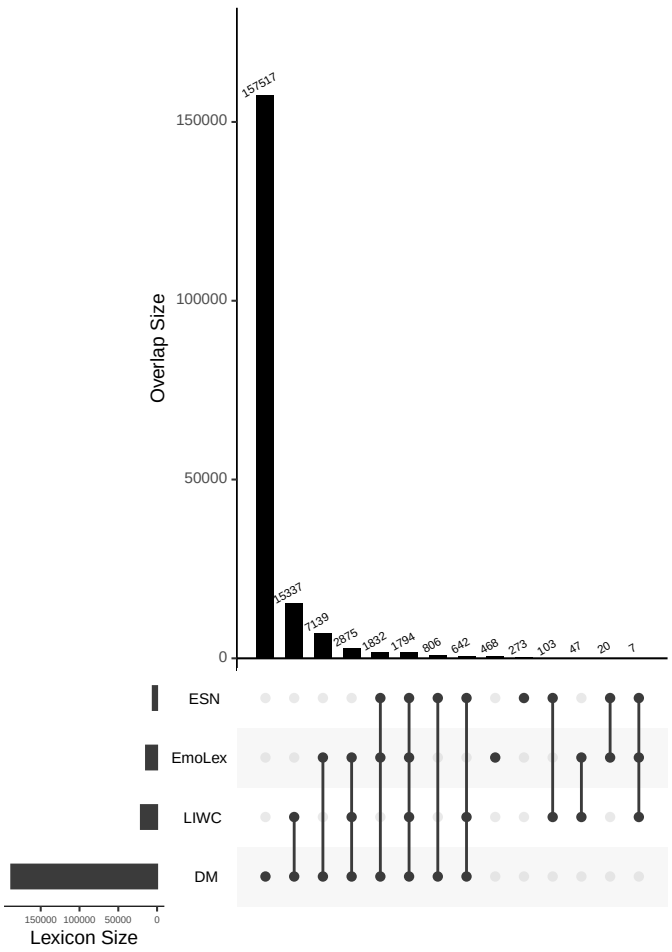
Figure 5.1
Hypothesized Category Overlap



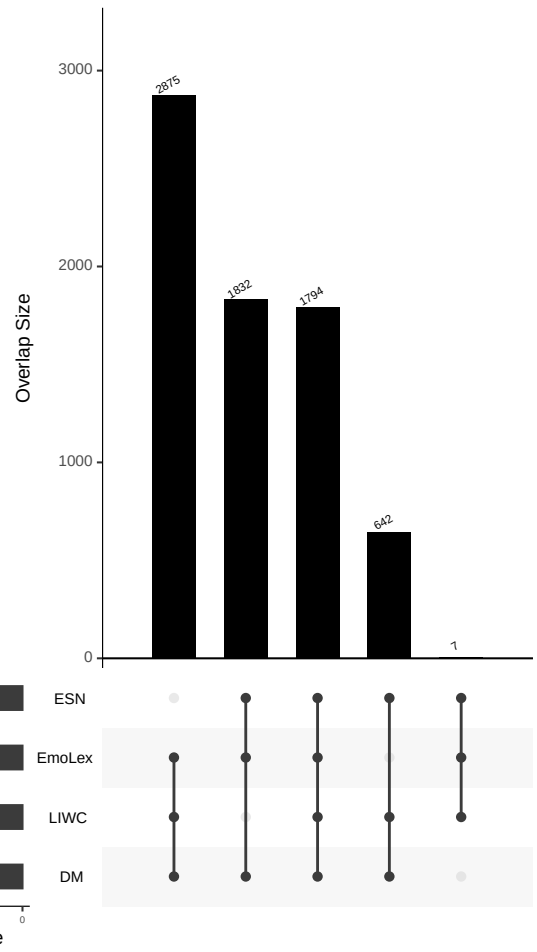
Note. Dashed lines indicate variables that do not share the same name as the group they are hypothesized to belong to. The Negative and Positive groups are included for clarity.

Figure 5.2
Sets of Overlapping Words Across Lexicons.

(a) *Raw Lexicon Overlap*



(b) *CompLex Overlap*



Note. UpSet plots are more accurate alternatives to Venn diagrams (Gehlenborg, 2019). The histogram on the bottom left section of each plot shows the relative size of each lexicon. The main upper histogram displays the size of the sets, i.e., how many words overlap. The dots underneath the main histograms indicate lexicon membership of each set. For example, the first bar in the CompLex figure indicates that ESN, LIWC, and DM share 2,875 words.

Table 5.1
CompLex Sample

Word	Happy-DM	Joy-EmoLex	Joy-ESN	PosEmo-LIWC
baby	0	1	1	0
mother	0	1	NA	0
new	0	NA	1	0

Note:
Shown are three rows and selected positive emotion variables from CompLex. The word "baby" is found in all lexicons, while "mother" does not appear in ESN and "new" does not appear in EmoLex. For DM, the Chance transformation is shown.

Therefore, CompLex only included 7,150 words that appeared in at least 3 of the 4 lexicons. Of these words, 5,356 (74.91%) are missing observations from one of the four lexicons. Therefore, 16.01% of the data points are missing. The rates of missingness by lexicon can be seen in Table 5.2. While missing data can bias IRM parameters (Thomas et al., 2016), the large size of this dataset compared to its missing rate is a significant protective factor (Enders & Bandalos, 2001; Schafer & Graham, 2002) as does using maximum likelihood model estimators (Zhang & Walker, 2008). However, it should be noted that the results of this study are only applicable to the overlapping words between lexicons. It is a reasonable assumption, though, that if the overlapping words between lexicons function similarly, then the unique words of each lexicon likely measure the same emotions in the same way as their observed counterparts.

In summary, CompLex consists of 7,150 observations of twenty-six different emotion variables from four lexicons. All emotion categories consist of binary data, except for DM variables which contain either binary or polytomous data. The next section details any alterations or transformations done to the lexicons to create CompLex.

Table 5.2
CompLex Missingness

	Original Size	Entries in CompLex	Percent Missing in CompLex
DM	187,942	7,143	0.1%
EmoLex	14,182	6,508	8.98%
ESN	5,477	4,275	40.21%
LIWC	20,805	5,318	25.62%

Note:
The original size of LIWC indicates how many words found in the other three lexicons were scored by LIWC, not how many unique stems are in the LIWC dictionary.

DepecheMood++

The DepecheMood++ token lexicon (DM) required transformation before it could be used in multivariate analyses because it contains probability weights that sum to one across rows (i.e., is closed). This is compositional data which lies in a mathematical simplex rather than in real euclidean space (Aitchison, 1982; Greenacre, 2021). Without special consideration, the analysis of compositional data reliably distorts the relationships between variables, typically producing spurious correlations (Pearson, 1896). Because the data is closed, when one the value of one variable increases then values in some other variable(s) must also decrease. While factor analytic and item response models have been developed for compositional data, there are significant limitations. Generally, these methods require either isometric or additive logratio transformations where the ratio between each variable and a reference variable are modeled (Brown, 2016; C.-W. Chen et al., 2021; Coenders et al., 2011; Filzmoser et al., 2009). As a result, these transformations take D dimensional components to produce D-1 ratio components in real space. The purpose of this study is the interpretation of lexicons through latent trait modeling. Because the relationships of ratio variables such as $ln(\frac{happy}{angry})$ alongside standard non-ratio variables cannot be easily interpreted, analyzing the DM lexicon with compositional data methods would force its exclusion.

However, the compositional nature of DM is rarely considered in applied studies.

Typically, DM is applied by summing the raw probability weights of each word in the text, by binarizing the lexicon and then summing the binarized scores, or by calculating similarity measures based on euclidean distances (e.g., Agrawal et al., 2018; Gollapalli et al., 2020; Mejova & Kalimeri, 2020; Vorakitphan et al., 2021). Sums of compositional data are still compositional, and DM scores do not lie in euclidean space. I have seen no comment on this in papers using the DM lexicon. It is unknown how compositional sums may bias ED analyses.

Because the raw probability weights are unamenable to standard methods of analysis, some transformation must be applied if the DM lexicon would be included in the present study. Two different transformation methods were explored: polytomization and binarization. The results of these two transformations were compared in the Confirmatory analysis.

Polytomous Transformation. In the Polytomous transformation, each continuous DM emotion variable was binned from 1 (lowest quantile) to 5 (highest quantile) to produce polytomous scores. However, Don't Care, Annoyed, and Afraid had extreme positively skewed distributions where the first and second quantiles contained values that were all equal to zero. Therefore, these variables have scores from one to four instead of one to five.

This transformation allowed for words to be associated with multiple emotions. However, it does not completely address the compositional nature of the data as the continuous weights are simply rescaled into discrete scores. Because the polytomous transformation bins the continuous weights, it may not reduce any spurious correlations between variables. Polytomous data can be easily handled by IRMs, though, which is a significant advantage over continuous probability weights.

The polytomomous scores are still transformed probability weights, which may cause issues in interpretation and analysis. Because the original probability weights are closed, the weights do not represent the strength of the word-emotion association. Instead, they

represent the probability that the word will be associated with only that emotion when encountered in text. A word with three probability weights of 0.333 has an equal chance of expressing those three emotions when written. Likewise, a word with two probability weights of 0.50 has an equal chance of expressing either emotion. A word with a high probability weight may be intensely indicative of that emotion or never associated with any of the other emotions.

Therefore, words with multiple associations are likely penalized when expressed either with probability weights or the polytomous transformation. A word that is highly representative of two emotions will necessarily have lower scores than a word that is highly representative of one emotion. This is in contrast to the other lexicons which use binary scoring that allows for multiple associations without penalty. The polytomous scores may therefore have a non-linear relationship with the other lexicons because higher scores do not necessarily indicate "more" of the underlying latent factor.

Chance Transformation. In the Chance transformation, the DM lexicon was dichotomized based on a "chance" threshold. A word that is associated with no emotions would have equal probability weights for each of the eight emotion variables, that is, each variable would have a value of $1/8 = 0.125$. "Chance" was therefore assigned at a cut-off of 0.125. Scores above chance were assigned a value of 1, while those below chance were assigned 0. This method allowed for multiple emotion associations per word and is reflective of how the DM lexicon is typically applied in practice.

However, this method removed all information regarding the relative sizes of the probability weights. A score of 1 only indicated that it was more likely than "chance" that the word was associated with that emotion. The risk of spurious correlations was also still present, though the binarization may have reduced it. Because a relatively low threshold must be met to be associated with a variable, the conjoint rise and fall of probability weights may be minimized. That is, a word with multiple emotion associations will not be penalized

as much as it would in the Polytomous transformation. This brings the meaning of the DM scores closer in line with those of the other lexicons. In contrast, this lower threshold removes information about relative probabilities and thus may reduce measurement quality.

EmoSenticNet

All multiword phrases were removed from ESN, leaving 5,477 entries (41.5% of its total) before merging with CompLex.

NRC EmoLex

EmoLex's Positive and Negative categories were not used as they are measures of valence, not emotion. No other alterations to EmoLex were performed.

LIWC

All words from DM, ESN, and EmoLex were run through the LIWC program. Any words scored as multiple words or that LIWC indicated was not its dictionary (a score of 0 on its "DIC" variable) were removed and considered not present in LIWC. LIWC raw scores are the percentage of words that match its dictionary per category. Therefore, each single word received a score of either 100(%) or 0(%) on each emotion category; these were transformed into 1 and 0, respectively, to produce dichotomous data. The LIWC words used in CompLex do not represent the totality of the LIWC dictionary but do represent all the words that overlap with at least three of the four lexicons.

5.2 Item Response Models

Traditionally, item response theory models (IRT, IRMs) were developed to model how people responded to questions on a test. One of the earliest and most well-known logistic

models for dichotomous test data was proposed by Rasch (1960). The field of IRT was heavily influenced by Lord & Novick (1968) and further developed by many others including Birnbaum, Wright, and Masters. IRMs are a part of the general linear model family and share many similarities with factor analytic techniques. Multidimensional IRMs can be viewed as the ordinal counterpart to multivariate factor analysis (Cai, 2010; Jöreskog & Sörbom, 1993; Reckase, 1997). The analysis of ordinal data is one of the defining characteristics of IRMs, alongside providing specific fit measures for both persons and items, the ability to equate across test forms, and enhanced generalizability from the sample population (Embretson & Reise, 2000).

Fundamentally, IRMs model the likelihood that a specific, ordinal response pattern is generated from a distribution of items (Schmidt & Embretson, 2012). While IRMs are traditionally used to model how people respond to test items, it is not necessary for the response patterns to be associated with people, nor the items with questionnaires. To analyze the comprehensive lexicon (CompLex), words in the lexicons take the place of "persons" and lexicon variables serve as "items". Each word in CompLex can be viewed as an observation of the set of lexicon emotion variables. The likelihood of a word being associated with a variable is then described by the variable's association with its latent factor(s). The latent factors represent the shared variance between the emotion variables.

The Multidimensional Two Parameter Logistic Model

All IRMs were fit using the `mirt` package in R (Versions 1.35.1 and 4.1.3, respectively, Chalmers, 2012; R Core Team, 2021). Both the confirmatory and exploratory methods relied on multidimensional two parameter logistic models (M-2PLs) when the data was fully dichotomous (i.e., with the Chance DM transformation) and a multidimensional graded response model when there was ordinal data (i.e., with the Polytomous DM transformation) (Birnbaum, 1968; Samejima, 1969). Their multidimensional forms were defined in `mirt` by

Chalmers (2012) and differ only in the inclusion of threshold parameters.

Let x_{ij} represent the associations between a word/person $i = (1, \ldots, N)$ and a lexicon item $j = (1, \ldots, n)$. There are m latent factors $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{im})$. There are two sets of freely estimated item parameters: slope $\boldsymbol{\alpha}_j = (\alpha_{j1}, \ldots, \alpha_{jm})$ and item intercept d_j. The likelihood of a word being associated with a lexicon category is

$$\Phi(x_{ij} = 1|\boldsymbol{\theta}_i, \boldsymbol{\alpha}_j, d_j) = \frac{1}{1 + exp[-D(\boldsymbol{\alpha}_j^T \boldsymbol{\theta}_i + d_j)]}$$

where D is a scaling parameter typically equal to 1.702. For ordinal data, the unique categories for item j are represented by C_j with corresponding intercepts $\boldsymbol{d}_j = (d_1, \ldots, d_{(C_j-1)})$ (Samejima, 1969). The boundary of response probabilities is defined as

$$\Phi(x_{ij} \geq 0|\boldsymbol{\theta}_i, \boldsymbol{\alpha}_j, d_j) = 1,$$

$$\Phi(x_{ij} \geq 1|\boldsymbol{\theta}_i, \boldsymbol{\alpha}_j, d_j) = \frac{1}{1 + exp[-D(\boldsymbol{\alpha}_j^T \boldsymbol{\theta}_i + d_1)]},$$

$$\Phi(x_{ij} \geq 2|\boldsymbol{\theta}_i, \boldsymbol{\alpha}_j, d_j) = \frac{1}{1 + exp[-D(\boldsymbol{\alpha}_j^T \boldsymbol{\theta}_i + d_2)]},$$

$$\ldots$$

$$\Phi(x_{ij} \geq C_{ji}|\boldsymbol{\theta}_i, \boldsymbol{\alpha}_j, d_j) = 0$$

Following, these boundaries create the conditional probability for the response $x_{ij} = k$ to be

$$\Phi(x_{ij} = k|\boldsymbol{\theta}_i, \boldsymbol{\alpha}_j, d_j) = \Phi(x_{ij} \geq k|\boldsymbol{\theta}_i, \boldsymbol{\alpha}_j, d_j) - \Phi(x_{ij} \geq k + 1|\boldsymbol{\theta}_i, \boldsymbol{\alpha}_j, d_j)$$

If Ψ is the collection of all item parameters and expressing the data in indicator form as

$$\chi^{(X_{ij})} = \begin{cases} 1, & \text{if } x_{ij} = 1 \\ 0, & \text{otherwise} \end{cases}$$

the conditional distribution for the ith n x 1 response pattern vector, $\mathbf{x}_i$, is

$$L_l(\mathbf{x}_i|\Psi, \boldsymbol{\theta}) = \prod_{j=1}^{n} \Phi(x_{ij} = 1|\Psi, \boldsymbol{\theta}_i)^{\chi^{(x_{ij})}}$$

Assuming a multivariate normal distributional form $g(\boldsymbol{\theta})$, the marginal distribution is

$$P_l(\Psi|\mathbf{x}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} L_l(\mathbf{x}_i|\Psi, \boldsymbol{\theta}) g(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

where there are m-fold integrals. With $\mathbf{X}$ representing the complete N x n data matrix, the observed likelihood equation is

$$L(\Psi|\mathbf{X}) = \prod_{i=1}^{N} \left[\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} L_l(\mathbf{x}_i|\Psi, \boldsymbol{\theta}) g(\boldsymbol{\theta}) d\boldsymbol{\theta} \right]$$

The Metropolis-Hastings Robbins-Monro (MH-RM) algorithm was used to estimate all models (Cai, 2010). MH-RM is an efficient maximum likelihood estimator that easily handles both confirmatory and exploratory multidimensional IRMs with missing data (Cai, 2010). The variance of latent factors were fixed to 1 for model identification purposes, as is typical for IRMs.

Model Evaluation and Interpretation

One advantage of IRMs is the microscopic level of detail that they provide for evaluating the fit of models, persons, and items. This includes person and item fit, test and item information, empirical curves, local dependence, differential item functioning, and many other fit statistics. Not all of these are necessary or informative for this analysis, especially in

a multidimensional, non-questionnaire setting. Here I will describe what measures were used, and, in particular, how they are interpreted for words and emotion variables. Theses methods of evaluation and interpretation apply to both the confirmatory models and exploratory models.

Overall Model Fit. Nested model comparisons were performed using the Akaike information criterion (AIC), Bayesian information criterion (BIC), and likelihood-ratio χ^2 tests, as appropriate (Akaike, 1974; Schwarz, 1978; Vuong, 1989). Model fit was evaluated using M_2, root mean square error of approximation (RMSEA), standardized root mean square residual (SRMSR), the Tucker-Lewis index (TLI), and the Bentler Comparative Fit Index (CFI) (Bentler, 1990; Browne & Cudeck, 1993; F. F. Chen, 2007; Tucker & Lewis, 1973). Model fit statistics can only be calculated using complete rows (n = 1,794, 25.09% of CompLex). The M_2 statistic is a proxy for chi-squared goodness of fit testing for multidimensional IRMs (Maydeu-Olivares & Joe, 2006).

Universal cut-offs for fit indices are problematic because acceptance and rejection rates can vary widely based on model complexity, sample size, and estimation methods (F. Chen et al., 2008; Marsh et al., 2004; Maydeu-Olivares & Joe, 2014; Xia & Yang, 2019). With a sample size this large, values of RMSEA ≥ 0.05 may indicate mis-specification, though values ≤ 0.05 do not ensure proper specification (F. Chen et al., 2008; Kim & Yoon, 2011; Maydeu-Olivares & Joe, 2014). SRMSR is generally expected to ≤ 0.10, though cut-offs below this provide more accurate indications of model specification (Kline, 2005; Shi et al., 2018). TLI and CFI $\geq .90$ are generally considered acceptable, though many advocate for higher cut-offs of ≥ 0.95 (Sharma et al., 2005; West et al., 2012). Thus, fit statistics in this dissertation are used for model comparison and to evaluate fit, but they are not proof of model fit.

Correlations between factors indicate that their constituent variables share associations with the same words. It is expected that similar emotions (e.g., negatively valenced) would

be correlated as there are words that truly indicate both emotions or that are easy to "mix-up". However, factors may also correlate if there is little distinction between them. Thus, correlations between factors were examined as a form of model fit. Correlations that were excessively high (e.g., $|r| > 0.70$) could indicate that factors should be combined in the confirmatory analysis. Correlations that seemed excessively low, such as between *Anger* and *Disgust* in the confirmatory analysis, could indicate mis-measurement by the lexicons.

Item Parameters and Fit. Under the M-2PL model, the lexicon variables/items received two types of parameters: slope(s) and intercept. Ordinal variables (i.e., when using the Polytomous DM transformation) also have additional threshold parameters; these will not be examined as they are not of great theoretical interest to this dissertation. In contrast to traditional IRT parameterization, `mirt` uses slope-intercept parameterization which is appropriate for estimating multidimensional models and prevents model solutions where parameters converge to infinity.

The interpretation of item slope α is the same as the IRT discrimination parameter a. Higher values indicate stronger association with the latent factor. The relationship between factors and items were explored through both factor loadings and the α parameters. The slope of a given item (α_{ij}) is related to its factor slope (f_{ij}) via the item's uniqueness (u_{ij})

$$\alpha_{ij} = D * \frac{f_{ij}}{\sqrt{u_{ij}}}$$

Each item had a single intercept (d) parameter estimated. It is the probability of an item being associated with a word when $\theta_1 = \theta_2 = \cdots = 0$. Thus, more negative values of d indicate items with relatively infrequent associations with words, while more positive values of d indicate items that are more frequently associated with words. d is largely not influenced by factor associations. The relationship between d and the IRT parameter b in a unidimensional model is expressed as

$$d = -ab$$

Item fit was evaluated using the Signed-χ^2 test ($S - X^2$) and its associated RMSEA (Kang & Chen, 2007; Orlando & Thissen, 2000, 2003). $S - X^2$ Type-I error rate is fairly robust to large sample sizes (Orlando & Thissen, 2000). Poor fit indicates that the pattern of associations between words and the lexicon variable is incongruent with associations found by similar lexicon variables. This can indicate that the lexicon category measures its latent factor poorly, the factor loadings are mis-specified, or because the lexicon category "disagrees" with what words should belong to its underlying latent factor. Item fit can only be calculated using complete rows.

Unmodeled correlations between items, also known as local dependence, is known to bias IRMs (DeMars, 2006). Local dependence can be examined through item residuals. Standardized χ^2 residuals were examined using Cramér's V (W.-H. Chen & Thissen, 1997). The `mirt` package provides a signed version of Cramér's V where the interpretation of the magnitude is the same as the classic version and the sign expresses the direction of the bivariate relationship (Chalmers, 2018). Like many fit statistics, there are not universally agreed upon cutoffs for Cramér's V. Guidelines proposed by Cohen (2013) suggest that Cramér's $V > |0.50|$ is a large effect, Cramér's $V > |0.30|$ is a medium effect, and Cramér's $V > |0.10|$ is a small effect.

Word Parameters and Fit. Factor scores (latent trait ability estimates) are calculated for every factor per word and are on a Z-score metric. Words that are associated with all items within a factor receive higher factor scores. Words that are associated with none or few variables receive lower scores on the corresponding factors. A high trait score indicates that a word is a consistent indicator of that emotion factor among lexicons. While some words may be associated with multiple emotions, most words should not be associated with all factors. A

word that has many high factor scores is associated with many different emotions; this would indicate that the word is a poor distinguisher of emotions. If many words have many high factor scores, this may indicate incorrect model specification. In any case, the distribution of factor scores may lead to insights on the overlap of emotion terminology across categories. Factor scores were estimated using the Bayesian maximum a-posteriori estimation method (MAP). MAP is appropriate for high dimensional models (Chalmers, 2012).

Word fit is a measure of the consistency with which words are associated to lexicon items. Word fit is calculated using the Zh statistic from Drasgow et al. (1985). Zh is a standardized value of lz for ordinal data for both uni- and multidimensional models. Zh values above $|2.00|$ can be used to identify aberrant response patterns, though this is not a strict cut-off (Felt et al., 2017). Low word fit indicates that a word's vector of associations among lexicon items is unlikely, given the fitted model, and is typically more of an issue than high fit. Word fit cannot be calculated with missing data, therefore only complete cases are examined. There is some evidence that calculating person fit with complete cases results in less bias than when using imputed data (Zhang & Walker, 2008).

6 Confirmatory Approach

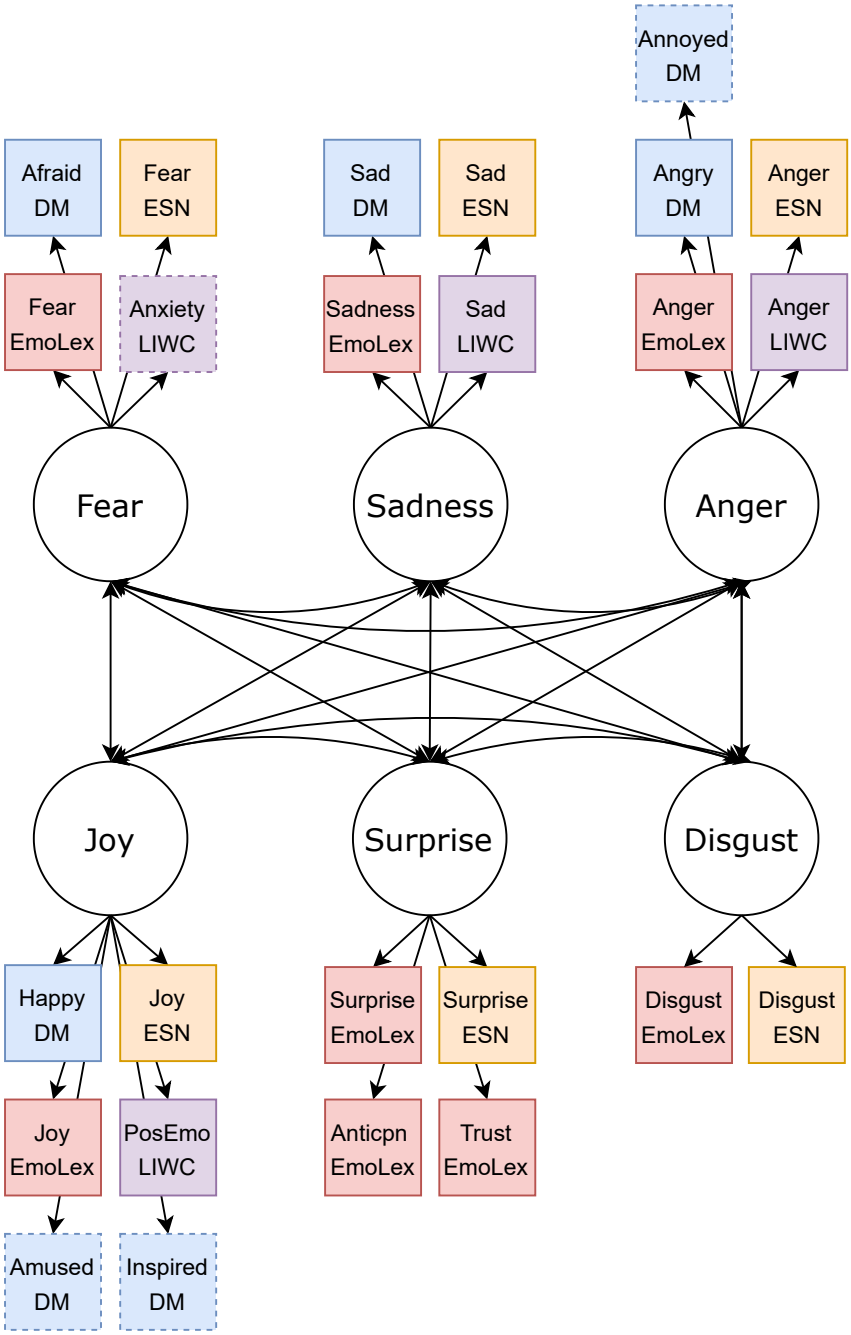
The purpose of the confirmatory approach was to test the assumption that same-named emotion lexicon variables measured the same emotion constructs. This belief is intrinsic in how emotion lexicons are created and used in research. This assumption is tested by imposing a strict hypothesized factor structure onto the lexicons using item response models. If the pattern of word-emotion associations from each lexicon is similar, this should be reflected as shared factor variance. The basis of this method is that the shared association across same-named lexicon variables best reflects emotion constructs as they exist in text. That is, the overlap between all lexicons is the best measurement of that emotion. Lexicons that share more measurement association with the factors thus have stronger evidence for construct validity.

6.1 Methods

The hypothesized factor structure forces same-named lexicon variables to load onto their corresponding discrete emotion factor. An illustration of the hypothesized item response model (IRM) structure can be seen in Figure 6.1. The hypothesized structure had an oblique simple structure where items only loaded onto one factor. There were no cross-loadings allowed, but the factors were free to correlate.

There were four conceptual emotion factors that were easily constructed from four or more lexicon variables. An *Anger* factor was constructed from Angry-DM, Annoyed-DM, Anger-ESN, Anger-EmoLex, and Anger-LIWC. Because *Annoyed* is not typically a distinct emotion used or found in empirical research, I hypothesized that Annoyed-DM measured *Anger*. *Fear* was constructed from Afraid-DM, Fear-ESN, Fear-EmoLex, and Anxiety-LIWC. *Anxiety* is not traditionally a basic or simple emotion. However, I hypothesized that Anxiety-LIWC would be most closely related to the other *Fear* variables

Figure 6.1
Hypothesized Structure of the Confirmatory Model



Note. Dotted lines indicate variables with names that do not directly match the factor but are still hypothesized to belong to that factor.

as Anxiety-LIWC may functioned as "Fear-LIWC". *Sadness* was constructed from Sad-DM, Sad-ESN, Sadness-EmoLex, and Sad-LIWC. *Joy* included Happy-DM, Amused-DM, Inspired-DM, Joy-ESN, Joy-EmoLex, and PosEmo-LIWC.

Because each lexicon used a different set of emotions, there were several variables that did not neatly coalesce into factors. Of these, there were two obvious *Disgust* lexicon variables (Disgust-ESN and Disgust-EmoLex), two obvious *Surprise* variables (Surprise-ESN and Surprise-EmoLex), and three without obvious counterparts (Trust-EmoLex, Anticipation-EmoLex, and Don't Care-DM). As this was a confirmatory-style model constructed on the basis of category overlap between lexicons, removing items with no counterparts would be appropriate. Therefore, Don't Care-DM was not included in the confirmatory analysis as it was hypothesized to have little relation to the other factors. This hypothesis was instead tested in the exploratory section.

However, categories such as Surprise-ESN, Surprise-EmoLex, Trust-EmoLex, and Anticipation-EmoLex had a common thread. While Plutchik separates *Trust*, *Anticipation*, and *Surprise* into separate emotions, Ekman's set only contains *Surprise*. It is possible that these lexicon variables contained similar words related to unexpected experiences, especially because research on the organization of English emotion words often reports just a single *Surprise* cluster (Jack et al., 2016; Jackson et al., 2019). I investigated this by creating a *Surprise* factor constructed of Surprise-ESN, Surprise-EmoLex, Trust-EmoLex, and Anticipation-EmoLex.

Finally, Disgust-ESN and Disgust-EmoLex were placed under a tentative *Disgust* factor. While two items are generally not enough to identify a factor (Tabachnick et al., 2007), this can be overcome if the items are highly correlated with each other (i.e., $|r| > .70$) and have low correlations with other variables (Worthington & Whittaker, 2006). Therefore, this factor would not be kept if their loadings were not sufficient.

Limited model comparisons and adjustments were made to improve model fit. Thus

while I refer to this as a confirmatory analysis, it is not strictly so (Jöreskog & Sörbom, 1993; Worthington & Whittaker, 2006). All changes were limited to those that were in the spirit of the hypothesized structure. That is, all changes were made in pursuit of a simple model structure where same-named variables were related to each other through their single, shared emotion factor. For example, changes in the factor structure of *Anger* were limited to the removal of lexicon categories with low loadings, and cross-loadings between factors were not added. All modifications to the confirmatory model structure were done before the exploratory section, that is, without knowledge of the best fitting exploratory structure.

I began the confirmatory analysis by comparing which DM transformation produced the best fitting IRM. Three multidimensional models were compared using the hypothesized confirmatory structure: one model using the Chance transformation, one model using the Polytomous transformation, and one model that excluded all DM variables. The chosen transformation was then also used in the exploratory section.

Hypotheses

For both the confirmatory and exploratory sections, I hypothesized that the highest "quality" lexicon based upon fit and loadings would be LIWC, followed by EmoLex, DM, and then ESN. LIWC had the most stringent inclusion criteria during its construction; every word was screened for inclusion by emotion researchers. EmoLex was also constructed with statistical oversight into its internal reliability, yet the word-emotion associations were based upon crowd-sourcing. Both DM and ESN were constructed with little researcher oversight or intervention. Therefore, DM and ESN were hypothesized to show the lowest factor loadings and fit due to their more liberal, unsupervised association criteria.

LIWC would likely show the highest item intercepts (difficulties). Even though 74.38% of words in CompLex are in LIWC's dictionary, 79.9% of them are not associated with any emotional category. In comparison, 62.26% of CompLex words from EmoLex have no

emotion association, while all words DM and ESN are associated with at least one emotion variable.

For the confirmatory model in particular, I hypothesized that it would not achieve acceptable fit based on global fit indices. While certain factors with many indicators (e.g., *Anger*, *Sadness*) may have strong loadings, there would likely be many variables that do not align with their same-name factor. If the model did show acceptable global fit, it would be after the removal of poor fitting items. I also expected significant correlations between factors because discrete emotions have been found to be inter-related, and items were not allowed to cross-load onto factors. Larger factor correlations were hypothesized to be found between negatively valenced emotions than between the negatively and positively valenced emotions.

6.2 Results

DepecheMood++ Transformations

Chance vs. Polytomous Results. Model fit for both transformations was extremely poor (Table 6.1). RMSEA and SRMSR were equivalent for both models; as SRMSR and RMSEA values were generally equal in the confirmatory section, only RMSEA will be discussed. TLI and CFI were slightly higher for the Polytomous transformation compared to the Chance transformation ,however neither approached acceptable levels. Both models had M_2 p-values $< .001$, indicating poor fit or misspecification. Information-based fit indices and likelihood

Table 6.1
Model Fit for Chance and Polytomous Transformations

	M_2	*df*	*p*	RMSEA [95% CI]	SRMSR	TLI	CFI
Chance	9545	260	<.001	0.14 [0.14-0.14]	0.14	0.24	0.34
Polytomous	8403	241	<.001	0.14 [0.13-0.14]	0.14	0.30	0.40
No DM	7870	120	<.001	0.19 [0.19-0.19]	0.18	0.09	0.29

ratio-tests could not be used for comparison as each model used a different transformation of the data.

Both models converged but failed the second-order test, indicating that either the model solutions were not a maximum or the information matrices were inaccurate. However, higher-order multidimensional IRMs may also fail the second-order test because of the stochastic process employed by the MHRM estimation algorithm in `mirt` (Paek & Cole, 2020). Unless otherwise noted, all the following models in the confirmatory section converged but failed the second-order test[5].

Five of the six DM variables showed higher loadings onto their same-name factors using the Chance transformation as compared to the Polytomous transformation (Table 6.2). In particular, Amused-DM increased from 0.09 in the Polytomous model to 0.28 in the Chance model, Inspired-DM increased from 0.20 to 0.35, and Sad-DM increased from 0.25 to 0.40. Accordingly, the Chance model had a slightly higher proportion of variance explained (48.53%) than the Polytomous model (46.74%). No factor loadings for non-DM variables showed substantial differences between the two transformations, with the greatest changes seen among Joy-EmoLex (reduced by 0.04 in the Chance model) and Sad-ESN (increased by 0.05 in the Chance model). Both the *Joy* and *Sadness* factors had DM variables with substantially higher loadings in the Chance model. Therefore it is not surprising that the loadings of the other variables in this factor would be influenced by the shift in shared variance.

The overall pattern of factor correlations was broadly similar between the two transformations, though their magnitudes differed (Table 6.3). The negative emotion factors were positively correlated with each other in each model, with no large differences between the two transformations. The negative emotion factors were negatively correlated with

[5]Increasing the number of draws for the MHRM information approximation and calculating the parameter information matrix using post-convergence approximation (FMHRM) can improve model estimation (Paek & Cole, 2020). However, neither of these methods led to the passing of the second-order test here or for any other model in the confirmatory or exploratory sections.

Table 6.2
Factor Loadings of the Chance, Polytomous, and No DM Models

	Chance		Polytomous		No DM	
	λ	h^2	λ	h^2	λ	h^2
Joy Factor						
Amused-DM	0.28	*0.08*	0.09	*0.01*		
Happy-DM	0.31	*0.10*	0.27	*0.07*		
Inspired-DM	0.35	*0.12*	0.20	*0.04*		
Joy-EmoLex	0.83	*0.68*	0.87	*0.76*	0.97	*0.95*
Joy-ESN	0.94	*0.89*	0.92	*0.85*	0.90	*0.80*
PosEmo-LIWC	0.76	*0.58*	0.78	*0.61*	0.76	*0.58*
Surprise Factor						
Anticipation-EmoLex	0.85	*0.72*	0.84	*0.70*	0.82	*0.67*
Surprise-EmoLex	0.70	*0.50*	0.69	*0.48*	0.68	*0.46*
Surprise-ESN	-0.32	*0.10*	-0.32	*0.10*	-0.34	*0.11*
Trust-EmoLex	0.67	*0.45*	0.70	*0.49*	0.71	*0.50*
Anger Factor						
Angry-DM	0.35	*0.12*	0.29	*0.09*		
Annoyed-DM	0.13	*0.02*	0.19	*0.04*		
Anger-EmoLex	0.98	*0.95*	0.99	*0.97*	0.99	*0.99*
Anger-ESN	0.69	*0.47*	0.65	*0.42*	0.66	*0.43*
Anger-LIWC	0.82	*0.67*	0.82	*0.67*	0.80	*0.65*
Fear Factor						
Afraid-DM	0.35	*0.12*	0.23	*0.05*		
Fear-EmoLex	0.95	*0.91*	0.96	*0.92*	0.98	*0.96*
Fear-ESN	0.69	*0.48*	0.65	*0.43*	0.65	*0.42*
Anxiety-LIWC	0.66	*0.44*	0.64	*0.41*	0.64	*0.41*
Sadness Factor						
Sad-DM	0.40	*0.16*	0.25	*0.06*		
Sadness-EmoLex	0.93	*0.86*	0.95	*0.91*	0.99	*0.97*
Sad-ESN	0.81	*0.65*	0.76	*0.58*	0.74	*0.55*
Sad-LIWC	0.84	*0.71*	0.83	*0.70*	0.83	*0.69*
Disgust Factor						
Disgust-EmoLex	0.94	*0.88*	0.93	*0.86*	0.96	*0.91*
Disgust-ESN	0.69	*0.47*	0.68	*0.47*	0.67	*0.45*

Table 6.3
Factor Correlations among Chance, Polytomous, and No DM Models

	Joy	Surprise	Anger	Fear	Sadness
Chance					
Surprise	0.53				
Anger	-0.69	-0.06			
Fear	-0.68	0.01	0.75		
Sadness	-0.75	-0.23	0.60	0.71	
Disgust	-0.68	-0.17	0.75	0.65	0.62
Polytomous					
Surprise	0.60				
Anger	-0.58	-0.01			
Fear	-0.59	0.05	0.78		
Sadness	-0.67	-0.15	0.65	0.73	
Disgust	-0.66	-0.21	0.71	0.61	0.60
No DM					
Surprise	0.70				
Anger	-0.53	-0.09			
Fear	-0.50	0.00	0.75		
Sadness	-0.61	-0.20	0.62	0.69	
Disgust	-0.59	-0.21	0.75	0.63	0.61

the *Joy* factor and had small negative or no correlations with *Surprise*. The correlations between *Joy* and *Fear*, *Anger*, and *Sadness* were greater in magnitude in the Chance model than in the Polytomous model. The correlations between *Surprise* and *Anger*, and *Surprise* and *Sadness* were also greater in magnitude in the Chance model.

No DM Results. The model without the DM lexicon was not an improvement over the other two models. Model fit was still unacceptable and all metrics were worse for the No DM model than either of the other two models (Table 6.1). Factor loadings of non-DM variables were not substantially different from either the Chance or Polytomous models (Table 6.2). The proportion of variance explained by the No DM model was higher (63.89%), though this largely reflects the removal of the DM variables which had low communalities. The change in communalities of the non-DM variables was mixed, with some communalities increasing and others decreasing. The correlation between *Joy* and *Surprise* was higher than in either DM model, and the correlations between *Joy* and the negative emotions were smaller in magnitude (Table 6.3).

Summary of Transformations. Neither the Chance nor the Polytomous transformation showed adequate model fit. In fact, removing DM entirely decreased model fit, suggesting that the DM lexicon was not the primary issue.

DM variables in the Chance model generally had higher communalities than in the Polytomous model. The large increase in explained variance for the two *Joy* DM variables in particular may be due to their associations with each other, and how this is expressed (or masked) in the Polytomous transformation. As mentioned, probability weights may have a non-linear relationship with the emotion factors because the DM variables form a closed set. If a word is associated with all three *Joy* categories, then its associated probability weights would be lower than a word that is only associated with one DM category. In this way, higher scores do not necessarily indicate stronger relationships with the emotion factors. In

contrast, the other lexicons do not penalize multiple associations. Though dichotomiziation is generally less preferred than polytomous binning (and neither are optimal (Altman & Royston, 2006)), the Chance lexicon may have reduced noise present in the Polytomous transformation.

Similar to the factor loading pattern, the factors in the Chance model generally had greater magnitude correlations than in the Polytomous or No DM models. This is most evident in the correlations between the positively and negatively valenced emotions. The pattern of more extreme correlations in the DM models is also found among *Disgust* and *Surprise* - both of which do not contain any DM variables. This suggests that measurement of the *Joy* factor is changing when the DM variables are added. These correlations are inline with theory that positively valenced emotions are negatively correlated with negatively valenced emotions and follow the patterns seen in the No DM model. Thus, though spurious correlations are a known side effect when treating compositional data as real-valued numbers, the correlation patterns seen are likely not entirely driven by statistical artifact.

Based on theses results, the Chance transformation was chosen. Though its global fit was lower, neither transformation showed adequate fit. However, the DM variables showed stronger association with the other lexicons using the Chance transformation. If we assume that DM measures similar emotion constructs as the other lexicons, then the transformation that maximizes these relationships is the most appropriate to use. Because a goal of the confirmatory method is to keep as many lexicon variables as possible and the No DM model fit worse than either DM models, the removal of DM entirely was not considered.

Confirmatory Model Comparisons

After choosing the Chance DM transformation, a series of model comparisons were performed in order to improve the confirmatory model's fit. I will refer to the Chance model created above as the Base model. The Base model follows the original hypothesized model

exactly; it contains six emotion factors onto which all same-named lexicon variables load. However, the fit of the Base model is extremely poor. I proceeded with a series of alterations to the model in order to improve fit, including adding testlets, removing under-identified factors, and removing low-loading variables. Any improvements onto the Base model aimed to preserve the model's discrete emotion structure.

Testlets. The exceptionally poor fit of the Base model suggested that the factor structure fundamentally did not reflect the data. Because the variables come from four separate lexicons, it was possible that the variables within a lexicon were substantially related to each other outside of the emotion factors. Large amounts of local dependence negatively impacts model fit. An examination of the Base model's residuals showed extremely high dependencies between items within the same lexicon. For example, eighteen pairs of variables had signed Cramér's V coefficients greater than $|0.30|$, of which two were above $|0.50|$. All were all among the EmoLex and ESN variables. Though most were within a lexicon, five pairs were between ESN and EmoLex variables. In total, 32% of variable pairs had coefficients above $|0.10|$.

To ameliorate this local dependence, testlet factors were added to the Base model. In IRMs, testlets are factors that account for variance between items that is outside of the latent dimensions of interest (DeMars, 2006). Traditionally, the term "testlet" refers to a group of questions about the same topic on a test, such as a set of reading comprehension items that use the same passage. In this project, each lexicon can be considered a testlet because all items come from the same source and may share variance due to that source. Nuisance testlets in multidimensional IRMs can be modeled using separate latent factors uncorrelated with the dimensions of interest (Eckes & Baghaei, 2015).

Four testlet factors were created, one for each of the four lexicons. Variables from each lexicon loaded onto their associated lexicon testlet and no other testlet. Each testlet was uncorrelated with any other factor in the model and factor loadings were freely estimated.

Table 6.4
Model Fit during Confirmatory Model Comparisons

Model	M_2	*df*	*p*	RMSEA [95% CI]	SRMSR	TLI	CFI	AIC	BIC
Base	9545	260	< .001	0.14 [0.14-0.14]	0.14	0.24	0.34	107945	108392
Testlet	4730	235	< .001	0.10 [0.10-0.11]	0.12	0.59	0.68	103914	104532
$\text{Testlet}_{\text{No LIWC}}$	4573	239	< .001	0.10 [0.10-0.10]	0.12	0.61	0.69	103914	104505
No Disgust	4423	246	< .001	0.10 [0.09-0.10]	0.12	0.64	0.70	104424	104967
Positive Factor	3492	244	< .001	0.09 [0.08-0.09]	0.11	0.72	0.77	103826	104383

Though there had been residual covariance between EmoLex and ESN variables, a correlation between the factors was not estimated because it was not theoretically implied. The means and variances of the testlet factors were fixed to zero and one, respectively, for model identification purposes.

The Testlet model showed a statistically significant improvement from the Base model, $\chi^2(25, 7{,}150) = 4{,}081.13$, $p < .001$. While model fit indices improved, the model still did not meet acceptable standards (Table 6.4). TFI and CFI in particular were still quite poor. The Testlet model explained 65.30% of the total variance, which was an increase of 17 percentage points from the Base model.

The addition of the testlets improved the residual covariances between items. The number of residuals with a standardized Cramér's V above |0.30| fell to six, with none above |0.50|. Five of the six involved Joy-ESN, and one was between Joy-EmoLex and Trust-EmoLex. The percent of residual co-variances above |0.10| in the Testlet model (31%) was not substantially lower than the Base model (32%). The EmoLex and ESN testlets explained the most variance out of all the factors, discrete emotion or testlet, in either the Testlet or Base Model. There was a median increase in h^2 of 0.14 across all variables.

The largest changes in the loadings of the emotion factors was found among the ESN and EmoLex variables where there were median reductions of -0.19 (SD = 0.13) and -0.21 (SD = 0.13) (Table 6.5). Smaller changes in the emotion factor loadings were seen among the DM (Mdn $\Delta\lambda = 0.00$, SD = 0.06) and LIWC variables (Mdn $\Delta\lambda = 0.06$, SD = 0.04).

Table 6.5
Factor Loadings of the Testlet Model

	Discrete Emotion Factors						Testlet Factors				
	Joy	Surpr.	Anger	Fear	Sadn.	Disgust	DM	EMX	ESN	LIWC	h^2
Joy Variables											
Amused-DM	0.22						-0.58				*0.38*
Happy-DM	0.28						-0.37				*0.22*
Inspired-DM	0.32						-0.52				*0.37*
Joy-EmoLex	0.86							0.49			*0.99*
Joy-ESN	0.65								-0.76		*1.00*
PosEmo-LIWC	0.83									0.19	*0.72*
Surprise Variables											
Anticipation-EmoLex		0.65						0.52			*0.69*
Surprise-EmoLex		0.39						0.71			*0.65*
Surprise-ESN		-0.07							0.36		*0.13*
Trust-EmoLex		0.76						0.18			*0.61*
Anger Variables											
Angry-DM			0.40				0.65				*0.58*
Annoyed-DM			0.21				0.20				*0.09*
Anger-EmoLex			0.68					0.72			*0.98*
Anger-ESN			0.48						0.76		*0.81*
Anger-LIWC			0.95							-0.05	*0.90*
Fear Variables											
Afraid-DM				0.43			0.42				*0.36*
Fear-EmoLex				0.65				0.73			*0.95*
Fear-ESN				0.60					0.64		*0.76*
Anxiety-LIWC				0.69						0.20	*0.51*
Sadness Variables											
Sad-DM					0.40		0.35				*0.28*
Sadness-EmoLex					0.71			0.65			*0.93*
Sad-ESN					0.86				0.29		*0.83*
Sad-LIWC					0.91					0.01	*0.82*
Disgust Variables											
Disgust-EmoLex						0.77		0.56			*0.92*
Disgust-ESN						0.53			0.77		*0.87*
Prop. Var. Explained	*0.08*	*0.05*	*0.06*	*0.07*	*0.09*	*0.04*	*0.06*	*0.11*	*0.10*	*0.00*	

Note:
Blank entries indicate parameters that were fixed to zero.

Both the DM and ESN factors seemed reflect valence dimensions where positively valenced emotions loaded with opposite signs to negatively valenced emotions. Loading magnitudes were substantial for both of these factors with many magnitudes between 0.40 and 0.70. Though the DM variables did not shown large changes in their discrete emotion factor loadings, the addition of the DM testlet caused a median increase in their communalities of 0.24. It is unclear if these testlets truly reflect shared variance that is unique to the lexicons, or if they are acting as proxies for emotion dimensions (e.g., valence) or other relationships that are not otherwise captured among the discrete emotion factors.

There was not an obvious interpretable pattern of loadings for the EmoLex or LIWC testlets. All EmoLex variables positively loaded on to the EmoLex testlet. Most loadings for the EmoLex testlet ranged from 0.49 to 0.73, except for Trust-EmoLex ($\lambda = 0.18$). Factor loadings for the LIWC testlet were small, and the testlet explained less than 1% of the total variance. However, the explained variance of the LIWC variables still noticeably increased. The median change in h^2 for the LIWC variables was 0.13, with the largest changes seen in Anger-LIWC ($\Delta h^2 = 0.23$) and PosEmo-LIWC ($\Delta h^2 = 0.14$).

Removing the LIWC testlet did not significantly reduce model fit, $\chi^2(4, 7{,}150) = 8.40$, $p = .078$ (Table 6.4). This indicated that the LIWC variables did not meaningfully co-vary with each other outside of the emotion factors. Therefore, the LIWC testlet factor was removed from the Testlet model for parsimony.

Summary. The testlets had a noticeable positive impact on the confirmatory model. While these testlets were not proposed in the original hypothesized discrete emotion structure, model fit and variable communalities substantially improved from the Base model. Model fit was still poor, however. The LIWC testlet was removed as it contributed little, leaving testlets for DM, ESN, and EmoLex. All variables and all paths between them and the six discrete emotion factors remained. This will be referred to as the $\text{Testlet}_{\text{No LIWC}}$ model.

Disgust Factor. Next, the removal of the *Disgust* factor was investigated in order to improve model fit. With only two variables, the *Disgust* factor in the $\text{Testlet}_{\text{No LIWC}}$ model was under-identified. The *Disgust* factor loadings were also reduced after the addition of the testlet factors, from 0.94 to 0.77 for Disgust-EmoLex and 0.69 to 0.53 for Disgust-ESN. Two methods of removal were investigated. In one, the *Disgust* factor was removed, but the loadings from Disgust-ESN and Disgust-EmoLex to their respective lexicon testlets remained. This is referred to as the No Disgust model. In the other, the *Disgust* factor was removed and Disgust-ESN and Disgust-EmoLex were assigned to the *Anger* factor; their lexicon testlets loadings remained. This is referred to as the Disgust-Anger model.

The $\text{Testlet}_{\text{No LIWC}}$ model fit better than the No Disgust model, $\chi^2(7,\ 7{,}150) = 524.18$, $p < .001$) (Table 6.4). AIC and BIC comparisons also preferred the $\text{Testlet}_{\text{No LIWC}}$ model. However, model fit indices were slightly better for the No Disgust model than the $\text{Testlet}_{\text{No LIWC}}$ model. Overall, RMSEA was still too high, while CFI and TLI were too low. Because including the *Disgust* factor was preferred by the likelihood ratio test and AIC/BIC, and it aligned with the original confirmatory model, the $\text{Testlet}_{\text{No LIWC}}$ model was preferred over the No Disgust model.

The Disgust-Anger model fit worse than No Disgust model based on TLI (.63 vs .64) and CFI (.69 vs .70); both had equivalent RMSEA and SRMSR. Though *Disgust* and *Anger* are closely linked linguistically, the *Disgust* factor did not have a larger correlation with *Anger* ($r = 0.49$) than with the other factors in the $\text{Testlet}_{\text{No LIWC}}$ model as was hypothesized. This may explain why the Disgust-Anger model was not an improvement. Because the *Disgust* factor aligned with the original confirmatory model, the $\text{Testlet}_{\text{No LIWC}}$ model was kept as the best performing model.

Combined Positive Factor. Next, the the *Surprise* Factor was investigated. In the $\text{Testlet}_{\text{No LIWC}}$ model, the *Surprise* factor had a correlation of 0.73 with *Joy*. Combining the two factors into one *Positive* factor improved all aspects of model fit as well as AIC/BIC

Table 6.6
Model Fit when Removing Paths and Variables

Paths Removed	Variables Removed	M_2	*df*	*p*	RMSEA [95% CI]	SRMSR	TLI	CFI	AIC	BIC
Annoyed-DM		3500	245	$< .001$	0.09 [0.08-0.09]	0.11	0.72	0.77	103763	104312
Surprise-ESN, Annoyed-DM		3463	246	$< .001$	0.09 [0.08-0.09]	0.11	0.72	0.77	103763	104306
Happy-DM, Surprise-ESN, Annoyed-DM		3632	247	$< .001$	0.09 [0.08-0.09]	0.11	0.71	0.76	103984	104520
Surprise-ESN, Annoyed-DM	Joy-ESN	2131	224	$< .001$	0.07 [0.07-0.07]	0.10	0.78	0.82	101075	101597
Annoyed-DM	Surprise-ESN, Joy-ESN	2076	202	$< .001$	0.07 [0.07-0.07]	0.10	0.78	0.83	99194	99702

Note:
For 'Paths Removed', only paths between the named variable and their discrete emotion factor were removed. Testlet paths remained.

(Table 6.4). None of the negatively valenced emotion factors had correlations with each other above 0.54, so combining them was not considered.

Variables with Low Emotion Factor Loadings and Communalities. Next, I investigated if removing poorly loading items would improve model fit. The goal was to remove the paths of items that did not related well to their hypothesized emotion factor. With such a large sample size (N = 7,150), smaller factor loadings are less problematic and communalities of less than .40 can be acceptable (Fabrigar & Wegener, 2011; MacCallum et al., 1999). Conversely, large sample sizes can bias likelihood ratio tests towards including all paths in the model even if the paths are functionally zero. I chose to examine items that had communalities and emotion factor loadings less than 0.30. There were three variables that fit this criteria: Annoyed-DM (h^2 = 0.09, λ_{Anger} = 0.19), Surprise-ESN (h^2 = 0.13 , $\lambda_{Positive}$ = -0.04), and Happy-DM (h^2 = 0.22, $\lambda_{Positive}$ = 0.27). Removing these three variables would not leave any factor with less than three indicators.

Starting with the lowest communality variables, I removed the paths between variables and their discrete emotion factors sequentially. Removing the Annoyed-DM path from the Anger factor reduced AIC and BIC while also slightly increasing Annoyed-DM's h^2

(Table 6.6). Though these models were nested, the likelihood ratio test returned an invalid, negative χ^2 value that cannot be compared to the χ^2 distribution. This may be caused by issues with model convergence or because the data were not multivariate normal (Satorra & Bentler, 2010). However, based on the improved information criteria and low variable communality, I removed the path between Annoyed-DM and *Anger* from the model (referred to as $\text{Testlet}_{\text{No A-DM}}$).

Removing the path between Surprise-ESN and the *Positive* factor from $\text{Testlet}_{\text{No A-DM}}$ did not show a statistically significant difference in fit, ($\chi^2(1, 7{,}150) = 2.74$, $p = .098$). AIC and BIC were similar for each model. Therefore, this path was removed (referred to as $\text{Testlet}_{\text{No A-DM or S-ESN}}$).

Removing the path between Happy-DM and the *Positive* factor from the $\text{Testlet}_{\text{No A-DM or S-ESN}}$ model decreased fit, ($\chi^2(1, 7{,}150) = 222.91$, $p < .001$). AIC and BIC were higher for the model without Happy-DM. Therefore, the path between Happy-DM and the *Positive* factor remained.

Only paths belonging to Annoyed-DM and Surprise-ESN and their emotion factors were removed for the new best performing model ($\text{Testlet}_{\text{No A-DM or S-ESN}}$).

Item Fit. After evaluating the factor loadings, the fit of the items was examined via $S - X^2$ and RMSEA. There were only four items that $S - X^2$ indicated had appropriate fit: Fear-ESN, Anger-ESN, Disgust-ESN, and Sad-LIWC. All others were flagged by $S - X^2$ as mis-fitting. As $S - X^2$ is fairly robust to large sample sizes, this mis-fit likely cannot be attributed to test over-sensitivity (Orlando & Thissen, 2000).

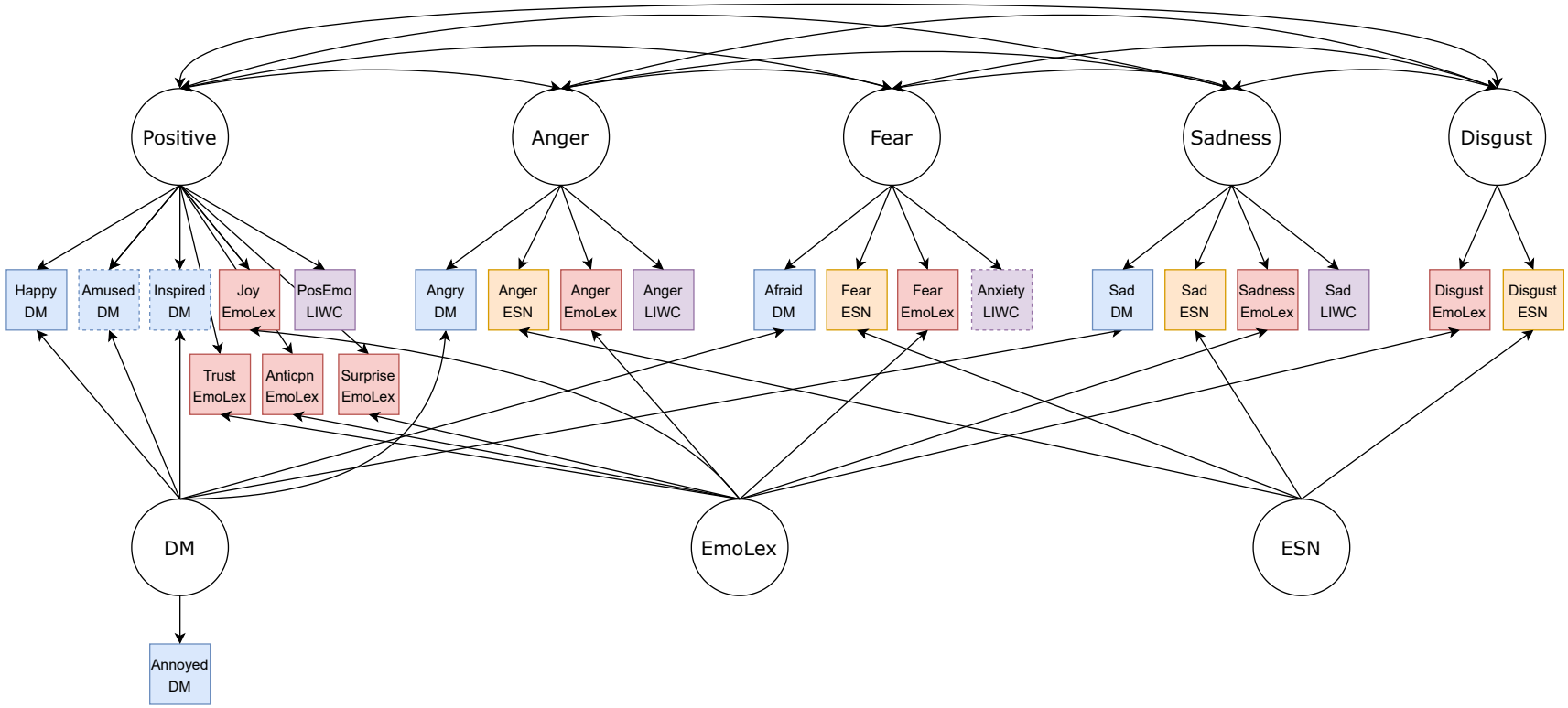
Joy-ESN had particularly poor item fit compared to the other variables, RMSEA = 0.17, and $S - X^2(9) = 454.56$. In addition, all residuals with a standardized Cramér's V above $|0.30|$ involved Joy-ESN; Joy-ESN had unaccounted for negative relationships with Anger-EmoLex, Fear-EmoLex, Sadness-EmoLex, Disgust-EmoLex, and Sad-ESN. Because

of Joy-ESN's particularly poor fit, Joy-ESN was removed from the model entirely. After removing Joy-ESN, there were no residuals with a standardized Cramér's V above $|0.30|$ and overall model fit improved (Table 6.6).

After removing Joy-ESN, h^2 of Surprise-ESN fell to approximately zero. Surprise-ESN now just had a path to the ESN testlet factor. This implied that Surprise-ESN did not share any variance with the remaining ESN variables or the *Positive* variables. Surprise-ESN was thus removed from the model entirely for parsimony. This did not substantially changed the fit statistics of the model nor any of the estimated parameters. This concluded changes to the structure of the confirmatory model.

Figure 6.2

Factor Structure of the Final Model



Note. Dotted lines indicate variables with names that do not directly match the factor but still load onto that factor.

Final Model

The best fitting confirmatory model included lexicon-specific testlets for DM, ESN, and EmoLex, a collapse of the *Joy* and *Surprise* factors into one *Positive* factor, the removal of the Annoyed-DM emotion factor loading, and the removal of the Joy-ESN and Surprise-ESN variables entirely. This will be referred to as the Final model. An illustration of the structure of the Final model can be seen in Figure 6.2. Though the Final model was the best fitting model in the confirmatory section, overall model fit was still quite poor (Table 6.7). No index met acceptable cut-off values.

The Final model explained 64.6% of the total variance. EmoLex variables were very well explained by the model, with median $h^2 = 0.94$ (Table 6.8). ESN and LIWC variables also performed very impressively under the Final model with median $h^2 = 0.78$ and 0.76, respectively. For LIWC, this is particularly noteworthy because there was not an LIWC testlet factor; the communality was entirely due to the discrete emotion factors. The DM variables had the lowest loadings on their discrete emotion factors and the lowest overall median h^2 at 0.27.

The largest loadings on the *Positive* factor came from Joy-EmoLex ($\lambda = 0.91$), and PosEmo-LIWC ($\lambda = 0.82$) (Table 6.8). The three *Surprise* EmoLex variables (Anticipation, Surprise, and Trust) all had higher loadings than the three *Positive* DM variables. The highest loadings on the *Anger* and *Sadness* factors came from Anger-LIWC ($\lambda_{Anger} = 0.95$) and Sad-LIWC ($\lambda_{Sadness} = 0.92$). The *Fear* variables from EmoLex, ESN, and LIWC all had similar sized loadings around 0.68.

The factor correlations of the Final model showed a similar pattern as the Base model,

Table 6.7
Final Model Fit

	M_2	*df*	*p*	RMSEA [95% CI]	SRMSR	TLI	CFI	AIC	BIC
Final Model	2076	202	$< .001$	0.07 [0.07-0.07]	0.10	0.78	0.83	99194	99702

Table 6.8
Factor Loadings of the Final Model

	Discrete Emotion Factors					Testlet Factors			
	Positv.	Anger	Fear	Sadn.	Disgust	DM	EMX	ESN	h^2
Positive Variables									
Amused-DM	0.18					-0.53			*0.31*
Happy-DM	0.22					-0.40			*0.21*
Inspired-DM	0.32					-0.53			*0.38*
Joy-EmoLex	0.91						0.41		*0.99*
PosEmo-LIWC	0.82								*0.68*
Anticipation-EmoLex	0.59						0.56		*0.66*
Surprise-EmoLex	0.36						0.73		*0.67*
Trust-EmoLex	0.69						0.22		*0.52*
Anger Variables									
Angry-DM		0.24				0.77			*0.65*
Annoyed-DM						0.32			*0.10*
Anger-EmoLex		0.71					0.69		*0.98*
Anger-ESN		0.57						0.66	*0.77*
Anger-LIWC		0.95							*0.90*
Fear Variables									
Afraid-DM			0.39			0.34			*0.27*
Fear-EmoLex			0.68				0.71		*0.96*
Fear-ESN			0.69					0.57	*0.79*
Anxiety-LIWC			0.68						*0.46*
Sadness Variables									
Sad-DM				0.42		0.25			*0.24*
Sadness-EmoLex				0.77			0.61		*0.96*
Sad-ESN				0.75				-0.53	*0.85*
Sad-LIWC				0.92					*0.85*
Disgust Variables									
Disgust-EmoLex					0.79		0.54		*0.91*
Disgust-ESN					0.63			0.60	*0.76*
Prop. Var. Explained	*0.12*	*0.07*	*0.08*	*0.10*	*0.04*	*0.07*	*0.12*	*0.06*	

Note:
Blank entries indicate parameters that were fixed to zero.

Table 6.9
Final Model Factor Correlations

	Positive	Fear	Anger	Sadness	Disgust	DM	EMX
Fear	-0.59						
Anger	-0.60	0.49					
Sadness	-0.63	0.57	0.39				
Disgust	-0.53	0.38	0.60	0.43			
DM	0	0	0	0	0		
EMX	0	0	0	0	0	0	
ESN	0	0	0	0	0	0	0

Note:
Factor correlations of zero were fixed to be zero during model estimation.

but with different magnitudes (Table 6.9). The *Positive* factor still had moderate negative correlations with the four negatively valenced emotions. These correlations were generally the same or larger in magnitude than in the original Base model. In comparison, the correlations within the negative emotion factors were smaller than in the Base model; the median inter-factor correlation was 0.46, as compared to 0.68 in the Base model. Overall, the negatively valenced emotions had smaller associations with each other than with the *Positive* factor.

Item Parameters. Item fit was poor overall. There were only two items where $S - X^2$ indicated appropriate fit: Sad-LIWC and Disgust-ESN (Table 6.10). The lowest median RMSEAs were associated with the ESN variables (0.02), followed by LIWC (0.03), DM (0.04), and EmoLex (0.06). Local dependence was still present between items, though in smaller amounts than the previous model iterations. There were no residuals with a standardized signed Cramér's $V > |0.30|$, though 26% of possible variable parings had a Cramér's $V > |0.10|$. Residuals $> |0.10|$ were generally found among groups of emotion variables, for example between the *Disgust* and *Anger* variables.

Table 6.10
Item Fit of the Final Model

	$S - X^2$	*df*	*p*	RMSEA
Positive Variables				
Amused-DM	83.28	10	< .001	0.06
Happy-DM	95.75	10	< .001	0.07
Inspired-DM	77.03	10	< .001	0.06
Joy-EmoLex	104.14	8	< .001	0.08
PosEmo-LIWC	78.28	10	< .001	0.06
Anticipation-EmoLex	58.54	9	< .001	0.06
Surprise-EmoLex	19.34	9	.022	0.03
Trust-EmoLex	107.74	10	< .001	0.07
Anger Variables				
Angry-DM	30.49	11	.001	0.03
Annoyed-DM	46.89	11	< .001	0.04
Anger-EmoLex	90.61	10	< .001	0.07
Anger-ESN	18.03	9	.035	0.02
Anger-LIWC	31.10	8	< .001	0.04
Fear Variables				
Afraid-DM	25.19	10	.005	0.03
Fear-EmoLex	85.60	10	< .001	0.06
Fear-ESN	29.22	9	< .001	0.04
Anxiety-LIWC	22.98	9	.006	0.03
Sad-DM	41.97	10	< .001	0.04
Sadness Variables				
Sadness-EmoLex	58.03	10	< .001	0.05
Sad-ESN	21.98	10	.015	0.03
Sad-LIWC	6.97	8	.540	0.00
Disgust Variables				
Disgust-EmoLex	55.84	10	< .001	0.05
Disgust-ESN	13.08	9	.159	0.02

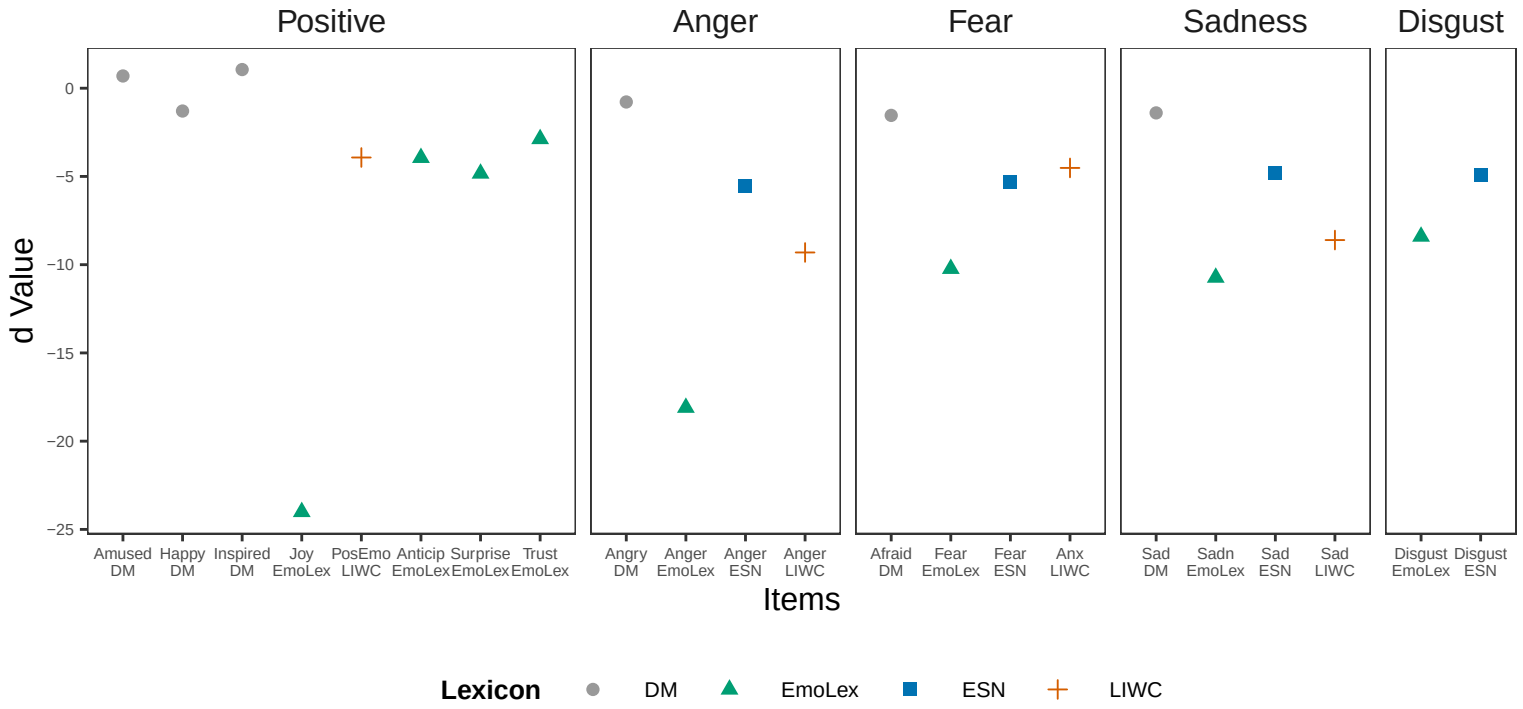
The distribution of item intercepts (d) can be seen in Figure 6.3. Lower item intercepts indicate that an item is less likely to be associated with a word. The median intercepts among each lexicon were: DM = -0.78 (SD = 1.06), ESN = -5.12 (SD = 0.34), EmoLex = -9.32 (SD = 7.35), and LIWC = -6.56 (SD = 2.76). Contrary to my hypothesis, EmoLex had the lowest item intercepts instead of LIWC; LIWC had the second lowest median item intercepts. DM items had the highest intercepts, indicating that words in CompLex were more likely to be associated with DM than with other lexicons. This is not entirely surprising as DM had the highest coverage in CompLex and all words were associated with at least one emotion. The relatively higher intercepts of the *Positive* variables may indicate that most words in CompLex were *Positive* words.

The distribution of item slopes (α) for the emotion factors can be seen in Figure 6.4. Slope parameters were estimated for each factor that an item was associated with. Higher values indicate more discriminant association with the latent factor. Similar to the intercepts, EmoLex variables had the highest emotion factor slopes (Mdn = 5.16, SD = 4.99), followed by LIWC (Mdn = 3.25, SD = 1.56), ESN (Mdn = 2.38 SD = 0.57), and DM (Mdn = 0.69, SD = 0.19). EmoLex variables generally had the strongest ties to their corresponding factor. DM variables were the least aligned with their corresponding factors. The slope parameters for the *Positive* variables were generally lower than those of the other four emotion factors.

Several of Joy-EmoLex's parameters were more extreme than those of the other items. Joy-EmoLex had a very high slope for the *Positive* factor ($\alpha_{Positive}$ = 15.83) and a very low intercept (d = -24.01). Joy-EmoLex's low uniqueness and high association with the *Positive* factor contributes to it's extreme α. Taken all together, while it was relatively rare for words to be associated with Joy-EmoLex, words that were associated with Joy-EmoLex were very likely to also be associated with other *Positive* variables. Joy-EmoLex's testlet factor slope (α = 7.06) was higher than some EmoLex variables, but was not the most extreme (Mdn = 4.12, SD = 3.06).

Figure 6.3

Intercept (d) Parameters for the Final Model



Note. Items are grouped by factor for interpretation, Only one d parameter is estimated per item, regardless of model structure.

Figure 6.4
Slope (α) Parameters for the Final Model

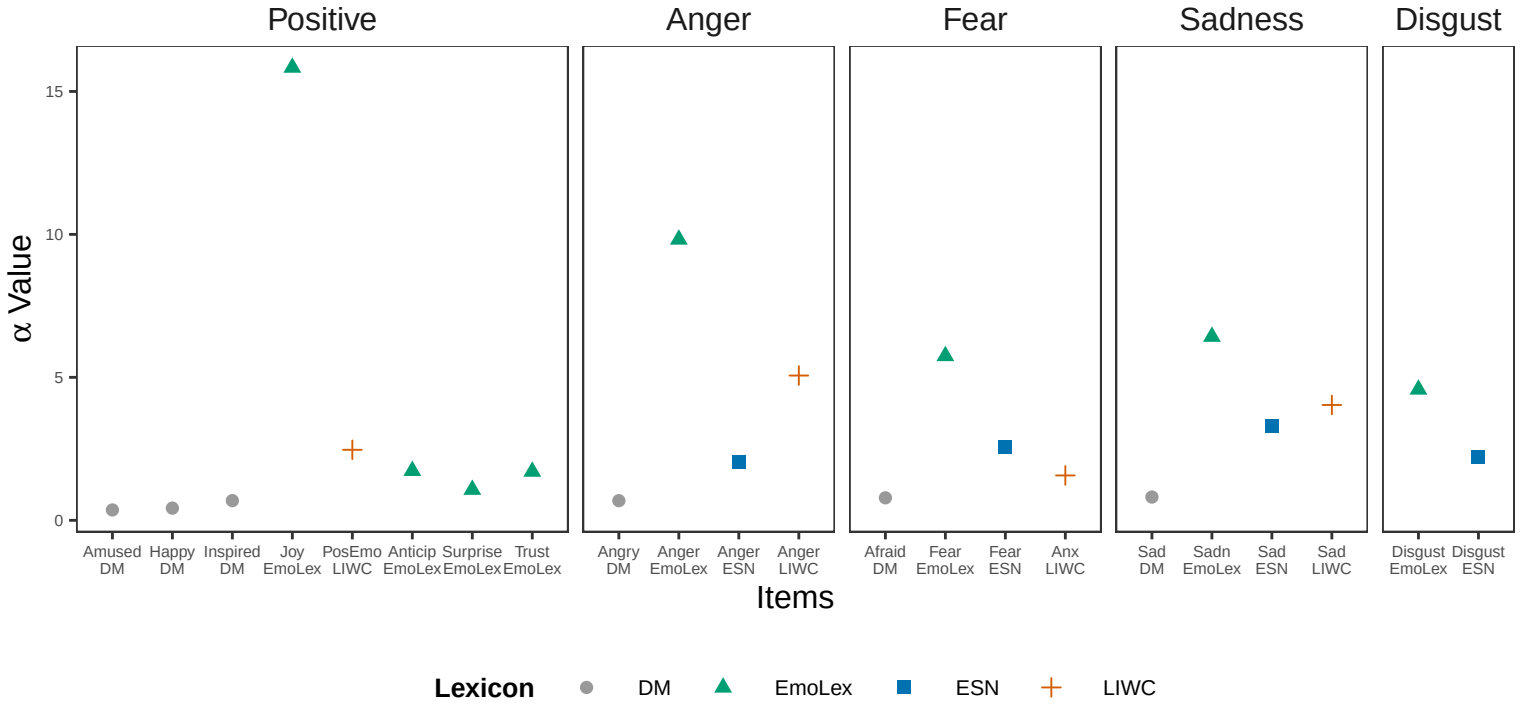
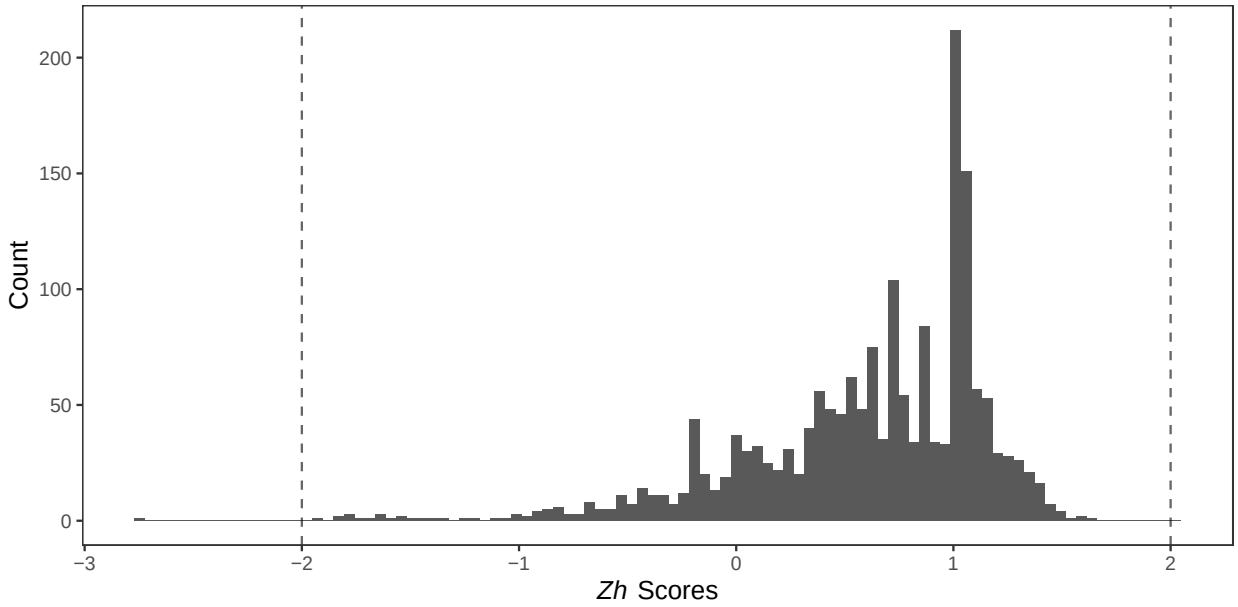


Figure 6.5
Distribution of Word Fit based on Zh



Note. Suggested cut-off lines are drawn at -2 and 2.

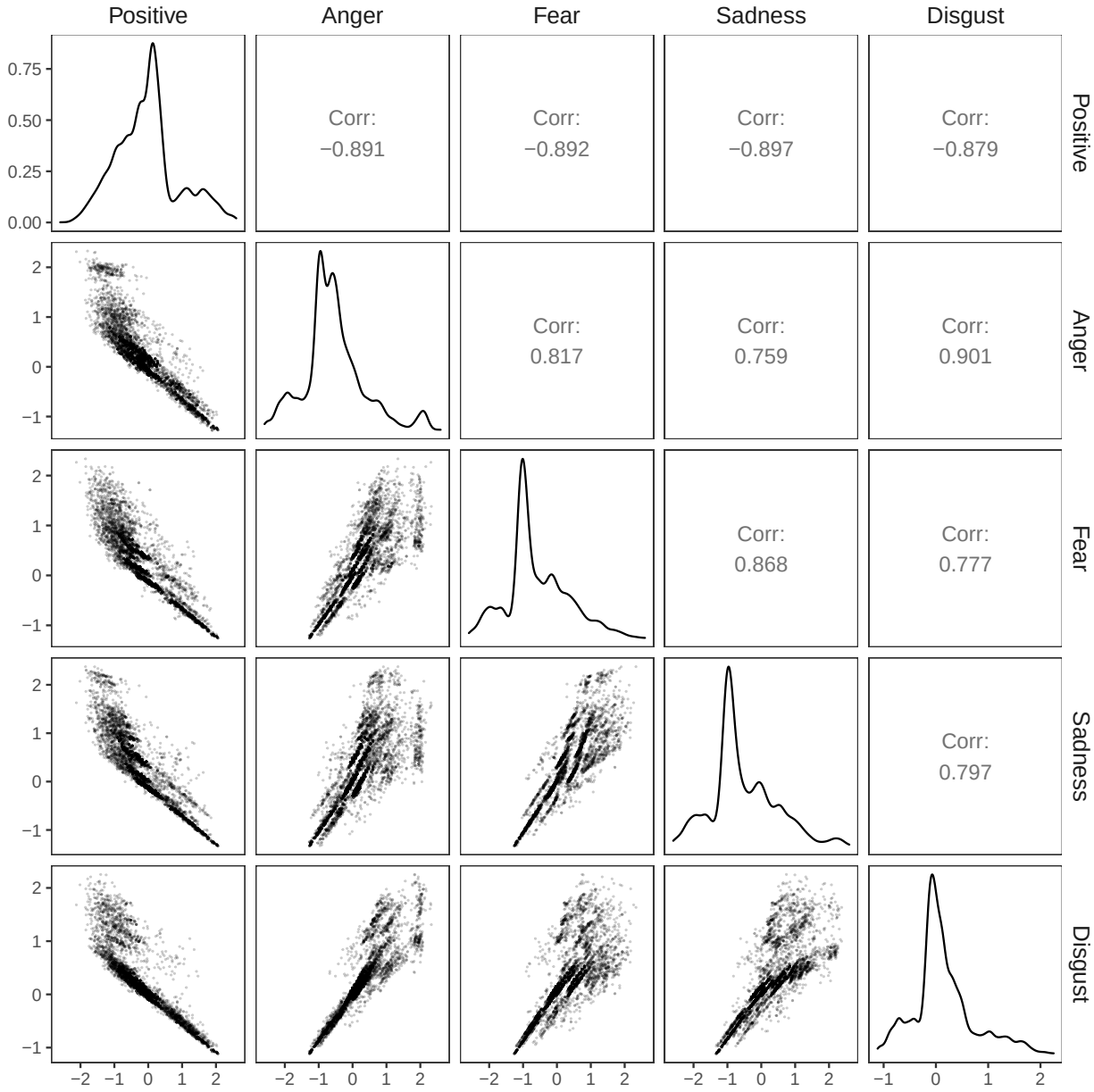
Word Scores and Fit. Most word-emotion association vectors were consistent with the model based on Zh (Figure 6.5). Only one word had a Zh scores < -2.00, indicating mis-fit. No words had Zh scores > 2.00, which is associated with over-fit. The low rate of extreme Zh scores reflects positively on the Final model, but the global model fit was still very poor.

In the Final model, every word received a factor score for each of the five emotion dimensions and the three testlet dimensions[6]. Factor scores were estimated using MAP, as is appropriate for multidimensional models, and are interpreted akin to standardized Z-scores (Chalmers, 2012). Figure 6.6 shows the distribution of the factor scores for the emotion factors and the factor scores' relationships to each other. For identification purposes during model estimation, the mean of each factor was fixed to zero and the standard deviations to one. The largest spread was found in the distribution of *Positive* emotion scores (SD = 0.74) as well as the most extreme negative scores (minimum = -2.12). The SDs of the distributions of *Anger*, *Fear*, *Sadness*, and *Disgust* were smaller (0.57 < SD's < 0.67) and

[6]Factor scores failed to converge for four words: "blossom", "determinate", "fondness", and "pastry".

Figure 6.6

Relationships Between Emotion Factor Scores



Note. Lower plots are scatterplots of each pair of factor scores, along the diagonal are the densities of each factor score, while upper plots show the Spearman correlation between the factor score pairs.

had less extreme minimums between -1.33 and -1.12. The distributions for the negatively valence emotions had heavier positive tails.

The words with the most extreme scores per factor can be seen in Table 6.11. The highest scoring words of each factor clearly support the factor labels. The *Positive* factor featured words like "enjoy" and "happy", while *Fear* featured words like "tense" and "terrorism". The top scoring words for *Disgust* suggested that Disgust-EmoLex and Disgust-ESN contained terms relating to moral disgust ("idiot", "liar") and core disgust ("crap", "poisoning"). The top *Disgust* words also had high *Anger* scores, reflecting the close linguistic connection between these two emotions. It should be noted that factor scores indicate the likelihood that the word is associated with that emotion; they do not necessarily signify intensity. For example, "misplace" is one of the top scoring words for *Anger* likely due to the phrase "misplaced anger" and not just because misplacing an object is so intensely infuriating.

While the top words were generally unique per factor, the lowest scoring words were not. The top scoring *Positive* words also received the lowest scores for each of the four negative emotions. Accordingly, the correlations between the *Positive* factor scores and those of the four negative emotions were all $\leq$ -0.88 (Figure 6.6). It is very clear that words could either be positively or negatively valenced in the Final model. This likely contributes to the heavy negative tail of the *Positive* score distribution.

Of note, there were substantial differences between the factor correlations estimated by the model and the correlations between the estimated factor scores. While the general pattern among the correlations between the estimated emotion factor scores and the model calculated factor correlations was similar, the magnitudes were twice as large among the estimated factor scores. These differences may indicate that there were issues with model specification or in the estimation of the factor scores (Grice, 2001). This may be a results of the discrete nature of the lexicons; only two of the lexicons (DM and EmoLex) allowed for substantial levels of multiple word-emotion associations. The scatter plots in Figure 6.6

Table 6.11

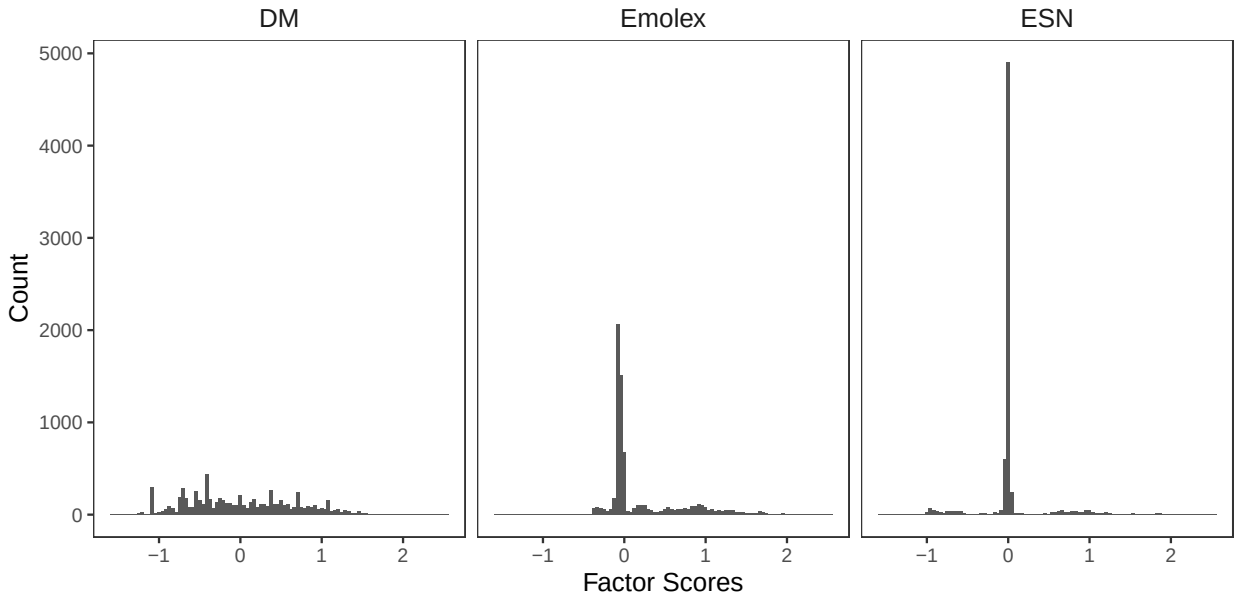
Representative Words of Each Emotion Factor

Words with the Highest Factor Scores						Words with the Lowest Factor Scores					
	Factor Scores						Factor Scores				
	Positive	Anger	Fear	Sadn.	Disgust		Positive	Anger	Fear	Sadn.	Disgust
Positive Factor											
Champion	2.04	-1.26	-1.24	-1.32	-1.11	Destroyer	-2.12	2.31	2.27	1.63	1.68
Compliment	2.06	-1.27	-1.26	-1.33	-1.12	Kill	-1.84	1.74	2.05	1.86	1.21
Enjoy	2.04	-1.26	-1.24	-1.32	-1.10	Misery	-1.85	1.34	2.34	2.38	1.63
Flirt	2.06	-1.27	-1.26	-1.33	-1.12	Ruinous	-1.83	1.28	1.66	2.30	1.60
Happy	2.04	-1.26	-1.24	-1.32	-1.10	Sterile	-1.83	0.97	1.59	2.30	1.49
Improve	2.04	-1.26	-1.24	-1.32	-1.11	Threatening	-1.84	2.16	2.18	1.22	1.98
Magnificent	2.06	-1.27	-1.26	-1.33	-1.11	Traitor	-2.01	2.11	1.86	2.07	2.06
Perfection	2.06	-1.27	-1.26	-1.33	-1.12						
Anger Factor											
Criticize	-1.59	2.29	1.24	1.52	1.61	Bless	1.89	-1.28	-1.20	-1.25	-1.08
Destroyer	-2.12	2.31	2.27	1.63	1.68	Blessing	1.89	-1.28	-1.20	-1.25	-1.07
Destructive	-1.63	2.15	1.36	0.81	1.58	Compliment	2.06	-1.27	-1.26	-1.33	-1.12
Extinguish	-1.79	2.33	1.38	1.58	1.31	Faith	1.89	-1.28	-1.20	-1.25	-1.08
Injustice	-1.56	2.30	1.02	1.55	1.71	Flirt	2.06	-1.27	-1.26	-1.33	-1.12
Misplace	-1.52	2.17	1.05	1.43	1.94	Magnificent	2.06	-1.27	-1.26	-1.33	-1.11
Threatening	-1.84	2.16	2.18	1.22	1.98	Perfection	2.06	-1.27	-1.26	-1.33	-1.12
Fear Factor											
Battlefield	-1.46	1.80	2.17	0.91	0.99	Champion	2.04	-1.26	-1.24	-1.32	-1.11
Destroyer	-2.12	2.31	2.27	1.63	1.68	Compliment	2.06	-1.27	-1.26	-1.33	-1.12
Insecurity	-1.07	0.91	2.16	1.06	0.60	Enjoy	2.04	-1.26	-1.24	-1.32	-1.10
Misery	-1.85	1.34	2.34	2.38	1.63	Flirt	2.06	-1.27	-1.26	-1.33	-1.12
Tense	-1.27	1.04	2.18	1.52	0.82	Happy	2.04	-1.26	-1.24	-1.32	-1.10
Terrorism	-1.68	1.54	2.17	1.42	1.33	Improve	2.04	-1.26	-1.24	-1.32	-1.11
Threatening	-1.84	2.16	2.18	1.22	1.98	Magnificent	2.06	-1.27	-1.26	-1.33	-1.11
						Perfection	2.06	-1.27	-1.26	-1.33	-1.12
Sadness Factor											
Abandon	-1.38	0.81	1.39	2.30	1.23	Champion	2.04	-1.26	-1.24	-1.32	-1.11
Deprivation	-1.67	1.24	1.61	2.27	1.57	Compliment	2.06	-1.27	-1.26	-1.33	-1.12
Inadequate	-1.48	1.04	1.85	2.34	1.03	Flirt	2.06	-1.27	-1.26	-1.33	-1.12
Misery	-1.85	1.34	2.34	2.38	1.63	Improve	2.04	-1.26	-1.24	-1.32	-1.11
Overwhelmed	-1.58	0.91	1.90	2.37	1.00	Magnificent	2.06	-1.27	-1.26	-1.33	-1.11
Ruinous	-1.83	1.28	1.66	2.30	1.60	Perfection	2.06	-1.27	-1.26	-1.33	-1.12
Sterile	-1.83	0.97	1.59	2.30	1.49	Succeeding	2.02	-0.96	-1.20	-1.33	-0.97
Disgust Factor											
Crap	-1.24	1.83	0.78	0.99	2.24	Champion	2.04	-1.26	-1.24	-1.32	-1.11
Humiliating	-1.79	1.93	1.63	1.35	2.17	Compliment	2.06	-1.27	-1.26	-1.33	-1.12
Idiot	-1.40	1.84	0.75	0.74	2.25	Enjoy	2.04	-1.26	-1.24	-1.32	-1.10
Liar	-1.48	1.88	0.92	1.18	2.11	Flirt	2.06	-1.27	-1.26	-1.33	-1.12
Mockery	-1.44	1.85	0.86	0.85	2.11	Happy	2.04	-1.26	-1.24	-1.32	-1.10
Poisoning	-1.70	1.94	1.34	1.31	2.21	Improve	2.04	-1.26	-1.24	-1.32	-1.11
Wench	-1.44	1.56	0.75	1.13	2.13	Magnificent	2.06	-1.27	-1.26	-1.33	-1.11
						Perfection	2.06	-1.27	-1.26	-1.33	-1.12

Note:

Shown are seven words associated with the most extreme scores per factor. Anger and Sadness have eight lowest words shown due to ties in factor scores, and Positive has eight highest words. Factor scores are interpreted akin to z-scores.

Figure 6.7
Distributions of Lexicon Testlet Factor Scores



reflect what was seen from the qualitative examination of the most extreme factor scores: there are near perfect correlations at the lower end of each pair of factor scores. In contrast, there is more diversity among the highest scoring words on the negative factors, forming cone shaped scatter plot distributions.

Though the testlet factors were not of substantive interest, their distributions reflected basic information about the lexicons. The majority of words in EmoLex are not associated with any emotion. Correspondingly, most words have factor scores around zero (Figure 6.7). The same pattern is seen in the distribution of ESN variables, but instead because Joy-ESN and Surprise-ESN were removed. Words relating to these variables would have zeros on all remaining ESN variables. In contrast, every word in the non-transformed DM lexicon received non-zero scores on all of the eight emotions. Thus, the DM testlet factor scores had a more uniform distribution with no peak at zero. While the testlet factors were orthogonal in the Final model, there were some small correlations between their estimated factor scores and the emotion factor scores (Mdn = 0.12, min = -0.16, max = 0.19). There was no obvious pattern in these correlations.

Table 6.12
Unidimensional Model Fit

	M_2	*df*	*p*	RMSEA [95% CI]	SRMSR	TLI	CFI
Positive	64	9	< .001	0.06 [0.05-0.07]	0.04	0.88	0.93
Anger	155	5	< .001	0.13 [0.11-0.15]	0.09	0.60	0.80
Fear	1	2	.72	0.00 [0.00-0.03]	0.05	1.00	1.00
Sadness	1	2	.66	0.00 [0.00-0.04]	0.05	1.00	1.00

Unidimensional Models

As a final exploration of the confirmatory structure, I ran separate unidimensional models. The purpose of these models was to provide a brief and broad overview of how each discrete emotion factor functioned on its own because the fit of the Final model was so poor. Separate unidimensional models were run for the *Positive* variables[7], *Anger* variables, *Fear* variables, and *Sadness* variables. Each of these unidimensional models passed the second-order test; previously, none of the multidimensional models passed the second-order test.

The *Fear* and *Sadness* models had near perfect fit across all indices, while the *Positive* model had middling fit (Table 6.12). The *Anger* model performed the worst with results comparable to the multidimensional models. The *Sadness* model had the highest communalities, while the communalities for many of the other variables was quite low. Factor loadings were quite similar for most variables in both their unidimensional model and in the Final model. This suggests that while one of the core issues in the Final Model was the improperly modeled relationships between emotion groups (e.g., enforced simple structure), some of the variables still may not be closely related to each other.

[7]The unidimensional model with only *Joy* variables did not pass the second-order test. Therefore, all positive variables were combined, as in the Final model.

Table 6.13
Factor Loadings of the Unidimensional Models

	λ	h^2
Positive Model		
Amused-DM	0.18	0.03
Happy-DM	0.23	0.05
Inspired-DM	0.34	0.12
Joy-EmoLex	1.00	0.99
Joy-ESN	0.55	0.31
PosEmo-LIWC	0.74	0.55
Anticipation-EmoLex	0.75	0.57
Surprise-EmoLex	0.63	0.40
Surprise-ESN	-0.24	0.06
Trust-EmoLex	0.71	0.50
Anger Model		
Angry-DM	0.44	0.19
Annoyed-DM	0.33	0.11
Anger-EmoLex	0.80	0.64
Anger-ESN	0.60	0.36
Anger-LIWC	0.93	0.87
Fear Model		
Afraid-DM	0.37	0.14
Fear-EmoLex	0.74	0.55
Fear-ESN	0.76	0.58
Anxiety-LIWC	0.74	0.55
Sadness Model		
Sad-DM	0.35	0.12
Sadness-EmoLex	0.88	0.77
Sad-ESN	0.78	0.61
Sad-LIWC	0.93	0.86

6.3 Discussion

The hypothesized confirmatory model structure did not reflect the relationships present among the lexicons. Significant changes were made to the original hypothesized model: lexicon specific testlet factors were added, the *Joy* and *Surprise* factors were combined, two variables were removed entirely (Joy-ESN and Surprise-ESN), and one variable lost its path to its assigned emotion factor (Annoyed-DM to *Anger*). However, model fit was still poor even with these dramatic changes. Two broad conclusions can be seen: simple structure is likely inappropriate for the inter- and intra-lexicon relationships, and DM had a weak relationship to the other lexicons.

Model Structure

Considering the extremely poor model fit, it is clear that discrete emotion factors with simple structure does not accurately represent the lexical relationships. The lack of cross-loadings is likely a significant factor in the poor model fit. Simple structure is a strong assumption here, especially considering that the variables came from different lexicons and that discrete emotions are actually inter-related.

The addition of the testlets made the single largest improvement to the fit of the confirmatory model. One significant drawback is that they made the interpretation of the individual factors harder. It is not immediately obvious what the testlets measure, besides variance that is specific to the individual lexicons. However, it is quite obvious that without the testlets, the confirmatory model would have fit far worse.

A different factor structure may have been appropriate. When examining the factor scores, it is clear that words can either be associated with the *Positive* factor, the negative emotions, or none at all. Perhaps a bi-factor structure where there is a single valence factor for all variables alongside the individual discrete emotion factors would better approximate

the data. The need for a valence factor is supported by the DM and ESN testlet factors; both seemed to represent valence.

Emotion Factors

Despite the poor model fit, there is evidence that same-named emotion variables did measure similar constructs. Most variables from EmoLex, ESN, and LIWC had positive medium or large loadings on their associated factor. Half of the unidimensional models also showed appropriate fit, though inter-lexicon agreement was higher for some emotions (*Sadness*) and poorer for others (*Anger*). Thus, if emotions are viewed separately, as they are in most analyses, some level of similar measurement would be expected. However, I still would not assert that all lexicon measures are interchangeable. I will not speak at length on the interpretation of the emotion factors as the model fit was so poor and the exploratory model reveals more; I will touch on notable results, though.

I had originally hypothesized that there would be a close connection between *Anger* and *Disgust*. This hypothesis was supported in some ways, but not others. As mentioned, *Anger* and *Disgust* are closely related linguistically. Yet of the four lexicons, only EmoLex and ESN had separate *Disgust* variables; it was not clear if DM and LIWC included *Disgust* terms in their *Anger* measures. Despite their close linguistic connection, the factor correlation between *Anger* and *Disgust* was not particularly different from the other variables. Further, though the model fit best with a *Disgust* factor, removing the *Disgust* factor entirely was better than placing the *Disgust* variables under the *Anger* factor. This suggests that there is still a substantial difference between the *Anger* variables and the *Disgust* variables. Yet, the estimated factor scores of *Anger* and *Disgust* were highly correlated to each other. This is evidence tempered, though, as all the estimated factor scores were highly correlated to each other and the global model fit was particularly poor.

Notably, the *Surprise* variables seemed to have just as much in common with the

Joy variables than with themselves. Even with such a large sample size, there was not a substantial difference when the *Surprise* and *Joy* factors were combined into *Positive* or kept separated. This connection is not entirely surprising as *Surprise* is often considered a positive emotion in ED research (Mohammad & Turney, 2013; Poria et al., 2013b). The combination of *Surprise* and *Joy* together may reflect the similarities found between positive words/emotions (Jack et al., 2016; Rozin et al., 2010; Rozin & Royzman, 2001; Schrauf & Sanchez, 2004), or poor differentiation of constructs between the lexicons.

However, it may also be caused by the lack of *Surprise* categories in the DM and LIWC lexicons. As described earlier, when lexicons do not contain the same categories, the meaning of the overlapping categories can change. It could be that PosEmo-LIWC and the positive DM variables contain elements of *Surprise* that are split off separately in ESN and EmoLex. The Final model did not provide much clarity into the differences among the *Positive* variables. Regardless, it calls into question how many different positively valenced categories should be included in an emotion lexicon.

Lexicons

DepecheMood++. As hypothesized, DM performed particularly poorly, regardless of the transformation chosen. The smallest communalities were almost all found among the DM variables. DM did not seem to measure the same constructs as the other lexicons. In fact, Annoyed-DM did not seem to measure *Anger* at all, contrary to my hypothesis.

It is difficult to pinpoint the exact issue with DM. Reader ratings of news articles may simply not generalize, or be too noisy. However, its compositional nature likely also contributed to its low associations. For example, the closed scoring of the raw DM may have censored any one *Positive* variable's true relationship with the latent factor because there were three *Positive* variables competing against each other. In support of this, the DM variables with the highest communalities were the only DM variables on their factors.

It is also possible that the threshold used in the Chance transformation was too low and introduced too much noise. Regardless, there is little evidence to support the use of the DM lexicon for text analysis.

NRC EmoLex. EmoLex was one of the better lexicons, as I had hypothesized. Besides the *Surprise* variables, all EmoLex variables had factor loadings between 0.68 and 0.91 on their emotion factors. EmoLex also had the highest slope (α) parameters for these same variables, indicating better discrimination. That is, when words are associated with EmoLex's emotion variables, they are likely to be associated with other same-named emotions.

The measurement capabilities of Surprise-EmoLex, Trust-EmoLex, and Anticipation-EmoLex were ambiguous, however. They were quite highly related to each other and the *Joy* variables. Because they did not have obvious counterparts in other lexicons, aside from Surprise-ESN which was removed, I cannot conclude from this analysis what constructs they measure. Whether they measure distinct constructs of their own is yet to be determined, though their positive valence is very obvious.

EmoSenticNet. In line with my hypotheses, ESN was not as "good" a lexicon as EmoLex or LIWC, though it performed better than DM. Two of the ESN variables were removed from the model entirely. The four remaining ESN variables (Anger, Fear, Sad, and Disgust) generally performed well. They typically tied in loading strength to the EmoLex variables, though their factor discrimination (slopes) were lower than both EmoLex and LIWC.

Joy-ESN was a particularly curious variable. It had a *Positive* factor loading comparable to the those of Surprise-EmoLex and Trust-EmoLex, but far below that of Joy-EmoLex and PosEmo-LIWC. It seems that while Joy-ESN did measure some useful degree of *Joy* or *Positive*-ness, it was not closely related to Joy-EmoLex and PosEmo-LIWC. Joy-ESN may have closely measured valence as it had large, negative local dependencies with several negative ESN and EmoLex variables.

After the removal of Joy-ESN, Surprise-ESN had no shared variance with any other variable. In the unidimensional *Positive* model, Surprise-ESN had a negative factor loading. This suggests that Surprise-ESN does not measure positively valenced surprise, as did Surprise-EmoLex or Anticipation-EmoLex. A sample of the words within Surprise-ESN does not provide clarity on what it measures; for example, Surprise-ESN contains words like "lesbian", "proxy", "gospel", and "migraine", alongside more obvious *Surprise* words like "strange", "abrupt", and "unpredictable".

LIWC. In line with my hypotheses, LIWC variables had some of the highest loadings on each emotion factor. It is also telling that there was no need for an LIWC lexicon testlet; there was no meaningful covariance between the LIWC variables outside of the emotion factors.

As hypothesized, Anxiety-LIWC does seem to measure *Fear* as well as Fear-EmoLex and Fear-ESN. Each of the *Fear* variables had the same loadings onto the *Fear* factor in the Final model (0.68) and in the unidimensional *Fear* model (≈ 0.75). However, they only shared about half their variance with each other, based on the pattern of communalities in the Final mode and the unidimensional model. Thus, the relationships between Anxiety-LIWC and *Fear* is not completely settled.

7 Exploratory Analysis

To the best of my knowledge, it is currently unknown how emotion lexicons relate to each other outside of head-to-head classification and prediction competitions. The exploratory method of this dissertation sought to fill this gap by examining the innate relationships within and among emotion lexicons through latent factor modeling. Unlike the confirmatory analysis, this analysis was not constrained by a hypothetical structure. Instead, the best fitting item response model (IRM) structure was sought regardless of how the lexicon variables are believed to be related. First, the number of factors in the data was estimated using Exploratory Graph Analysis, then a series of IRMs were run based on this information, and finally the best fitting model was examined using two different rotations.

7.1 Methods

The number of dimensions within CompLex was estimated using Exploratory Graph Analysis (EGA) via the EGAnet package in R (Version 1.0.0, Golino & Christensen, 2021). EGA is a network-based method of identifying the dimensional structure of a dataset (Golino & Epskamp, 2017). EGA has high accuracy in recovering the number of factors underlying a dichotomous dataset and may be especially appropriate for CompLex considering the large sample size, anticipated number of factors per variable, and anticipated high inter-factor correlations (Golino et al., 2020). Two EGA estimation methods, the Gaussian graphical model (GGM, Lauritzen, 1996) and the Triangulated Maximally Filtered Graph (TMFG, Christensen et al., 2019; Massara et al., 2017), were compared using relative fit statistics (RMSEA, SRMSR, TLI, CFI) and the total entropy fit index with Von Neumman entropy (Golino et al., 2021). When both methods agree, there is a high likelihood that the identified structure is accurate (Golino et al., 2020). I hypothesized that the GGM method would have better fit based upon the simulations in Golino et al. (2020).

I then fit a series of exploratory multidimensional two parameter logistic models (M-2PL, exploratory IRMs) using the results of the EGA as a starting point for the number of factors. Technical information on the M-2PL can be found in the General Methods section. Because this analysis was not constrained by a hypothetical structure, poorly performing items were able to be freely removed to improve model fit. I continued to use the binary Chance transformation of the DM lexicon as I did in the confirmatory section.

Two different rotations were examined: oblimin and bifactor. An oblimin rotation attempts to separate the data into an oblique, simple structure. Based on the results of the confirmatory section and on past research that found separate but correlated emotion dimensions, I chose to begin with an oblimin rotation for model selection (Jack et al., 2016; Mohammad, 2018; Warriner et al., 2013).

After the final exploratory model was chosen, it was also examined using a bifactor rotation. In a bifactor model, each variable loads onto a general factor and one group factor. For a model with three factors, the factor loading matrix would be

$$\Lambda = \begin{bmatrix} * & 0 & * \\ * & 0 & * \\ * & 0 & * \\ * & * & 0 \\ * & * & 0 \\ * & * & 0 \end{bmatrix}$$

Exploratory bifactor rotations approximate this general and group factor structure, allowing all variables to load onto the general factor and encouraging perfect cluster structure for the group factors (Jennrich & Bentler, 2011). Both orthogonal and oblique exploratory bifactor rotations are available in `mirt` (Jennrich & Bentler, 2011, 2012). I hypothesized that the general factor would be interpretable as a valence factor because valence is one of

the most basic dimensions found in empirical research (e.g., Fontaine et al., 2007) and it seemed to be a significant structural component of the confirmatory results. The bifactor rotation may not produce true valence, arousal, and dominance dimensions, though, because it encourages simple structure in the group factors. However, general factors are otherwise quite rare to identify without a rotation that specifically fits one (Gorsuch, 2014).

Exploratory Hypotheses

As in the confirmatory analyses, I hypothesized that the highest "quality" lexicon based upon fit and loadings would be LIWC, followed by EmoLex, DM, and ESN. LIWC would likely show the highest factor associations and intercepts as it had more stringent inclusion criteria during its construction for what words could be associated with what emotions. Again, DM and ESN were hypothesized to show the lowest factor loadings and fit due to their more liberal, unsupervised association criteria.

Specific to the exploratory section, I hypothesized that the best fitting model would not align closely with the hypothesized confirmatory model. Further, fit would be significantly better for the exploratory model than the confirmatory model because the confirmatory model was so restrictive. One major distinction between the exploratory and confirmatory models was that the exploratory model allowed items to load onto multiple factors. Substantial cross-loadings were hypothesized as many words are actually associated with multiple emotions - especially for *Anger* and *Disgust*.

I hypothesized that factors in the exploratory model would likely not represent clean divisions of discrete emotions. Because lexicon variables were able to freely associate with each other in the exploratory model, the factors may represent either discrete or dimensional structures. That is, same-named emotion variables may all load strongly onto a single factor, representing a discrete structure, or they may have substantial cross-loadings on multiple factors representing emotion dimensions like valence. For example, instead of *Anger* and

Disgust variables loading onto one single factor, the *Anger* variables could be split between an *Anger* factor and a general negativity factor. A two factor model (one for positive emotions, one for negative emotions) could be the best fitting model. Such an outcome would imply that lexicons do not have clean divisions of words across emotions.

Based upon my own research experience with Don't Care-DM, I hypothesized that it actually represents a topic (politics) rather than an emotion. Therefore, I expected that Don't Care-DM would be removed from the exploratory model as it would show weak relationships to the other variables.

7.2 Results

Exploratory Graph Analysis

Both the EGA_{GGM} and EGA_{TMFG} fit with four clusters. Contrary to my hypothesis, the EGA_{TMFG} showed better performance across all fit indices including the total entropy fit index (-24.79 vs. -24.41). However, relative fit was quite poor for both EGA methods, with CFI and TFI ≈ 0.75.

The first cluster contained all the DM variables (Figure 7.1); this cluster was also present in EGA_{GGM}. The second cluster consisted of the three *Anger* variables, alongside Anxiety-LIWC and all other ESN variables except Sad-ESN. The third cluster contained the positive EmoLex variables alongside PosEmo-LIWC. The fourth cluster contained the three *Sadness* variables plus Disgust-EmoLex and Fear-EmoLex. Thus, though specific emotion groups could be seen in the clusters, the clusters were also largely defined by lexicons.

Exploratory Models

Using the results of the EGA_{TMFG} as a starting point, exploratory IRMs with four and five factors were fit. Neither model achieved acceptable fit across all indices, but they were

Figure 7.1
EGA_{TMFG} using all Lexicon variables

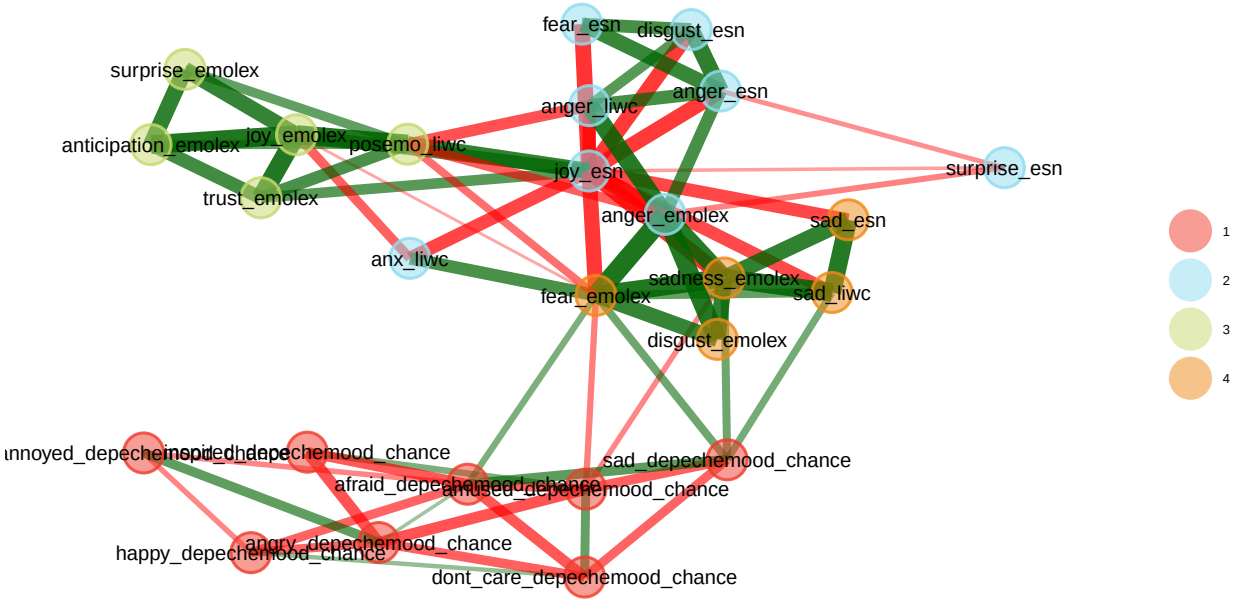


Table 7.1
Global Fit of the Exploratory Models

	M_2	*df*	*p*	RMSEA [95% CI]	SRMSR	TLI	CFI	AIC	BIC
All Variables									
Four Factor	1965	227	< .001	0.07 [0.06-0.07]	0.06	0.83	0.88	112556	113409
Five Factor	1347	205	< .001	0.06 [0.05-0.06]	0.06	0.88	0.92	111747	112751
No DepecheMood++									
Three Factor	920	102	< .001	0.07 [0.06-0.07]	0.08	0.89	0.92	46041	46515
Four Factor	524	87	< .001	0.05 [0.05-0.06]	0.06	0.93	0.96	45646	46223
Five Factor	527	73	< .001	0.06 [0.05-0.06]	0.06	0.91	0.96	45558	46231
No DepecheMood++, No Joy-ESN									
Three Factor	468	88	< .001	0.05 [0.04-0.05]	0.05	0.93	0.95	43419	43865
Four Factor	322	74	< .001	0.04 [0.04-0.05]	0.05	0.94	0.97	43217	43760
Five Factor	233	61	< .001	0.04 [0.03-0.05]	0.05	0.95	0.98	43167	43799

still a large improvement from the confirmatory models (Table 7.1). Overall, the five factor model showed better fit than the four factor model. Neither model passed the second-order test. Unless otherwise specified, all models in this section converged but did not pass the second-order test. Similar measures were taken as in the confirmatory section to encourage the models to pass the second-order test.

Based on the results of the confirmatory section and on past research that found correlated emotion dimensions, I chose to focus on an oblimin rotation during model selection. The factor memberships of both the four and five factor solutions did not entirely reflect those found in the EGA_{TMFG} (Tables 7.2 and 7.3). Primarily, the DM variables did not constitute one single factor. In both the four and five factor IRMs, the DM variables loaded onto two factors with the variables from ESN, EmoLex, and LIWC split between the remaining factors[8]. The factor membership of the five factor model was more similar to the EGA_{TMFG} than the four factor model because DM variables took up two factors on both. The first factor primarily contained the *Anger* and ESN variables, the second factor contained the positive EmoLex variables, and the third factor contained the *Sadness* variables. The lowest communalities were found among the DM variables (Mdn = 0.46, SD = 0.19), though Surprise-ESN also had particularly low communality. However, the communalities of the DM variables was still much higher than in the final confirmatory model (previously, Mdn = 0.27, SD = 0.17).

The five factor IRM was also run using the Polytomous DM variables. The Polytomous DM variables were still unrelated to the other lexicons and had lower communalities (Mdn = 0.12). This drop seemed to be largely driven by Amused-DM, Happy-DM, and Inspired-DM losing all relation to each other and the other variables. Model fit was higher using the Polytomous transformation (e.g., $TLI_{Poly} = 0.93$ vs. $TLI_{Chance} = 0.88$), however, this is likely due to the increased exclusion of the DM variables, as will be seen in the next section.

[8]This same pattern of separation was also seen using a bifactor rotation.

Table 7.2
Factor Loadings of the Five Factor Model

	$F1$	$F2$	$F3$	$F4$	$F5$	h^2
Amused-DM	0.13	-0.01	-0.08	**-0.42**	**-0.44**	*0.40*
Happy-DM	-0.01	0.07	-0.10	-0.19	**-0.36**	*0.22*
Inspired-DM	-0.11	0.17	0.11	**-0.31**	**-0.42**	*0.37*
Joy-EmoLex	-0.16	**0.92**	-0.05	-0.12	-0.11	*0.99*
Joy-ESN	**-0.65**	0.27	**-0.47**	-0.04	0.10	*0.99*
PosEmo-LIWC	-0.27	**0.55**	-0.20	-0.16	-0.09	*0.60*
Anticipation-EmoLex	0.04	**0.78**	-0.03	0.01	-0.04	*0.61*
Surprise-EmoLex	0.27	**0.78**	0.01	0.09	-0.08	*0.66*
Surprise-ESN	**0.33**	-0.15	-0.02	-0.03	-0.23	*0.15*
Trust-EmoLex	-0.19	**0.64**	-0.19	-0.03	0.21	*0.56*
Angry-DM	0.02	-0.12	-0.05	0.09	**0.81**	*0.71*
Annoyed-DM	0.03	-0.06	0.03	**-0.46**	**0.61**	*0.51*
Anger-EmoLex	**0.71**	0.27	**0.31**	-0.03	0.21	*0.89*
Anger-ESN	**0.95**	-0.09	-0.17	-0.03	-0.05	*0.79*
Anger-LIWC	**0.87**	-0.09	-0.05	-0.09	0.16	*0.79*
Afraid-DM	0.04	-0.12	-0.14	**0.71**	0.09	*0.54*
Fear-EmoLex	**0.49**	**0.37**	**0.38**	0.26	0.10	*0.83*
Fear-ESN	**0.74**	-0.09	-0.10	**0.32**	-0.08	*0.71*
Anxiety-LIWC	**0.38**	0.06	0.25	0.21	0.00	*0.39*
Sad-DM	-0.12	-0.05	0.18	**0.69**	0.00	*0.53*
Sadness-EmoLex	0.25	0.24	**0.75**	0.09	0.11	*0.87*
Sad-ESN	0.06	**-0.32**	**0.90**	-0.07	-0.03	*0.99*
Sad-LIWC	-0.21	0.00	**0.94**	0.08	-0.01	*0.81*
Disgust-EmoLex	**0.64**	0.18	0.27	0.03	0.05	*0.66*
Disgust-ESN	**0.44**	-0.01	0.16	0.18	0.18	*0.45*
Don't Care-DM	-0.21	0.07	-0.07	-0.11	-0.10	*0.12*
Rotated SS loadings	*4.62*	*3.36*	*3.10*	*1.85*	*1.81*	

Note:
An oblimin rotation was used. Loadings above |0.30| are bolded for ease of interpretation. Items are grouped in the table by their hypothesized emotion membership.

Table 7.3
Factor Correlations of the Five Factor Model

	$F1$	$F2$	$F3$	$F4$
$F2$	-0.07			
$F3$	0.39	-0.09		
$F4$	0.27	-0.06	0.25	
$F5$	0.19	-0.04	0.12	0.13

Without DepecheMood++. Because the DM variables were largely unassociated with the other lexicons in both the EGA and IRMs, a new set of IRMs were run without the DM lexicon. Three, four, and five factor exploratory IRMs were fit to cover the decreased number of variables. The four factor model fit better than either the three or five factor models, and also fit better than any model with DM (Table 7.1). All variables in the four factor model had communalities ≥ 0.39 (Mdn = 0.75, SD = 0.17).

As in the confirmatory section, Joy-ESN had noticeably poorer item fit than the other variables, $S - X^2(3) = 118.64$, $p < .001$, $\text{RMSEA}_{S-X^2} = 0.15$. Because of this, Joy-ESN was removed from the exploratory model, which improved model fit and interpretability. However, the results of the model with Joy-ESN will be briefly examined in order to provide insight into Joy-ESN.

The four factor model with Joy-ESN but without DM was examined via an oblimin rotation (Tables 7.4 and 7.5). The positively valenced variables all had substantial loadings onto the second factor except for Joy-ESN. Joy-ESN only had a loading of 0.18, while all others ranged between 0.57 and 0.94. Thus, Joy-ESN did not share meaningful variance with the positive variables. Instead, Joy-ESN showed a stronger relationship to third factor (λ_{F3} = -0.41) which had large positive loadings from the *Sadness* variables, and with the fourth factor (λ_{F4} = -0.68), which had positive loadings from a few *Fear* and *Anger* variables. The positive EmoLex variables did not show substantial relationships with these third and fourth factors; PosEmo-LIWC only had small negative loadings (λ_{F3} = -0.26, λ_{F4} = -0.29). It is evident that while Joy-ESN is not negatively valenced, it does not closely relate to the other

Table 7.4
Factor Loadings of the Four Factor Model, without DepecheMood++

	$F1$	$F2$	$F3$	$F4$	h^2
Joy-EmoLex	-0.08	**0.94**	-0.07	-0.12	*0.98*
Joy-ESN	-0.13	0.18	**-0.41**	**-0.68**	*1.00*
PosEmo-LIWC	-0.09	**0.57**	-0.26	-0.29	*0.64*
Anticipation-EmoLex	-0.08	**0.84**	0.02	0.17	*0.70*
Surprise-EmoLex	0.26	**0.79**	0.00	0.14	*0.70*
Surprise-ESN	**-0.42**	0.05	-0.07	**0.68**	*0.39*
Trust-EmoLex	-0.18	**0.65**	-0.13	-0.03	*0.53*
Anger-EmoLex	**0.88**	0.11	0.19	0.03	*0.93*
Anger-ESN	**0.47**	-0.11	-0.24	**0.62**	*0.78*
Anger-LIWC	**0.88**	-0.29	-0.18	0.05	*0.86*
Fear-EmoLex	**0.56**	**0.31**	**0.35**	0.23	*0.80*
Fear-ESN	0.26	0.05	-0.09	**0.73**	*0.72*
Anxiety-LIWC	0.04	0.15	0.22	**0.67**	*0.61*
Sadness-EmoLex	**0.45**	0.17	**0.70**	0.01	*0.88*
Sad-ESN	-0.13	-0.23	**0.87**	0.25	*0.97*
Sad-LIWC	0.04	-0.05	**0.93**	-0.16	*0.82*
Disgust-EmoLex	**0.64**	0.04	0.20	0.17	*0.67*
Disgust-ESN	**0.51**	-0.19	0.15	0.22	*0.55*
Rotated SS loadings	*3.37*	*3.35*	*2.73*	*2.67*	

Note:
An oblimin rotation was used. Loadings above |0.30| are bolded for ease of interpretation. Items are grouped in the table by their hypothesized emotion membership.

Table 7.5
Factor Correlations of the Four Factor Model, without DepecheMood++

	$F1$	$F2$	$F3$
$F2$	-0.06		
$F3$	0.30	-0.11	
$F4$	0.42	-0.12	0.30

Table 7.6
Item Fit of the Final Model

	$S - X^2$	*df*	*p*	RMSEA
Joy-EmoLex	3.87	2	.145	0.02
PosEmo-LIWC	22.53	2	< .001	0.08
Anticipation-EmoLex	13.31	3	.004	0.04
Surprise-EmoLex	12.08	3	.007	0.04
Surprise-ESN	7.38	2	.025	0.04
Trust-EmoLex	32.16	2	< .001	0.09
Anger-EmoLex	7.96	2	.019	0.04
Anger-ESN	21.71	3	< .001	0.06
Anger-LIWC	16.06	3	.001	0.05
Fear-EmoLex	10.85	2	.004	0.05
Fear-ESN	24.17	3	< .001	0.06
Anxiety-LIWC	14.29	3	.003	0.05
Sadness-EmoLex	13.14	2	.001	0.06
Sad-ESN	12.31	3	.006	0.04
Sad-LIWC	5.56	3	.135	0.02
Disgust-EmoLex	6.58	4	.160	0.02
Disgust-ESN	17.63	6	.007	0.03

positive variables. This same pattern was reflected when the model was examined using a bifactor rotation.

Without Joy-ESN (Final Model). After removing Joy-ESN, model fit improved (Table 7.1). The four factor model fit better than the three factor model. RMSEA, SRMSR, TLI, and CFI all indicated good fit for the four factor model. The five factor model only had two variables with loadings above |0.30| on the fifth factor using both the oblimin and bifactor rotations, and it had a higher BIC than the four factor model. Therefore, the four factor model without DM and Joy-ESN was chosen as the final exploratory model. Prior to describing the oblimin and bifactor rotations, aspects of the model that are not influenced by rotation will be touched on: variable communalities, item fit, and word/person fit.

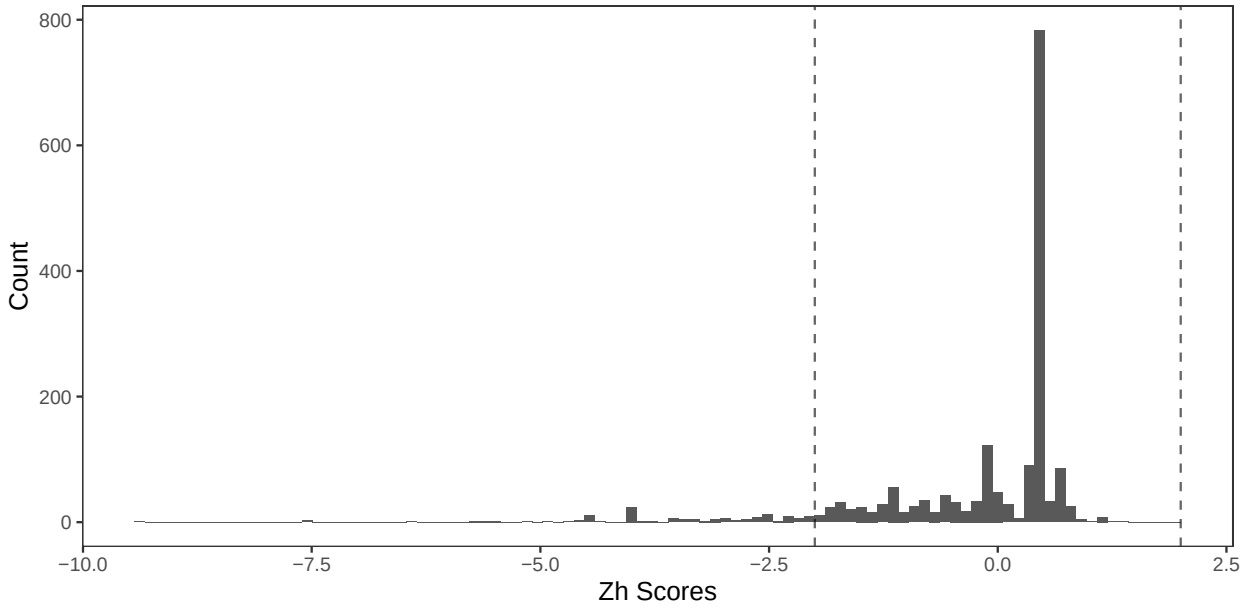
Variables generally shared substantial variance. The highest communalities were seen

among the LIWC variables (Mdn = 0.80, max = 0.89, min = 0.57), and EmoLex variables (Mdn = 0.75, max = 0.99, min = 0.55). Some ESN variables were well explained, while others were not (Mdn = 0.69, max = 0.79, min = 0.19).

Despite the other strong fit indices, M_2 was still significant, indicating potential mis-specification. Likely related to this, only three items were not flagged by $S - X^2$ as misfitting: Joy-EmoLex, Sad-LIWC, and Disgust-EmoLex (Table 7.6). All residuals had a standardized Cramér's $V \leq |.14|$.

Figure 7.2

Distribution of Word Fit based on Zh Scoress



In regards to word/person fit, 7.61% of words with complete observations had Zh scores < -2.00, suggesting atypical or misfit (Figure 7.2). Considering that this data comes from several different lexicons, and many atypical words do exist, this seemed to be a reasonable rate. No words had Zh scores > 2.00, which is associated with over-fit.

Rotations and Parameters of the Final Model

Table 7.7
Factor Loadings of the Final Model, Oblimin Rotation

	$F1$	$F2$	$F3$	$F4$	h^2
Joy-EmoLex	-0.02	**0.92**	-0.11	-0.24	*0.99*
PosEmo-LIWC	0.05	**0.54**	**-0.33**	**-0.47**	*0.73*
Anticipation-EmoLex	-0.18	**0.85**	0.02	0.19	*0.75*
Surprise-EmoLex	0.16	**0.81**	0.04	0.19	*0.70*
Surprise-ESN	**-0.39**	0.00	-0.15	**0.40**	*0.19*
Trust-EmoLex	-0.15	**0.64**	-0.15	-0.14	*0.55*
Anger-EmoLex	**0.84**	0.18	0.26	0.08	*0.96*
Anger-ESN	**0.65**	-0.13	**-0.35**	**0.32**	*0.69*
Anger-LIWC	**0.95**	-0.18	-0.11	-0.07	*0.86*
Fear-EmoLex	**0.39**	0.30	**0.38**	**0.46**	*0.85*
Fear-ESN	0.18	-0.11	-0.21	**0.81**	*0.79*
Anxiety-LIWC	-0.04	0.05	0.19	**0.70**	*0.57*
Sadness-EmoLex	**0.30**	0.13	**0.76**	0.16	*0.90*
Sad-ESN	0.05	-0.14	**0.86**	-0.15	*0.76*
Sad-LIWC	-0.10	-0.14	**0.94**	0.01	*0.89*
Disgust-EmoLex	**0.65**	0.01	0.25	0.13	*0.67*
Disgust-ESN	**0.43**	-0.12	0.18	0.08	*0.32*
Rotated SS loadings	*3.16*	*3.18*	*2.90*	*2.07*	

Note:
Loadings above |0.30| are bolded for ease of interpretation. Items are grouped in the table by their hypothesized emotion membership.

Table 7.8
Factor Correlations of the Final Model, Oblimin Rotation

	$F1$	$F2$	$F3$
$F2$	-0.04		
$F3$	0.24	-0.09	
$F4$	0.45	-0.08	0.24

Oblimin Rotation. First, the final exploratory model was examined using an oblimin rotation. Simple structure, where each variable loads strongly on only one factor, was largely achieved for 65% of variables: 6 from EmoLex (75%), 2 from ESN (50%), 3 from LIWC (75%). This was contrary to my hypotheses. Accordingly, a quick comparison to other rotations (promax, quartimin) showed largely similar factor structures, as is typical when simple structure is present in the data (Gorsuch, 2014, p. 216).

The first factor (*F1*) had the largest positive loadings from the three *Anger* variables starting with Anger-LIWC (λ = 0.95), and followed by Anger-EmoLex (λ = 0.84), and Anger-ESN (λ = 0.65). The linguistic connection between *Anger* and *Disgust* can be seen by the moderate loadings of Disgust-EmoLex (λ = 0.65), and Disgust-ESN (λ = 0.43) on this factor. Smaller loadings came from Fear-EmoLex (λ = 0.39), Sadness-EmoLex (λ = 0.30), and Surprise-ESN (λ = -0.39). An examination of factor scores shows that words relating to *Anger* and *Aggression* (e.g., "criticize", "murderous", "violent") scored highly on $F1$, while peaceful and positive words scored low (e.g., "courtship", "faith", "peace") Table 7.9 provides a snapshot of extreme scoring words on each factor; additional representative words from each factor can be seen in Appendix A.

The second factor had the largest loadings from Joy-EmoLex (λ = 0.92), Anticipation-EmoLex (λ = 0.85), Surprise-EmoLex (λ = 0.81), Trust-EmoLex (λ = 0.64), and PosEmo-LIWC (λ = 0.54). Interestingly, Fear-EmoLex also had a non-trivial loading ($\lambda = 0.30$). Overall, the second factor seemed to measure positive valence, especially among the EmoLex variables. This is supported qualitatively by the patterns of high scoring (e.g., "lovely", "supremacy", "opera") and low scoring words (e.g., "destroyer", "liar", "overwhelm").

The largest loadings for the third factor were found among the three *Sadness* variables. Sad-LIWC had the highest loading (λ = 0.94), followed by Sad-ESN (λ = 0.86), and Sadness-EmoLex (λ = 0.76). Smaller loadings were seen among Fear-EmoLex (λ = 0.38),

PosEmo-LIWC, (λ = -0.33), and Anger-ESN (λ = -0.35). Accordingly, high scoring words included “despair” and “misery”, while low scoring words included “adventurer” and “silly”.

The fourth factor was dominated by *Fear* alongside a more generalized negative dimension. The largest loadings came from Fear-ESN (λ = 0.81) and Anxiety-LIWC (λ = 0.70). Smaller loadings came from Fear-EmoLex (λ = 0.46), Surprise-ESN (λ = 0.40), Anger-ESN (λ = 0.32), and PosEmo-LIWC (λ = -0.47), Representative high scoring words included “anxiety” and “terrorism”, while low scoring words included “safe” and “freedom”.

The three negatively valenced factors all had correlations with each other between 0.24 and 0.45 (Table 7.8). The strongest relationship was between the first factor (*Anger*) and the fourth factor (*Fear*). The second factor (*Positive*) was relatively uncorrelated with the other three factors.

Item Parameters. The slope (α) parameters of each item calculated using an oblimin rotation are shown in Figure 7.3. Due to the cross-loadings across factors, the interpretation of the α parameters becomes slightly more complicated as the parameters are interpreted as a vector, rather than as discrete scalars. While higher and lower α values correspond with higher or lower factor loadings within an item, comparison between items is not as precise due to differences in h^2. For example, Joy-EmoLex has particularly large negative α values for the third and fourth factors despite its factor loadings both being $< |0.30|$. However, α_{F2} for Joy-EmoLex is a clear outlier, being much higher than the α values of any other item, due to Joy-EmoLex’s high positive association with $F2$ and its high h^2. Because factor loadings are a transformation of α values, the relationships between variables and factors are largely discussed above.

Figure 7.4 shows the intercept (d) parameters of each item. The d parameter for each variable is an intercept parameter; it is not influenced by rotation. It describes the basic probability that a word will be associated with the variable, regardless of factor associations. Therefore, the d parameter pattern is similar between the confirmatory and exploratory

Table 7.9
Representative Words of Each Factor using an Oblimin Rotation

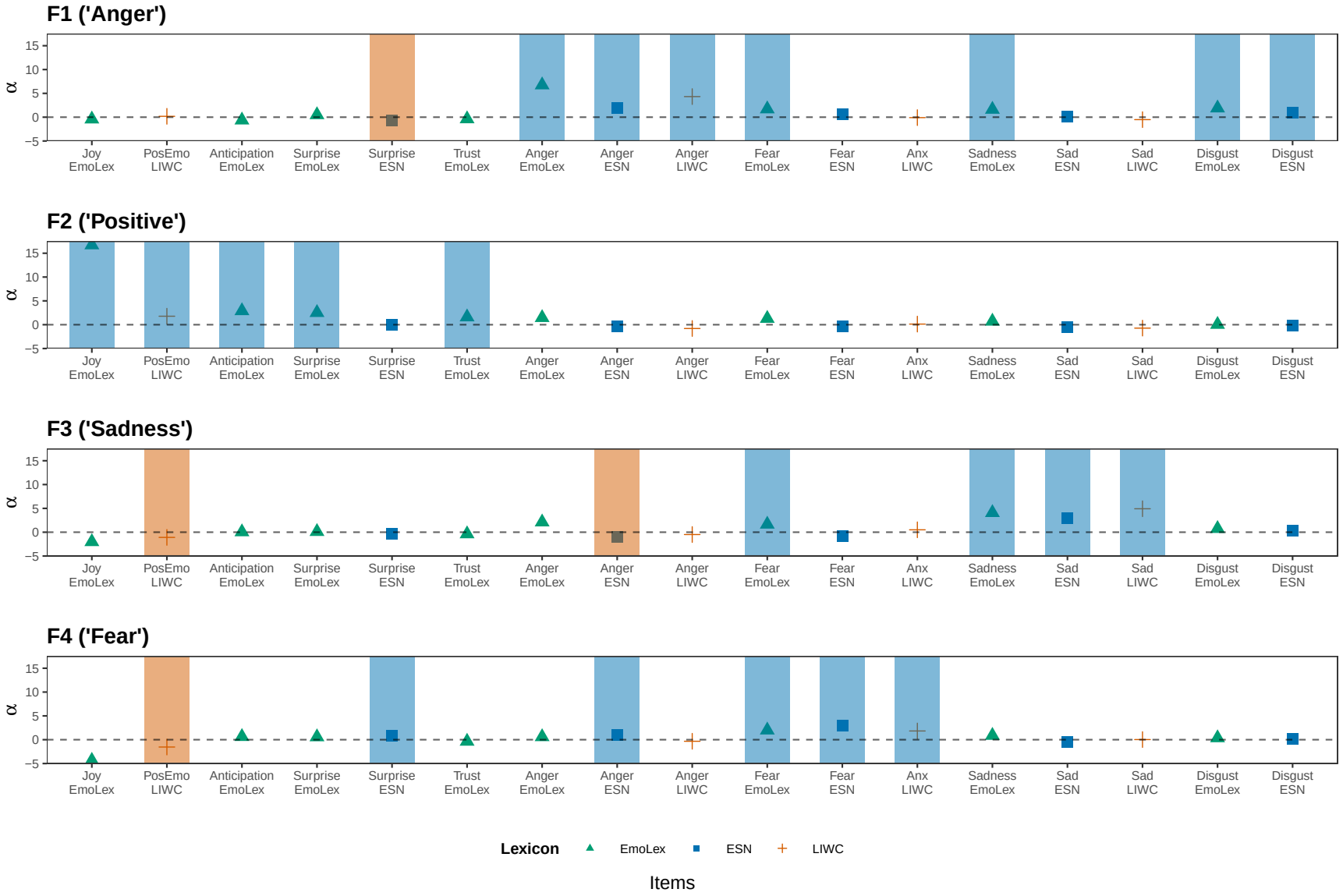
Words with the Highest Factor Scores					Words with the Lowest Factor Scores				
	$F1$	$F2$	$F3$	$F4$		$F1$	$F2$	$F3$	$F4$
F1 ("Anger")									
Criticize	3.24	-0.58	1.93	1.73	Champion	-1.06	2.47	-1.12	-0.82
Murderous	3.22	0.88	1.44	1.83	Courtship	-0.90	2.51	-0.66	-0.35
Nasty	3.28	-0.19	1.32	1.84	Faith	-0.78	2.47	-0.94	-0.99
Offender	3.39	-0.32	1.39	1.89	Pastor	-0.88	2.30	-0.83	-0.80
Prick	3.22	0.92	0.36	1.70	Peace	-0.78	2.47	-0.94	-0.99
Rape	3.49	-0.53	1.28	2.63	Perfect	-0.78	2.47	-0.94	-0.99
Violent	3.22	0.92	0.36	1.70	Treasure	-0.78	2.47	-0.94	-0.99
F2 ("Positive")									
Feeling	1.90	3.23	1.41	1.10	Destroyer	3.04	-1.01	1.08	2.27
Lovely	0.56	3.12	1.02	-0.04	Feudalism	3.03	-0.99	0.97	2.14
Opera	1.66	3.21	1.44	1.01	Humiliating	2.05	-1.12	1.36	1.21
Romance	0.95	3.27	1.25	0.55	Humiliation	2.03	-1.12	2.01	1.39
Supremacy	1.84	3.28	0.30	0.66	Liar	2.07	-1.03	1.26	0.58
Treat	1.97	3.29	1.25	0.76	Overwhelm	1.48	-1.50	1.56	2.52
Weight	1.14	3.17	1.49	0.95	Whine	1.20	-1.04	2.35	2.19
F3 ("Sadness")									
Despair	1.94	-0.36	3.36	1.45	Adventurer	1.38	0.48	-1.13	1.24
Hopelessness	1.94	-0.36	3.36	1.45	Bounty	0.95	2.79	-1.34	1.11
Hurtful	1.94	-0.36	3.36	1.45	Credit	0.23	1.34	-1.14	0.38
Lonely	1.94	-0.36	3.36	1.45	Equality	1.05	2.64	-1.40	0.82
Lose	1.88	0.76	3.38	1.50	Hardy	0.41	2.38	-1.20	-0.46
Misery	2.02	-0.49	3.36	2.25	Proud	0.29	2.56	-1.32	-0.31
Resign	1.94	-0.36	3.36	1.45	Silly	0.50	2.30	-1.36	-0.68
F4 ("Fear")									
Anxiety	1.92	1.30	1.51	2.86	Excellent	-0.57	2.29	0.32	-1.58
Catastrophe	3.15	0.79	1.24	2.83	Freedom	-0.76	2.19	-1.11	-1.42
Scold	3.05	-0.37	1.37	2.73	Hug	-0.76	2.19	-1.11	-1.42
Terrorism	2.69	-0.16	1.44	3.08	Kind	-0.76	2.19	-1.11	-1.42
Terrorist	2.37	1.06	1.56	2.94	Safe	-0.76	2.19	-1.11	-1.42
Terrorize	2.12	0.14	1.62	2.79	True	-0.76	2.19	-1.11	-1.42
Worrying	1.20	1.17	1.49	2.80	Wealth	-0.76	2.19	-1.11	-1.42

Note:
Shown are seven words associated with the most extreme scores per factor, and the suggested interpretation of the dimensions. Many words tied for the lowest scores of each factor. Therefore, a selection of the lowest seven were randomly chosen. Factor scores are interpreted akin to z-scores.

models. The median intercepts among each lexicon were: ESN = -3.82 (SD = 1.04), EmoLex = -5.16 (SD = 6.83), and LIWC = -6.44 (SD = 2.79). LIWC did have the lowest median intercept, as hypothesized. Again, Joy-EmoLex (d = -24.90) is a clear outlier as it was for the α parameter and in the confirmatory model. The combination of low d and high α again suggest that words that are associated with Joy-EmoLex are highly likely to be associated with other positively valenced variables.

Figure 7.3

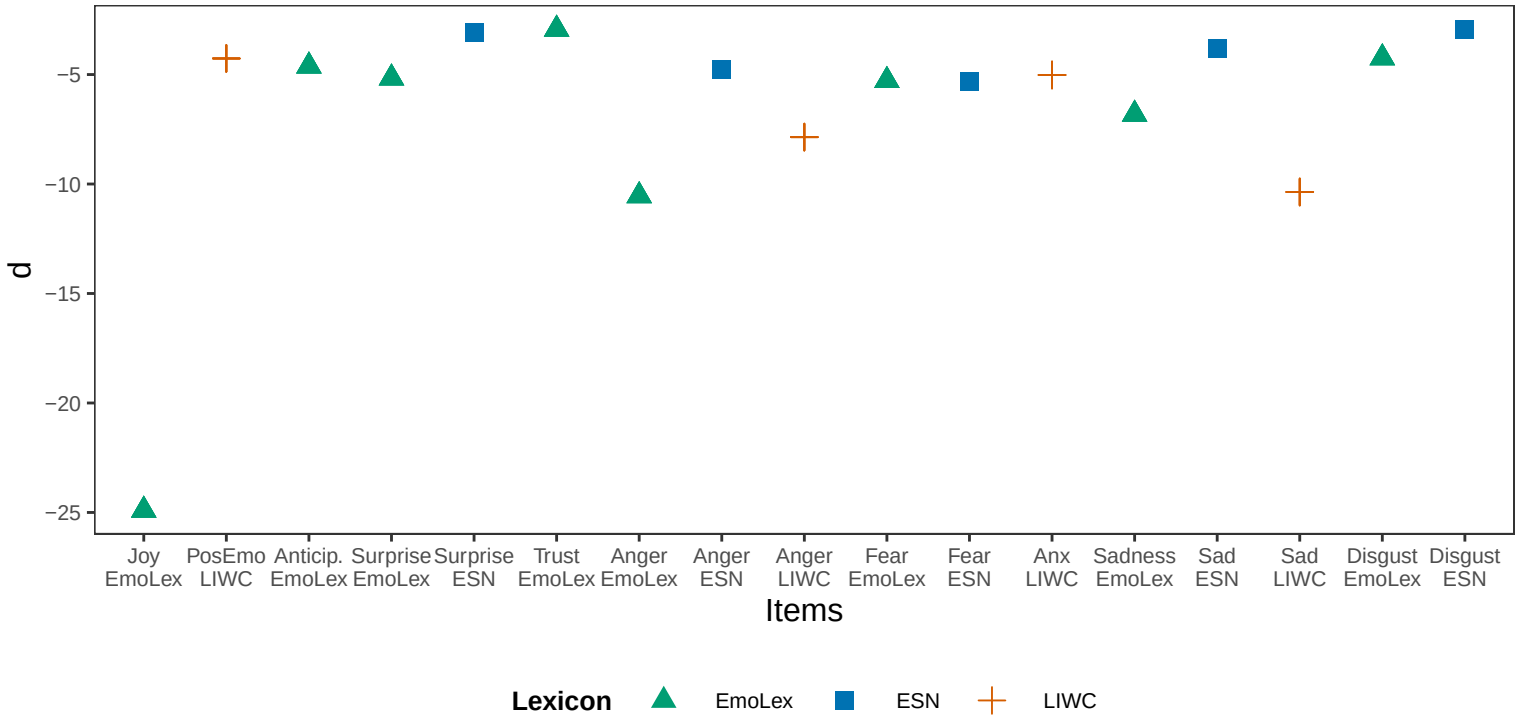
Oblimin Rotation Slope (α) Parameters of the Final Exploratory Model



Note. Blue bars indiciate variables which have a loading above 0.30 for that factor. Red bars indicate loadings below -0.30. Dashed lines at y = 0 are for visibility.

Figure 7.4

Oblimin Rotation Intercept (d) Parameters of the Final Exploratory Model



Word Parameters. In the exploratory model, every word[9] received a factor score for each of the four dimensions. The factor scores are interpreted akin to standardized Z-scores. The distributions of the factors' scores were quite similar to one another (see the diagonal element of Figure 7.5). For identification purposes during model estimation, the mean of each factor was fixed to zero and the standard deviations to one. Accordingly, the median of each factor score distribution was approximately zero with standard deviations between 0.72 and 0.82. It is a positive sign that tails were fairly light; most words should not be scoring extremely high on multiple dimensions.

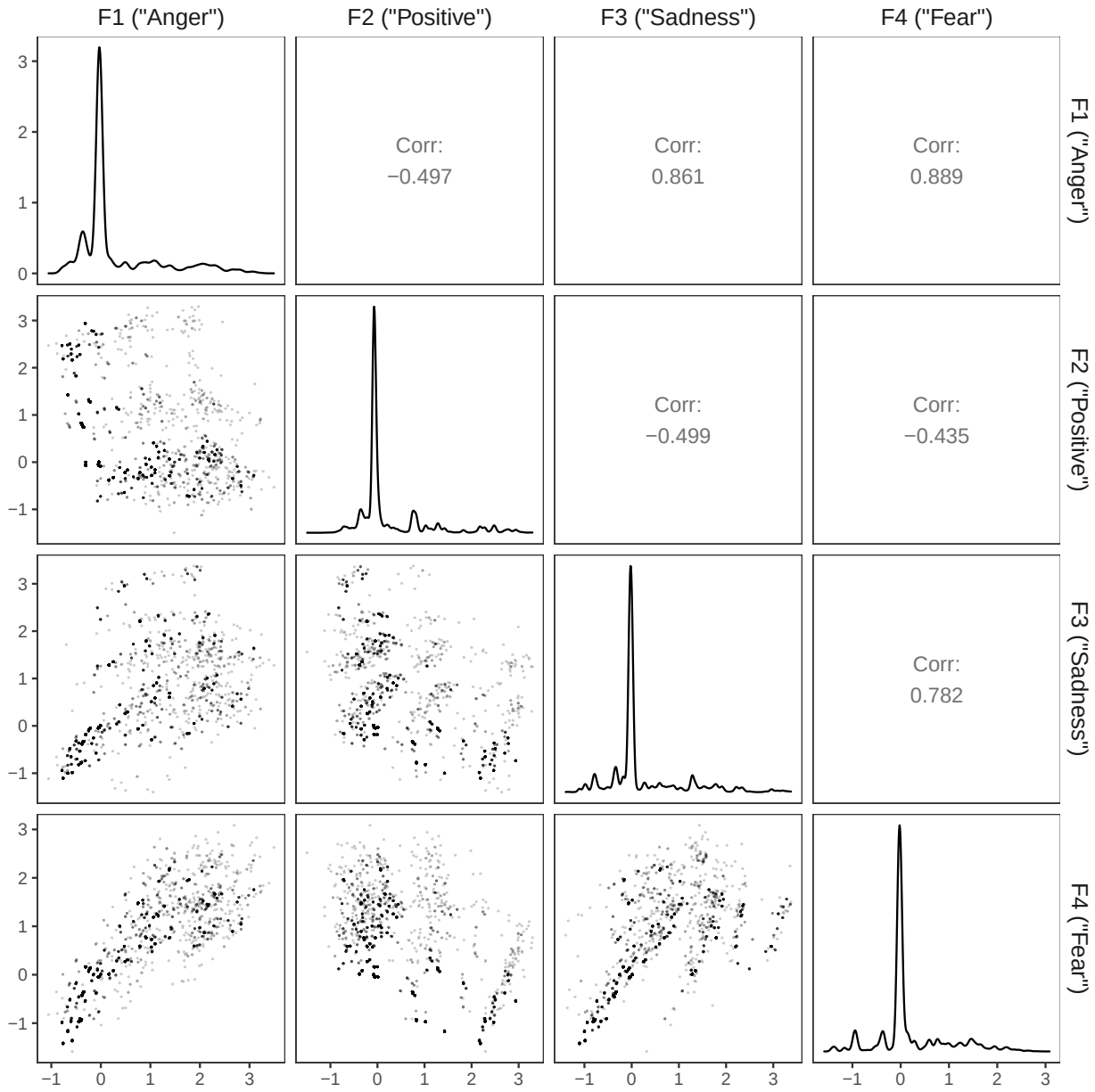
Unlike in the confirmatory model, the same positive words were not the "opposites" of all negative dimensions. The factor score correlations involving the positive factor $F2$ were much smaller than in the confirmatory section, and the scatter plots in Figure 7.5 are much more dispersed. While the cone shape is still present, it is found between fewer factors and the points are much more dispersed. This can also be seen qualitatively in Table 7.9 and Appendix A. The same exact words do not score the lowest on each of the negative dimensions, though there is still overlap.

$F2$ scores also were the least correlated with the other dimensions. Spearman correlations involving the $F2$ scores were $\approx$ -0.45, while the correlations between the other three dimensions ranged from 0.78 to 0.89. This is the opposite of what was found in the confirmatory model where the correlations between the *Positive* and the negative factors/scores were higher than between the negative factors/scores. Spearman correlations were used as several of the factor score scatter plots suggested non-linear relationships. As in the confirmatory section, the factor score correlations are all larger than the model estimated factor correlations.

[9]The factor scores of five words failed to converge successfully: "complain", "confidence", "intelligence", "procession", and "sweetheart."

Figure 7.5

Relationships Between Oblimin Factor Scores



Note. Lower plots are scatterplots of each pair of factor scores, along the diagonal are the densities of each factor score, while upper plots show the Spearman correlation between the factor score pairs.

Bi-Factor Rotation. The final exploratory model was also examined using an orthogonal bifactor rotation. While an oblique bifactor rotation was first attempted, the magnitude of the largest factor correlation was 0.12. Thus for simplicity, the orthogonal bi-factor rotation was used. There was not a substantial difference between the resulting factor loadings or the word factor scores with either rotation. The primary purpose of the bifactor rotation was to create a valence factor through the general factor.

As hypothesized, the general factor (G) did seem to measure valence (Table 7.10). All variables had a factor loading $> |.40|$ on the general factor except for Anticipation-EmoLex, Surprise-EmoLex, and Surprise-ESN. All the negatively valenced variables had positive loadings, while the positive variables had negative loadings. The strongest loadings came from the negatively valenced EmoLex variables ($0.75 < \lambda < 0.81$). A brief examination of the words with the highest factor scores (e.g., "agony", "catastrophe", "rape") and lowest factor scores (e.g., "harmony", "hug", kind", "tranquility") supports the interpretation of this factor as a valence dimension (Table 7.11, additional representative words from each factor can be seen in Appendix B). In addition, though the loadings on G are very different from any factor's loading pattern in the oblimin rotation, G factor scores correlate quite highly with the oblimin factor scores from $F1$ (*Anger*, $r = 0.93$), $F3$ (*Sadness*, $r = 0.81$), and $F4$ (*Fear*, $r = 0.95$). The correlation between G and $F2$ (*Positive*) was smaller, but still moderately sized (r = -0.41).

To support my interpretation of G as valence, I examined the correlations between the G factor scores and two different valence, arousal, dominance (VAD) lexicons: the NRC-VAD lexicon (Mohammad, 2018) and the Warriner et al. 2013 VAD lexicon[10]. NRC-VAD is the dimensional counterpart to EmoLex, while the Warriner et al. 2013 lexicon is a large crowdsourced VAD lexicon created by researchers in psychology and linguistics. There was a -0.73 correlation between G scores both valence measures (Table 7.12). G runs positive

[10]Only words that appeared in CompLex, NRC-VAD, and Warriner were included, n = 5422 (75.83% of CompLex). Both NRC-VAD and Warriner use continuous scores. NRC-VAD scores words between 0 to 1 on each dimension, while Warriner scores 0 to 7.

Table 7.10
Factor Loadings of the Final Model, Orthogonal Bifactor Rotation

	G	$Gr1$	$Gr2$	$Gr3$	h^2
Joy-EmoLex	**-0.49**	**0.86**	-0.02	0.12	*0.99*
PosEmo-LIWC	**-0.65**	**0.46**	-0.16	0.27	*0.73*
Anticipation-EmoLex	-0.18	**0.83**	-0.01	-0.18	*0.75*
Surprise-EmoLex	0.11	**0.83**	-0.06	0.02	*0.70*
Surprise-ESN	-0.05	-0.01	-0.16	**-0.40**	*0.19*
Trust-EmoLex	**-0.46**	**0.58**	-0.06	0.00	*0.55*
Anger-EmoLex	**0.81**	**0.31**	0.03	**0.44**	*0.96*
Anger-ESN	**0.60**	-0.02	**-0.51**	0.24	*0.69*
Anger-LIWC	**0.66**	-0.06	-0.26	**0.59**	*0.86*
Fear-EmoLex	**0.82**	**0.41**	0.10	0.00	*0.85*
Fear-ESN	**0.72**	0.01	**-0.45**	-0.26	*0.79*
Anxiety-LIWC	**0.65**	0.14	-0.04	**-0.36**	*0.57*
Sadness-EmoLex	**0.76**	0.22	**0.52**	0.07	*0.90*
Sad-ESN	**0.42**	-0.11	**0.76**	0.05	*0.76*
Sad-LIWC	**0.48**	-0.11	**0.80**	-0.11	*0.89*
Disgust-EmoLex	**0.75**	0.12	0.04	**0.31**	*0.67*
Disgust-ESN	**0.52**	-0.04	0.04	0.21	*0.32*
Rotated SS loadings	*5.81*	*3.03*	*2.09*	*1.24*	

Note:
Loadings above |0.30| are bolded for ease of interpretation. Items are grouped in the table by their hypothesized emotion membership.

Table 7.11

Representative Words of Each Factor using the Bifactor Rotation

Words with the Highest Factor Scores					Words with the Lowest Factor Scores				
	G	*Gr*1	*Gr*2	*Gr*3		*G*	*Gr*1	*Gr*2	*Gr*3
General Factor									
Agony	3.36	-0.05	0.58	-0.19	Freedom	-1.76	1.82	-0.23	0.55
Catastrophe	3.22	1.57	-0.69	0.51	Hug	-1.76	1.82	-0.23	0.55
Misery	3.21	0.29	1.89	-0.22	Kind	-1.76	1.82	-0.23	0.55
Rape	3.48	0.27	-0.71	1.06	Safe	-1.76	1.82	-0.23	0.55
Scold	3.33	0.41	-0.54	0.51	Tranquility	-1.76	1.82	-0.23	0.55
Terrorism	3.35	0.64	-0.52	-0.17	True	-1.76	1.82	-0.23	0.55
Traitor	3.53	0.06	0.12	0.49	Wealth	-1.76	1.82	-0.23	0.55
Group Factor 1									
Destination	0.89	3.42	0.99	-0.03	Discouragement	0.89	-0.61	2.41	-0.34
Feeling	1.42	3.65	0.62	0.75	Empty	0.89	-0.61	2.41	-0.34
Opera	1.27	3.61	0.75	0.59	Humiliating	2.16	-0.64	0.29	0.88
Romance	0.63	3.52	0.93	0.30	Neglect	0.89	-0.61	2.41	-0.34
Supremacy	0.82	3.55	-0.28	1.17	Sigh	0.89	-0.61	2.41	-0.34
Treat	1.22	3.67	0.58	1.14	Suffer	0.89	-0.61	2.41	-0.34
Weight	1.02	3.51	0.95	0.12	Yearn	0.89	-0.61	2.41	-0.34
Group Factor 2									
Bereavement	1.47	-0.30	2.68	-0.24	Adventurer	0.90	0.69	-1.97	0.34
Cry	1.47	-0.30	2.68	-0.24	Ballot	1.24	1.73	-1.72	-0.24
Dull	1.47	-0.30	2.68	-0.24	Bounty	0.23	2.91	-1.91	-0.01
Isolate	1.47	-0.30	2.68	-0.24	Equality	0.13	2.73	-1.88	0.37
Isolation	1.47	-0.30	2.68	-0.24	Invigorate	1.22	0.61	-1.86	0.52
Pity	1.47	-0.30	2.68	-0.24	Recreational	0.56	3.00	-1.77	0.27
Resignation	1.31	0.87	2.69	-0.26	Revenge	2.51	2.28	-1.77	0.87
Group Factor 3									
Asshole	2.08	-0.08	-0.44	1.95	Avalanche	2.27	1.63	-0.01	-1.13
Cheat	2.08	-0.08	-0.44	1.95	Nervous	1.92	1.73	-0.72	-1.34
Damn	2.16	-0.06	-1.02	1.91	Risk	1.38	1.87	-0.11	-1.26
Grating	1.18	1.34	-0.24	1.92	Risky	1.92	1.73	-0.72	-1.34
Playful	0.40	2.95	-0.35	1.93	Tense	2.01	-0.32	0.15	-1.40
Prejudice	1.88	0.00	-0.49	1.90	Uneasiness	1.97	1.43	0.13	-1.29
Shit	2.08	-0.08	-0.44	1.95	Worrying	2.34	1.79	0.08	-1.45

Note:

Shown are seven words associated with the most extreme scores per factor, and the suggested interpretation of the dimensions. Many words tied for the lowest scores of each factor. Therefore, a selection of the lowest seven were randomly chosen. Factor scores are interpreted akin to z-scores.

(low) to negative (high), hence the negative correlations. G also had a large negative correlation with Dominance-Warriner (r =-0.61), and a smaller negative correlation with Dominance-NRC (r =-0.34). As valence and dominance are typically correlated in practice, the correlations with valence and dominance supports the interpretation of G as a valence dimension (Jack et al., 2016; Mohammad, 2018; Warriner et al., 2013).

The first group factor ($Gr1$) contained the positive EmoLex variables and PosEmo-LIWC, as well as Anger-EmoLex and Fear-EmoLex. Each of these variables had positive loadings, though the positive EmoLex variables had higher loadings ($0.58 < \lambda < 0.86$) than PosEmo-LIWC ($\lambda = 0.46$), Anger-EmoLex ($\lambda = 0.31$), and Fear-EmoLex ($\lambda = 0.41$). High scoring words included "graduation", "opera", "powerful", "romance", "supremacy", and "winning"; Low scoring words included "humiliating", "liar", "neglect", "useless", and "sigh". If interpreted using VAD, $Gr1$ seems to range from positive, high dominance, high arousal words to negative, low dominance, low arousal words. That Fear-EmoLex and Anger-EmoLex load onto $Gr1$ may represent their high arousal and dominance aspects. Or, they may be present on *Gr1* due to their association with the other EmoLex variables; each of the three *Anger* and *Fear* variables were split up across the three group factors depending on which lexicon they came from.

The correlations between $Gr1$ scores and the VAD lexicon variables were all between 0.10 and 0.25. These correlations may be lower than G's for two reasons. First, $Gr1$ represents what the discrete emotion variables have in common, rather than previously empirically defined pure dimensions. If $Gr1$ can be defined through VAD, it is with a combination of dimensions, which would lower the correlations. Second, NRC-VAD and Warriner agree much more on valence ($r = 0.86$) than on arousal ($r = 0.64$) or dominance ($r = 0.39$). Thus, the VAD correlations and the proposed interpretations of the factor scores should both be viewed skeptically. There is ample room to apply the analysis methods of this dissertation to VAD lexicons.

Table 7.12
Correlations Between the Bifactor Factor Scores and VAD Lexicons

	Valence NRC	Valence Warriner	Arousal NRC	Arousal Warriner	Dominance NRC	Dominance Warriner
General Factor	-0.73	-0.73	0.35	0.27	-0.34	-0.61
Group 1 ("Positive")	0.19	0.18	0.22	0.25	0.22	0.10
Group 2 ("Sadness")	-0.29	-0.29	-0.07	-0.03	-0.31	-0.28
Group 3 ("Aggression")	-0.17	-0.16	0.23	0.20	-0.01	-0.09
Valence-Warriner	0.86					
Arousal-NRC	-0.30	-0.31				
Arousal-Warriner	-0.20	-0.19	0.64			
Dominance-NRC	0.53	0.40	0.26	0.15		
Dominace-Warriner	0.72	0.75	-0.27	-0.18	0.39	

The second group factor ($Gr2$) included the three *Sadness* variables as well as Anger-ESN and Fear-ESN. The three *Sadness* variables had positive loadings ($0.52 < \lambda < 0.80$), while there were negative loadings from Anger-ESN (λ = -0.51) and Fear-ESN (λ = -0.45). High scoring words included “bereavement”, “cry”, “dull”, “isolate”. Low scoring words included “adventurer”, “equality”, “revenge”, “invigorate”. $Gr2$ seems to represent *Sadness* (typically a low valence, low arousal, low dominance emotion) on the positive end, and higher arousal, higher dominance words on the negative end. Thematically, low $Gr2$ scores seemed to also reflect *will* or *efficacy* in opposition to *Sadness*, perhaps similar to the potency-control dimension identified in Fontaine et al. (2007). $Gr2$’s arousal/dominance pattern is moderately supported by the VAD correlations. Correlations between $Gr2$ and each of the valence and dominance variables were approximately -0.30.

The third group factor ($Gr3$) had a more eclectic pattern of loadings than $Gr1$ or $Gr2$. Positive loadings came from Anger-LIWC, Anger-EmoLex, and Disgust-EmoLex, while negative loadings came from Surprise-ESN and Anxiety-LIWC ($0.31 < |\lambda| < 0.59$). Unlike the *Anger* and *Fear* lexicon pairs on $Gr1$ and $Gr2$, Anger-LIWC and Anxiety-LIWC had opposite signs. The *Anger* variables and Disgust-EmoLex had opposite signs to Anxiety-LIWC and Surprise-ESN.

High scoring words on $Gr3$ included "asshole", "cheat", "prejudice", "shit", and "antagonistic". Low scoring words included "anxiousness", "nervous", "risk" and "uneasiness". These words and factor loadings seemed to range from high aggression to submission, with high arousal throughout. Aggression is not exactly synonymous with dominance; the NRC-VAD lexicon defined dominance as "in control of the situation, powerful, influential, important, autonomous" (Mohammad, 2018). There were small negative correlations between $Gr3$ and the valence VAD measures ($\approx$ -0.16), small positive correlations with the arousal measures ($\approx$ 0.20), and, surprisingly, little to no relationship to dominance.

Item Parameters. As in the confirmatory section and the oblimin rotation, Joy-EmoLex had extreme α parameters for the factors it loaded most strongly on (Table 7.6). The patterns of the other α parameters generally followed convention, with large positive α's associated with large positive factor loadings, and large negative α's with large negative loadings. The d parameter of the final exploratory model is discussed in the oblimin rotation section above as it is not influenced by rotation.

Word Parameters. For identification purposes during model estimation, the mean of each factor was fixed to zero and the standard deviations to one. This is reflected in the distributions of the factor scores[11], which all had medians at approximately zero (Figure 7.7). Similar to the oblimin rotation and in the confirmatory analysis, the majority of words had factor scores of approximately zero. The largest spread was seen in G (SD = 0.95), while the smallest was seen in $Gr3$ (SD = 0.41).

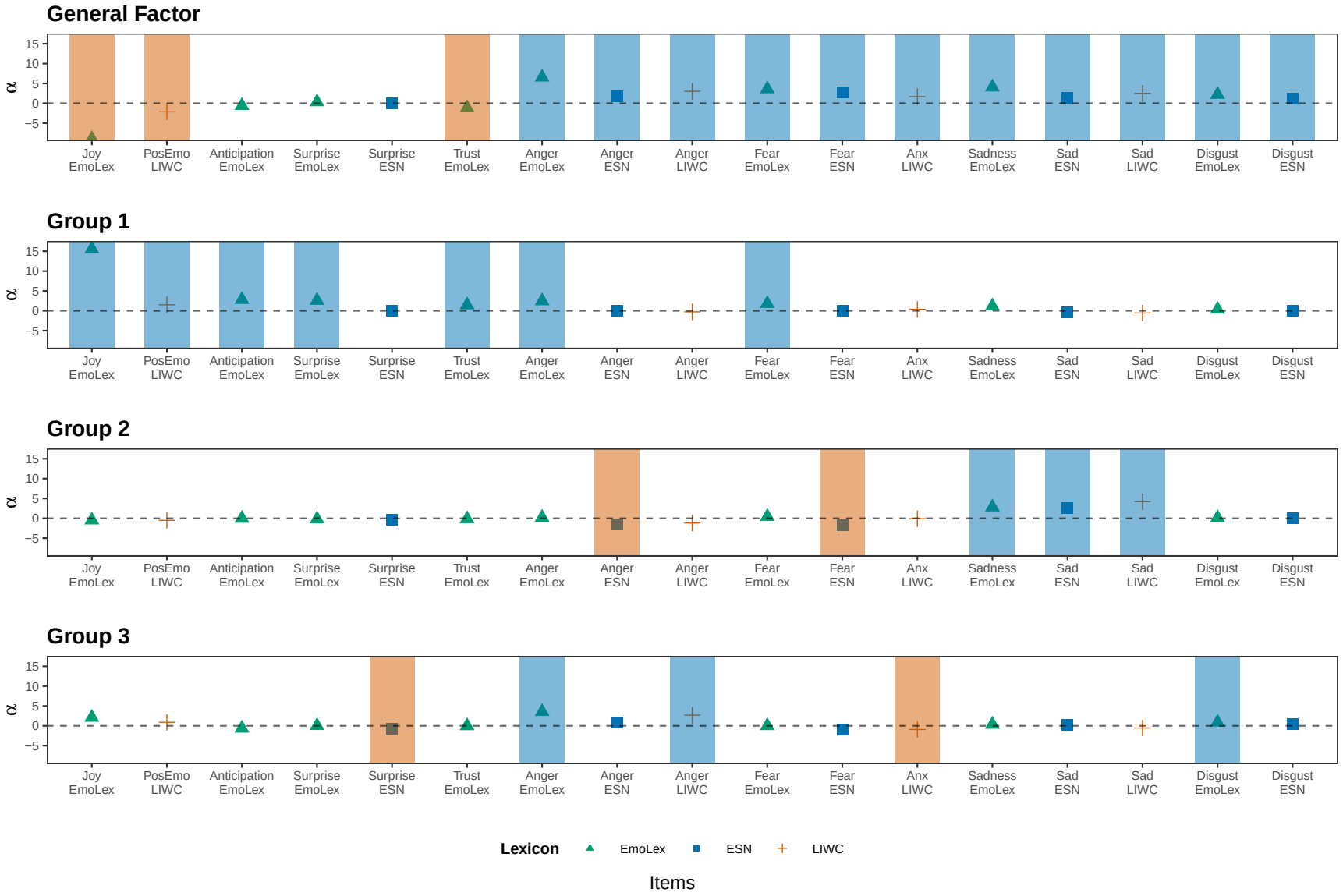
As the bifactor model was orthogonal, correlations between the factor scores were much lower than when calculated with the oblimin rotation. Yet, they were still present despite

[11]Bifactor factor scores for 16 words failed to converge: "angel", "birthday", "celebration", "erotic", "fortune", "gush", "heartfelt", "honeymoon", "independence", "labor", "medal", "organization", "present", "saint", "spa", and "sweetheart".

the orthogonality of the model. The largest Spearman correlation was found between G and $Gr2$ (*Sadness*), $r_s = 0.45$. $Gr3$ (*Aggression*) had the smallest correlations with any of the other factors.

Figure 7.6

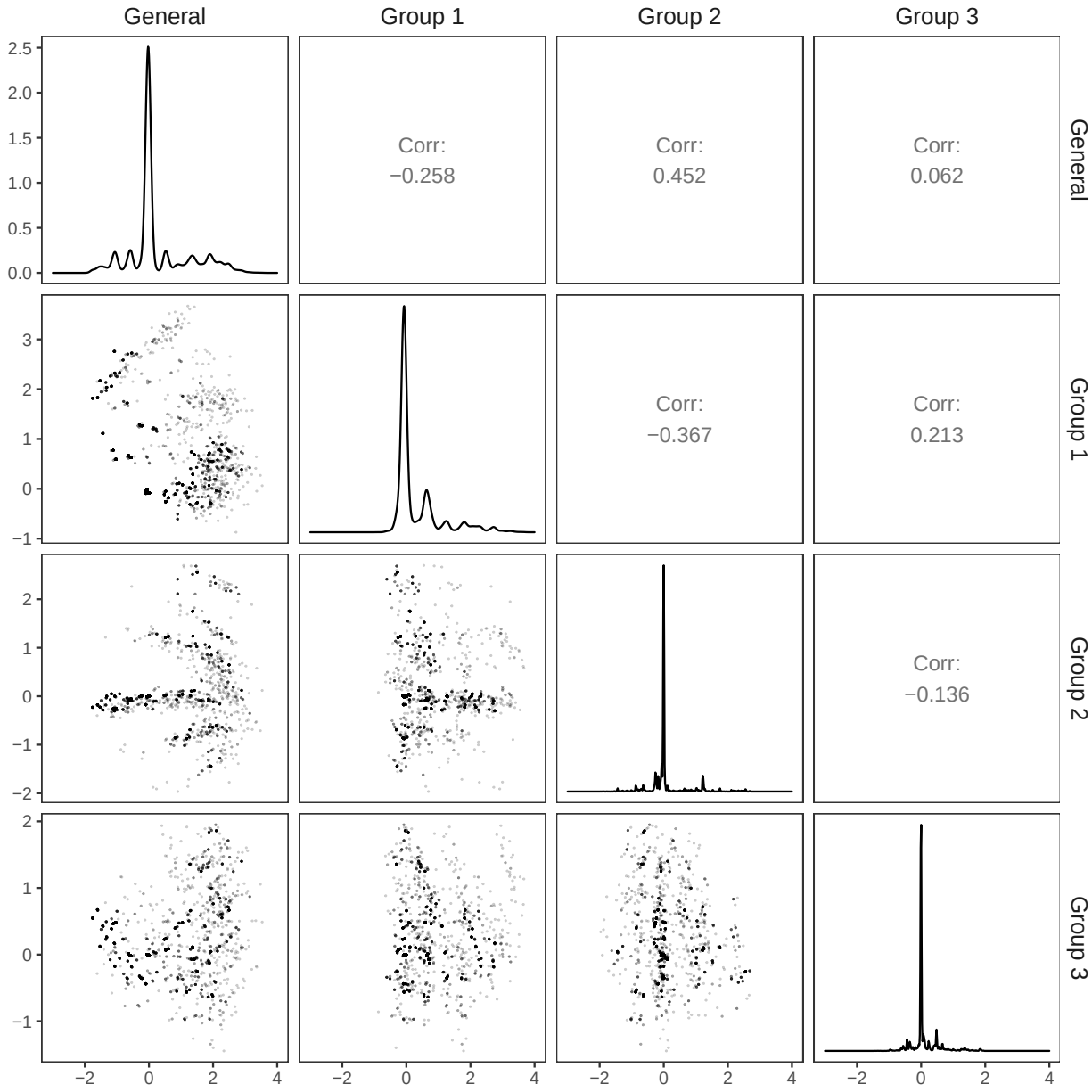
Bifactor Rotation Slope (α) Parameters of the Final Exploratory Model



Note. Blue bars indiciate variables which have a loading above 0.30 for that factor. Red bars indicate loadings below -0.30. Dashed lines at y = 0 are for visibility.

Figure 7.7

Relationships Between Bifactor Factor Scores



Note. Lower plots are scatterplots of each pair of factor scores, along the diagonal are the densities of each factor score, while upper plots show the Spearman correlation between the factor score pairs.

7.3 Discussion

The purpose of the exploratory analysis was to understand how measures of discrete emotions from different lexicons were related to each other. Unlike the confirmatory analysis, variables were able to freely associate, and a more liberal approach to excluding poorly fitting variables was taken. As hypothesized, acceptable fit was achieved by the exploratory model, though this was after the removal of the DM lexicon and Joy-ESN.

Two different rotations were used to interpret the model. The oblimin rotation organized emotions into their same-named groups, while the bifactor model was used extract a valence dimension and examine the remaining differences between variables. Broadly speaking, ESN, EmoLex, and LIWC variables seemed to be highly related to their same-named counterparts with some exceptions. The DM lexicon, however did not seem to relate to the other lexicons at all.

The oblimin rotation came surprisingly closes to replicating the hypothesized model used in the confirmatory section. Simple structure was largely achieved for over half the variables in the oblimin rotation. I had hypothesized that there would be many cross loadings, but this was not entirely supported.
While some variables did load onto multiple factors, the majority only loaded alongside their same-named counterparts.

Based on the results of the exploratory analysis, I would recommend that researchers use either the LIWC or EmoLex lexicon. Their loadings were typically the highest within same-named emotion groups, and they did not have aberrant variables like ESN. This will be discussed in greater detail in the final Discussion section.

DepecheMood++ Performance

There are several interesting takeaways from the exploratory model. First, the DM lexicon was wholly separate from ESN, EmoLex, and LIWC. No DM variable had a loading $> |0.18|$ on any factor besides the two DM factors. As briefly noted, this was true even when a bifactor rotation was used.

As in the confirmatory section, the exact combination of reasons for DM's separation is ambiguous; it is difficult to tease apart issues of mathematical measurement and construct validity. At the very least, the positive DM variables did load in opposite directions to the negative DM variables, but there was no connection to any other lexicon. When the DM variables have no relation to the other lexicons, it is doubtful that they measure the same constructs at all.

Internally, DM does seem to be somewhat consistent. The three positive variables generally loaded in the same way, suggesting that they do measure similar concepts. There was some separation between the negative DM variables, even when the model was reduced down to four factors. Angry-DM and Annoyed-DM did share membership on one factor, but Annoyed-DM also had substantial negative membership on the *Fear*/*Sadness* factor. Thus, Annoyed-DM may not be "*Anger*-lite" as was hypothesized. Don't Care-DM had no relation to any of the lexicon categories, as hypothesized. Perhaps it measures a distinct emotion category that has no relation to *Anger*, *Fear*, *Sadness*, or *Joy*, or perhaps it doesn't measure much of anything. Regardless, the general recommendation is clear: researchers should not apply the DM lexicon without careful consideration.

Bifactor Dimensional Structure

Using the bifactor rotation, the majority of lexicon variables could be understood through valence plus one other dimension. *G*, the valence dimension, had large positive loadings from the negative variables, and moderate negative loadings from the positive

variables. It is notable that there was still a *Positive* factor (*Gr1*) even with *G*. This indicates that the positive variables are not simply "not negative" words; they have unique measurement for positive affect that is separate from negative affect. In contrast, there was not an *Anger* or *Fear* factor. Only *Positive* and *Sadness* variables continued to share unique variance outside of the valence.

The bifactor group factors did not neatly represent other VAD dimensions. This is not entirely surprising for several reasons. First, the bifactor rotation encourages perfect clustering, or simple structure within the group factors. The VAD theory presupposes that every emotion can be described through the VAD dimensions, which is in conflict with simple structure. Second, the lexicons are built using discrete variables, not with dimensional structure. These discrete clusters are obvious within the patterns of factor loadings, especially for the *Sadness* variables. Despite this, the bifactor rotation did reveal some dimensional elements and measurement quirks among the factors, and was useful for understanding the relationships between variables. And finally, it is difficult to make firm conclusions about VAD within the bifactor rotation due to the disagreement between NRC-VAD and Warriner lexicons. Further research is necessary to understand how VAD lexicons relate to each other.

Relationships Among Same-Named Emotion Variables

Positive Variables. It is curious that the three *Joy* variables were quite different, considering the similarity often found among positive words and experiences (Alves et al., 2017; Rozin & Royzman, 2001). Joy-ESN did not consistently load alongside Joy-EmoLex and PosEmo-LIWC; rather it had negative loadings opposite to the *Anger* and *Fear* variables. Joy-EmoLex and PosEmo-LIWC were not perfect counterparts, either. Like Joy-ESN, PosEmo-LIWC had typically had loadings opposite to the *Anger* and *Fear*. Joy-EmoLex seemed to share just as much in common with Anticipation-EmoLex and

Surprise-EmoLex than with PosEmo-LIWC.

It does not seem to be the case that Joy-, Anticipation-, Surprise-, and Trust-EmoLex combine to measure PosEmo-LIWC. Rather, Joy-EmoLex seems to share variance with PosEmo-LIWC and Trust-EmoLex on one side, and Anticipation-EmoLex and Surprise-EmoLex on the other. This is significant as it indicates that same-named variables from high quality lexicons can still measure different entities, particularly when one lexicon has more fine-grained categories than another.

Of note, Anticipation-EmoLex and Surprise-EmoLex seemed to measure very similar constructs. About 70% of their variance was shared with the factor model, and as they have nearly equivalent factor loadings on each factor for each rotation, this 70% is likely shared with each other. It is unclear how much of their remaining unique variance is also shared outside of the factor model. While *Anticipation* and *Surprise* may be two different embodied experiences, they are also closely related and may not be easily differentiated in text. More investigation here is warranted, especially for applied situations.

In contrast, Surprise-ESN and Surprise-EmoLex had very little in common with each other. They did not load onto the same factors. Surprise-EmoLex seems to be more positively valenced than Surprise-ESN; Surprise-ESN seems to be more related to *Fear*. However, the communality of Surprise-ESN was quite low at 0.19 compared to 0.70 for Surprise-EmoLex. Surprise-ESN is simply distinct from all other lexicon variables. Because ESN and EmoLex are the only lexicons with *Surprise* variables, it is difficult to say which one is a better representation of *Surprise*. At the least, these two variables are not equivalent measurements of the same construct.

It is possible that the poor differentiation between the positive EmoLex variables is due to poor measurement rather than each being a closely related construct. EmoLex was crowdsourced; participants were asked whether or not a word was associated with the eight emotions. It is possible that while there was general consensus on which words belonged in

which category, this consensus was not meaningful (see the mix up of *Jealousy* and *Envy* in Haslam & Bornstein (1996)). While each variable may measure something slightly different, their separation may not be entirely meaningful or consistent in practice.

Anger and Fear Variables. While the *Anger* and *Fear* variables certainly shared much in common, there were again interesting patterns that suggested divergence. In the oblimin rotation, same-named variables generally loaded together onto their respective *Anger* and *Fear* factors. However, Fear-EmoLex had nearly as much common with *Fear* as it did with the *Positive*, *Anger*, and *Sadness* factors. Anger-ESN was also split, though less dramatically, between the *Anger*, *Sadness*, and *Fear* factors.

The bifactor rotation showed similar issues, suggesting that when general valence is removed from *Anger* and *Fear* variables, distinctions between the lexicons are revealed. All the *Anger* and *Fear* variables loaded substantially on G, as would be expected for strong negatively valenced emotions. However, there the similarities ended. This reflects past research that found that discrete emotions across lexicons may still be largely defined by valence rather than shared constructs (Kušen et al., 2017).

Anger-EmoLex and Fear-EmoLex seemed to be closely related to the positive EmoLex variables, either because they come from the same lexicon or because they measure a aspects of arousal and dominance not seen in their same-named counterparts. Anger-ESN and Fear-ESN share a unique opposition to *Sadness*, perhaps relating to *Agency/Potency*. In contrast, Anger-LIWC and Anger-EmoLex seem to share a higher association with *Aggression* that Anger-ESN does not have. Anxiety-LIWC did indeed seem to measure a certain amount of *Anxiety*. This is evidenced by its loading on *Gr3* which ranged from *Aggression* to *Anxiety*. Qualitatively, some of the words that appear in Anxiety-LIWC and not in Fear-ESN or Fear-EmoLex support this: “ashamed”, “doubt”, and “embarrass”. In summary, while there are certainly different flavors of *Anger* and *Fear* being measured by each lexicon, it is mostly visible when valence is purposely separated out.

Sadness Variables. The *Sadness* variables consistently stayed together in both the oblimin and bifactor rotations. However, Sad-ESN and Sad-LIWC seemed to have more *Sadness* specific variance than Sadness-EmoLex. Sadness-EmoLex either had a stronger valence component or was more closely related to the other EmoLex variables. This was evident in both rotations.

Disgust Variables. The *Disgust* variables were fairly similar to each other, but questions remain about their relationship. They both followed the *Anger* variables in the two rotations, and they had the same magnitudes as each other on the remaining factors. There were still substantial differences, though, as Disgust-EmoLex had twice the explained variance as Disgust-ESN, potentially because of the shared variance that EmoLex variables seem to have.

8 Discussion

The purpose of this dissertation was to understand how discrete emotions were measured across different emotion lexicons using multidimensional item response models (IRMs). This dissertation is predicated on the assumption that each lexicon attempts to measure the same discrete emotion constructs. The most accurate measurement of these constructs is then the shared measurement of these constructs - that is, words that all lexicons agree measure a given discrete emotion. This overlap is expressed through the IRM factor loadings and communalities.

To this end, two types of analyses were performed. In the confirmatory analysis, a hypothesized factor structure was imposed on the four lexicons where all same-named variables loaded onto the same emotion factor. This hypothesized structure is typically what is assumed by researchers when they use an emotion lexicon or compare results across studies. In the exploratory analysis, lexicon variables were freely allowed to associate with each other. The conclusion of both these analyses is that the EmoLex and LIWC lexicons seem to measure discrete emotion constructs most consistently. Several variables of ESN showed confusing patterns, and DM was unrelated to the other lexicons.

Table 8.1 shows a comparison of the models examined in the dissertation. It is clear that the exploratory models fit better than the confirmatory models; this is not surprising, as confirmatory models inherently place limitations on a model's structure. However, no confirmatory model reached acceptable fit. The core issue with the confirmatory models was likely the strict simple structure and the presence of the DM lexicon.

Table 8.1
Comparison of Dissertation Models

Type	Factors	Variables Removed	RMSEA [95% CI]	SRMSR	TLI	CFI
Great Fit						
Exploratory	Five Factors	All DM, Joy-ESN	0.04 [0.03-0.05]	0.05	0.95	0.98
Exploratory	**Four Factors**	**All DM, Joy-ESN**	**0.04 [0.04-0.05]**	**0.05**	**0.94**	**0.97**
Good Fit						
Exploratory	Four Factors	All DM	0.05 [0.05-0.06]	0.06	0.93	0.96
Exploratory	Five Factors	All DM	0.06 [0.05-0.06]	0.06	0.91	0.96
Exploratory	Three Factors	All DM, Joy-ESN	0.05 [0.04-0.05]	0.05	0.93	0.95
Fair Fit						
Exploratory	Three Factors		0.07 [0.06-0.07]	0.08	0.89	0.92
Exploratory	Five Factors		0.06 [0.05-0.06]	0.06	0.88	0.92
Exploratory	Four Factors		0.07 [0.06-0.07]	0.06	0.83	0.88
Poor Fit						
Confirmatory	**Combined Positive Factor + Testlets$_{\text{No LIWC}}$**	**Joy-ESN, Surpr-ESN, Annoyed-DM***	**0.07 [0.07-0.07]**	**0.10**	**0.78**	**0.83**
Confirmatory	Combined Positive Factor + Testlets$_{\text{No LIWC}}$		0.09 [0.08-0.09]	0.11	0.72	0.77
Confirmatory	No Disgust Factor + Testlets$_{\text{No LIWC}}$		0.10 [0.09-0.10]	0.12	0.64	0.70
Confirmatory	Testlets$_{\text{No LIWC}}$		0.10 [0.10-0.10]	0.12	0.61	0.69
Confirmatory	Testlets		0.10 [0.10-0.11]	0.12	0.59	0.68
Confirmatory	Hypoth. Structure		0.14 [0.14-0.14]	0.14	0.24	0.34

Note:
Confirmatory models comparing which low communality variables to remove that were not ultimately chosen are not shown for brevity. Bold entries indicate the final confirmatory and exploratory models.
*Annoyed-DM remained in this model, but lost its path to the Anger factor.

The biggest improvement in the confirmatory section came from the addition of the testlet factors, which allowed for variables to relate outside of their associated emotion factor. The second largest jump among the confirmatory models was the combination of the *Joy* and *Surprise* factors into one *Positive* factor. Indeed, the *Positive* factor was also found in the exploratory model using an oblimin rotation, alongside factors that measured *Anger*, *Fear*, and *Sadness*. When a bifactor rotation was used, a general valence factor emerged, alongside factors for positive affect, *Sadness*, and *Aggression*.

Across each model, it is evident that same-named emotions did generally group together on the same factors, indicating that the lexicons did measure similar constructs. However, there were still substantial differences found between same-named variables. The bifactor rotation suggested that some similarities were largely based on the shared valence of the discrete emotion, rather than what is unique to the emotion itself. This was particularly evident for the *Fear* and *Anger* variables. Thus, while the same-named emotion scores from different lexicons may overlap in practice, they may not always be actually measuring the same construct. Deeper discussions of the individual measurement qualities of each variable can be found in the exploratory section.

Based on this dissertation, it is difficult to determine if measurements using these lexicons can be genuinely compared. Because each lexicon measured each emotion sightly differently, their similarities in practice would be highly influenced by the emotions present in studied texts. Anxiety-LIWC, for example, does seem to measure aspects specific to *Anxiety* that are not simply *Fear*. If the studied text does not contain *Anxiety*-related words, then LIWC results may be comparable to EmoLex. However, if *Anxiety* is present, then the scores may diverge. Researchers should be cautious in comparing and interpreting results across studies.

8.1 Lexicons

NRC EmoLex and LIWC

Across both the confirmatory and exploratory sections, both EmoLex and LIWC showed the strongest construct validity. These two lexicons had the highest communalities, their variables loaded highly alongside their same-named counterparts from the other lexicons, and there were not glaring issues like in ESN and DM. However, neither were consistently 'better' across all discrete emotion groups, and it is evident that their same-named emotions do not perfectly match up between them. There are advantages and disadvantages to both, both from a measurement perspective and in how these lexicons can be applied to text.

First, LIWC and EmoLex differ in the number and kinds of discrete emotions they contain. The choice of a lexicon should be partly driven by the emotions that the researcher believes will be in the text. LIWC contains four emotion measures, while EmoLex contains eight. LIWC's emotion categories do not match well with either Ekman or Plutchik's theories, and may be too specific for some circumstances. For example, LIWC does not have a *Disgust* measure, and Anxiety-LIWC does not exactly measure *Fear*. However, the distinctions between some of EmoLex's categories is ambiguous. Anticipation and Surprise seem to be very similar to each other and the other positive variables, and all the EmoLex variables shared a curious amount of variance among them. Further investigation into the differentiation between the EmoLex categories is warranted.

LIWC has the advantage of long and accepted use in psychology; using LIWC may make psychological papers easier to compare and be accepted by the community. EmoLex is freely available to all researchers, though, and is not proprietary. Yet, EmoLex is simply a lexicon while LIWC is a program that uses combines its lexicon with stem matching. Researchers who use EmoLex need to match the text and lexicon using lemmatization or stemming. While free R packages like udpipe (Wijffels, 2022) make this easily available, it is not as

straightforward as LIWC's user interface.

EmoLex is bigger based upon unique words and lemmas. In the CompLex dataset, LIWC did cover almost all of EmoLex with its stems - however, I did not examine how many of these stem matches were accurate. It is possible that some of the lower associations between LIWC and the other lexicons occurred because LIWC over-matched stems to unrelated words. EmoLex may have increased coverage compared to LIWC when appropriate text cleaning (lemmatization) is applied.

The best choice, however, is likely consistency. Results from different studies will be more comparable if they are created from the same lexicon. If either EmoLex or LIWC is commonly used in a specific niche, researchers should consider continuing to use that lexicon to ensure comparability. However, the mixing of lexicons may also lend credence to increased validity of the results and generalizability. Thus the matter is not fully settled.

DepecheMood++

DepecheMood++ should not be used to measure emotion in text without further investigation. The DepecheMood++ lexicon did not relate to any other lexicon in either the Confirmatory or Exploratory analyses, and its compositional nature makes applications tricky. It is clear from the confirmatory and exploratory sections that the Polytomous transformation performed worse than the Chance transformation, though neither seemed optimal. It is possible that the issue lay with the transformations, and proper methods of compositional data analysis (CoDA) may have showed a closer match between DM and the other lexicons. However, considering that typical applications of DM do not use CoDA, further investigation may not be useful.

DM's division of *Joy* into multiple variables (Amused, Happy, Inspired) also contributed to it's poor measurement. The differences in factor loadings between the Chance and Polytomous transformations were most obvious for these three variables. Sad-DM and

Afraid-DM, which were not divided, and Angry-DM, who seemed to share no relationship with Annoyed-DM, all had higher associations with their same-named counterparts in the confirmatory model. While combining DM's split emotions into a single variable could be investigated in the future, there are presently better lexicons to use.

EmoSenticNet

ESN had a mixture of reliable and unreliable variables. Fear-ESN, Sad-ESN, and Anger-ESN generally loaded alongside their same-named counterparts to varying degrees. However, Joy-ESN did not relate well to PosEmo-LIWC or Joy-EmoLex, and seemed to rather indicate the absence of negative emotion. Surprise-ESN also had very little relation to other variables in the model. It is possible that Surprise-ESN's low communality occurred because it is measuring a completely distinct aspect of *Surprise* from EmoLex. However, without further investigation no conclusion can be made. Thus, while ESN does not show the deep issues that DM has, EmoLex and LIWC are better choices.

8.2 Reflection & Limitations

As far as I am aware, this is the first time that IRMs have been applied to study emotion lexicons. I would conclude that this novel application has merit. Logical assumptions and hypotheses about patterns of item parameters were born out, and interesting knowledge about the differences between lexicons was revealed. Many of these differences could only be revealed by examining the internal word-emotion associations of the lexicons rather than comparisons of prediction and classification. If the target texts of a classification competition did not contain *Anxiety*, for example, the differences between Anxiety-LIWC and the other *Fear* variables may not have been revealed.

However, there are always limitations. While good global model fit was achieved in the exploratory section, item fit was poor and there were inconsistencies between the model

factor correlations and the factor score correlations. Excellent model fit does not guarantee that the model is properly specified (F. Chen et al., 2008). There is indeed some conceptual mis-match between the data and the model. IRMs are based upon dimensional factors, but these lexicons were built upon discrete models of emotions. This may account for some of the mismatch between the estimated factor correlations and the word factor scores: because words in some lexicons are only associated with a single emotion, then they are never associated with any other emotion. IRMs cannot entirely reflect such discrete structures, though meaningful results were still obtained.

Further, this dissertation was largely concerned with construct validity using nomothetic span and convergent validity. This dissertation did not seek to answer the question of whether discrete emotion are real or useful constructs. As described in Robinson & Clore (2002), there is an important distinction between emotion as it is experienced in real time and the cultural, conceptual beliefs about emotion. Because the the discrete emotions framework is so pervasive in thought and it is the framework that each lexicon is built upon, it is not surprising that this structure was present in the IRMs. However, the organization of the IRMs does not necessarily guarantee that emotion is actually expressed discretely in text.

Yet, it is still unlikely that these lexicons do not measure anything useful about emotion. LIWC and EmoLex, for example, agreed fairly often even though one had heavy expert oversight and the other was largely based on the layperson ratings.
This convergence speaks to their construct validity. Further, if people write and think about emotion in a discrete framework, then these lexicons will accurately reflect that. I do not think that ED should wait for the field of psychology to agree upon one unified framework before ED techniques can be used.

There is still further research that can be done to understand how lexicons measure emotion in text. First, these analyses were performed directly on the shared words in the lexicons. Therefore, there are questions remaining when considering how the remaining

non-overlapping, un-examined words measure emotions. For example, the exploratory analysis suggests that Anger-ESN has lower levels of words related to *Aggression*. It is possible that when EmoLex and LIWC are expanded to their full size, they contain even more words related to *Aggression*, widening the measurement gap when applied to text. However, it is not unreasonable to assume that unique words from each lexicon are less likely to appear in the general written corpus, and thus would have less of an impact on measurement.

Similarly, differences between lexicons may be magnified or reduced depending on word frequency. Not all words appear in the same frequencies within text, nor are they the same across genres. Differences in texts will likely change how well an emotion is measured. Further research comparing lexicons on different genres and writers is sorely needed.

And finally, differentiation between discrete emotions within the lexicons may not always translate in practice. Simply because all lexicons agree that a given word is associated with a certain emotion does not make it so. I specifically refer to the conflation of *Anger* and *Disgust* in written and spoken communication. While these two emotions were distinct in the lexicons, there is ample evidence that their terminology is mixed up in practice (Jack et al., 2016; Nabi, 2002; Roseman et al., 1994; Royzman et al., 2014; Vicario et al., 2020). I do not believe that clear distinctions within the lexicons will always translate to clear measurement in practice.

8.3 Future Directions

While the models in this dissertation were primarily used to evaluate the lexicons, it is possible the IRMs may themselves be useful for emotion detection. The core premise of this dissertation and the use of IRMs is that the shared variance between lexicons represents the best measurement of the discrete emotions. Therefore, it is possible to use the final exploratory model as an ensemble learning method that produces a new, more comprehensive

lexicon through its factor scores. This factor-based lexicon combines information from each lexicon to indicate which words are most and least likely to indicate a certain facet of emotion. Further work is needed to test the accuracy of the final exploratory model as method of emotion detection.

In addition, there is ample room to repeat this method of analysis on Valence, Arousal, and Dominance (VAD) lexicons. VAD lexicons are the next most popular kind of lexicon used in emotion detection research. It is similarly unknown how VAD lexicons inter-relate, there seem to be substantial differences between popular VAD lexicons. It would also be interesting to examine both VAD and discrete emotion lexicons simultaneously to understand their associations.

8.4 Conclusion

Emotion detection encompasses a wide variety of methods that are used inside and outside of psychology. This dissertation examined one specific tool: general purpose emotion lexicons. While the three of the four lexicons generally matched up, I would recommend either using EmoLex or LIWC. There is still much that is not known about the measurement validity of emotion lexicons, though this dissertation has certainly revealed important distinctions. The field of psychology still has much to contribute methodologically to emotion detection, just as these tools can provide great insight into invisible cognitive processes.

9 References

Abdullah, M., Hadzikadicy, M., & Shaikhz, S. (2018). SEDAT: Sentiment and emotion detection in arabic text using CNN-LSTM deep learning. In *2018 17th IEEE nternational Conference on Machine Learning and Applications (ICMLA)* (pp. 835–840). IEEE.

Acheampong, F. A., Wenyu, C., & Nunoo-Mensah, H. (2020). Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, 2:e12189.

Agrawal, A., An, A., & Papagelis, M. (2018). Learning emotion-enriched word representations. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 950–961).

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, *44*(2), 139–160.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Altman, D. G., & Royston, P. (2006). The cost of dichotomising continuous variables. *BMJ*, *332*(7549), 1080.

Alves, H., Koch, A., & Unkelbach, C. (2017). Why good is more alike than bad: Processing implications. *Trends in Cognitive Sciences*, *21*(2), 69–79.

An, S., Ji, L.-J., Marks, M., & Zhang, Z. (2017). Two sides of emotion: Exploring positivity and negativity in six basic emotions across cultures. *Frontiers in Psychology*, *8*, 610.

Araque, O., Gatti, L., Staiano, J., & Guerini, M. (2018). DepecheMood++: A bilingual emotion lexicon built through simple yet powerful techniques. *arXiv Preprint arXiv:1810.03660*.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA).

Barrett, L. F. (2006a). Are emotions natural kinds? *Perspectives on Psychological Science*, *1*(1), 28–58.

Barrett, L. F. (2006b). Solving the emotion paradox: Categorization and the experience of emotion. *Personality and Social Psychology Review*, *10*(1), 20–46.

Barrett, L. F. (2017). Categories and their role in the science of emotion. *Psychological Inquiry*, *28*(1), 20–26.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238–246.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In M. R. Lord F. M. And Novice (Ed.), *Statistical theories of mental test scores* (pp. 397–472). Addison-Wesley Publishing.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1–8.

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, *114*(28), 7313–7318.

Brown, A. (2016). Thurstonian scaling of compositional questionnaire data. *Multivariate Behavioral Research*, *51*(2-3), 345–356.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & K. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Sage publications.

Cacioppo, J. T., Berntson, G. G., Larsen, J. T., Poehlmann, K. M., & Ito, T. A. (2000). The psychophysiology of emotion. In R. Lewis & J. M. Haviland-Jones (Eds.), *The handbook of emotion* (pp. 173–191). Guilford.

Cai, L. (2010). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, *35*(3), 307–335.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from

language corpora contain human-like biases. *Science*, *356*(6334), 183–186.

Cameron, C. D., Lindquist, K. A., & Gray, K. (2015). A constructionist review of morality and emotions: No evidence for specific links between moral content and discrete emotions. *Personality and Social Psychology Review*, *19*(4), 371–394.

Canales, L., & Martinez-Barco, P. (2014). Emotion detection from text: A survey. In *Proceedings of the Workshop on Natural Language Processing in the 5th Information Systems Research Working Days (JISIC)* (pp. 37–43).

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chalmers, R. P. (2018). Sign for residuals [online forum post]. In *Mirt Google Group*. https://groups.google.com/g/mirt-package/c/2lpp63VdJMY/m/F7CIMsr9AQAJ

Chen, C.-W., Wang, W.-C., Mok, M. M. C., & Scherer, R. (2021). A lognormal ipsative model for multidimensional compositional items. *Frontiers in Psychology*, *12*.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464–504.

Chen, F., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods & Research*, *36*(4), 462–494.

Chen, S.-Y., Hsu, C.-C., Kuo, C.-C., Ku, L.-W., et al. (2018). Emotionlines: An emotion corpus of multi-party conversations. *arXiv Preprint arXiv:1802.08379*.

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*, 265–289.

Chowdhury, G. G. (2003). Natural language processing. *Annual Review of Information Science and Technology*, *37*(1), 51–89. https://doi.org/https://doi.org/10.1002/aris.1440370103

Christensen, A. P., Cotter, K. N., & Silvia, P. J. (2019). Reopening openness to experience:

A network analysis of four openness to experience inventories. *Journal of Personality Assessment*, *101*(6), 574–588.

Clore, G. L., & Ortony, A. (1988). The semantics of the affective lexicon. In V. Hamilton, G. Bower, & N. Frijda (Eds.), *Cognitive perspectives on emotion and motivation* (pp. 367–397). Springer.

Coenders, G., Hlebec, V., & Kogovšek, T. (2011). Measurement quality in indicators of compositions. A compositional multitrait-multimethod approach. *Survey Research Methods*, *5*(2), 63–74.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge.

Cordaro, D. T., Sun, R., Keltner, D., Kamble, S., Huddar, N., & McNeil, G. (2018). Universals and cultural variations in 22 emotional expressions across five cultures. *Emotion*, *18*(1), 75–93.

Cowen, A. S., & Keltner, D. (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, *114*(38), E7900–E7909.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302.

Darwin, C. (1872). *The expression of the emotions in man and animals*. John Murray.

De Choudhury, M., Gamon, M., & Counts, S. (2012). Happy, nervous or surprised? Classification of human affective states in social media. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 6).

DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement*, *43*(2), 145–168.

DiGirolamo, M. A., & Russell, J. A. (2017). The emotion seen in a face can be a methodological artifact: The process of elimination hypothesis. *Emotion*, *17*(3), 538–546.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement

with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, *38*(1), 67–86.

Eckes, T., & Baghaei, P. (2015). Using testlet response theory to examine local dependence in c-tests. *Applied Measurement in Education*, *28*(2), 85–98.

Ekman, P. (1992a). An argument for basic emotions. *Cognition & Emotion*, *6*(3-4), 169–200.

Ekman, P. (1992b). Are there basic emotions? *Psychological Review*, *99*(3), 550–553.

Ekman, P. (2016). What scientists who study emotion agree about. *Perspectives on Psychological Science*, *11*(1), 31–34.

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, *17*(2), 124–129.

Ekman, P., Friesen, W. V., O'Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., Krause, R., LeCompte, W. A., Pitcairn, T., Ricci-Bitti, Pio E, Scherer, K., Tomita, Masatoshi, & Tzavaras, A. (1987). Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology*, *53*(4), 712–717.

Ekman, P., & Heider, K. G. (1988). The universality of a contempt expression: A replication. *Motivation and Emotion*, *12*(3), 303–308.

Ekman, P., Sorenson, E. R., & Friesen, W. V. (1969). Pan-cultural elements in facial displays of emotion. *Science*, *164*(3875), 86–88.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*(1), 179–197.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Psychology Press.

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling*, *8*(3), 430–457.

Fabrigar, L. R., & Wegener, D. T. (2011). *Exploratory factor analysis*. Oxford University

Press.

Felt, J. M., Castaneda, R., Tiemensma, J., & Depaoli, S. (2017). Using person fit statistics to detect outliers in survey research. *Frontiers in Psychology*, *8*, 863.

Filzmoser, P., Hron, K., Reimann, C., & Garrett, R. (2009). Robust factor analysis for compositional data. *Computers & Geosciences*, *35*(9), 1854–1861.

Fiske, A. P. (2020). The lexical fallacy in emotion research: Mistaking vernacular words for psychological entities. *Psychological Review*, *127*(1), 95–113.

Fontaine, J. R., Scherer, K. R., Roesch, E. B., & Ellsworth, P. C. (2007). The world of emotions is not two-dimensional. *Psychological Science*, *18*(12), 1050–1057.

Freud, S. (1901). The psychopathology of everyday life (Trans. A.A. Brill). In *Classics in the history of psychology*. Green, Christopher D. https://psychclassics.yorku.ca/Freud/Psycho/index.htm

Gehlenborg, N. (2019). *UpSetR: A more scalable alternative to Venn and Euler diagrams for visualizing intersecting sets*. https://CRAN.R-project.org/package=UpSetR

Gendron, M., Crivelli, C., & Barrett, L. F. (2018). Universality reconsidered: Diversity in making meaning of facial expressions. *Current Directions in Psychological Science*, *27*(4), 211–219.

Glenn, J. J., Nobles, A. L., Barnes, L. E., & Teachman, B. A. (2020). Can text messages identify suicide risk in real time? A within-subjects pilot examination of temporally sensitive markers of suicide risk. *Clinical Psychological Science*, *8*(4), 704–722.

Golino, H. F., & Christensen, A. P. (2021). *EGAnet: Exploratory graph analysis – a framework for estimating the number of dimensions in multivariate data using network psychometrics*. R package version 0.9.9

Golino, H. F., & Epskamp, S. (2017). Exploratory graph analysis: A new approach for estimating the number of dimensions in psychological research. *PloS One*, *12*(6), e0174035.

Golino, H. F., Moulder, R., Shi, D., Christensen, A. P., Garrido, L. E., Nieto, M. D.,

Nesselroade, J., Sadana, R., Thiyagarajan, J. A., & Boker, S. M. (2021). Entropy fit indices: New fit measures for assessing the structure and dimensionality of multiple latent variables. *Multivariate Behavioral Research*, *56*(6), 874–902.

Golino, H. F., Shi, D., Christensen, A. P., Garrido, L. E., Nieto, M. D., Sadana, R., Thiyagarajan, J. A., & Martinez-Molina, A. (2020). Investigating the performance of exploratory graph analysis and traditional techniques to identify the number of latent factors: A simulation and tutorial. *Psychological Methods*, *25*(3), 292–320.

Gollapalli, S. D., Rozenshtein, P., & Ng, S. K. (2020). ESTeR: Combining word co-occurrences and word associations for unsupervised emotion detection. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 1043–1056).

Gorsuch, R. L. (2014). *Factor analysis: Classic edition*. Routledge.

Greenacre, M. (2021). Compositional data analysis. *Annual Review of Statistics and Its Application*, *8*(1), 271–299.

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods*, *6*(4), 430–450.

Guerini, M., & Staiano, J. (2015). Deep feelings: A massive cross-lingual study on the relation between emotions and virality. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 299–305).

Hasan, M., Rundensteiner, E., & Agu, E. (2019). Automatic emotion detection in text streams by analyzing Twitter data. *International Journal of Data Science and Analytics*, *7*(1), 35–51.

Haslam, N., & Bornstein, B. H. (1996). Envy and jealousy as discrete emotions: A taxometric analysis. *Motivation and Emotion*, *20*(3), 255–272.

Jack, R. E. (2013). Culture and facial expressions of emotion. *Visual Cognition*, *21*(9-10), 1248–1286.

Jack, R. E., Garrod, O. G., & Schyns, P. G. (2014). Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. *Current Biology*, *24*(2), 187–192.

Jack, R. E., Sun, W., Delis, I., Garrod, O. G., & Schyns, P. G. (2016). Four not six: Revealing culturally common facial expressions of emotion. *Journal of Experimental Psychology: General*, *145*(6), 708–730.

Jackson, J. C., Watts, J., Henry, T. R., List, J.-M., Forkel, R., Mucha, P. J., Greenhill, S. J., Gray, R. D., & Lindquist, K. A. (2019). Emotion semantics show both cultural variation and universal structure. *Science*, *366*(6472), 1517–1522.

Jennrich, R. I., & Bentler, P. M. (2011). Exploratory bi-factor analysis. *Psychometrika*, *76*, 537–549.

Jennrich, R. I., & Bentler, P. M. (2012). Exploratory bi-factor analysis: The oblique case. *Psychometrika*, *77*(3), 442–454.

Jockers, M. L. (2015). *Syuzhet: Extract sentiment and plot arcs from text*. https://github.com/mjockers/syuzhet

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Scientific Software International.

Kang, T., & Chen, T. T. (2007). *An investigation of the performance of the generalized S-X2 item-fit index for polytomous IRT models*. ACT.

Kim, E. S., & Yoon, M. (2011). Testing measurement invariance: A comparison of multiple-group categorical CFA and IRT. *Structural Equation Modeling*, *18*(2), 212–228.

Kline, R. B. (2005). *Principles and practice of structural equation modeling*.

Kušen, E., Cascavilla, G., Figl, K., Conti, M., & Strembeck, M. (2017). Identifying emotions in social media: Comparison of word-emotion lexicons. In *2017 5th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW)* (pp. 132–137). IEEE.

Lamprinidis, S., Bianchi, F., Hardt, D., & Hovy, D. (2021). Universal joy: A data set and results for classifying emotions across languages. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (pp. 62–75).

Lasswell, H. D., Lerner, D., & Pool, I. de S. (1952). *Comparative study of symbols: An introduction*. Stanford University Press.

Lauritzen, S. L. (1996). *Graphical models* (Vol. 17). Clarendon Press.

Laville, M., Bouhandi, M., Morin, E., & Langlais, P. (2020). Word representations, seed lexicons, mapping procedures, and reference lists: What matters in bilingual lexicon induction from comparable corpora? In C. Goutte & X. Zhu (Eds.), *Advances in Artificial Intelligence* (pp. 349–355). Springer International Publishing.

Liu, C., Osama, M., & De Andrade, A. (2019). DENS: A dataset for multi-class emotion analysis. *arXiv Preprint arXiv:1910.11769*.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley Publishing Company.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*(1), 84–99.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing hu and bentler's (1999) findings. *Structural Equation Modeling*, *11*(3), 320–341.

Massara, G. P., Di Matteo, T., & Aste, T. (2017). Network filtering for big data: Triangulated maximally filtered graph. *Journal of Complex Networks*, *5*(2), 161–178.

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*(4), 713–732.

Maydeu-Olivares, A., & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research*, *49*(4), 305–328.

McNair, D. M., Lorr, M, & Droppleman, L. F. (1971). *Profile of mood states*. Educational and Industrial Testing Service.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, *54*(6).

Mejova, Y., & Kalimeri, K. (2020). COVID-19 on Facebook ads: competing agendas around a public health crisis. In *Proceedings of the 3rd ACM SIGCAS Conference on Computing and Sustainable Societies* (pp. 22–31).

Mesquita, B., Boiger, M., & De Leersnyder, J. (2016). The cultural construction of emotions. *Current Opinion in Psychology*, *8*, 31–36.

Mohammad, S. M. (2012). #Emotional tweets. In *SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)* (pp. 246–255).

Mohammad, S. M. (2018). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of The Annual Conference of the Association for Computational Linguistics (ACL)*.

Mohammad, S. M. (2021). Sentiment analysis: Automatically detecting valence, emotions, and other affectual states from text. In H. L. Meiselman (Ed.), *Emotion Measurement* (2nd ed., pp. 323–379). Woodhead Publishing.

Mohammad, S. M., Bravo-Marquez, F., Salameh, M., & Kiritchenko, S. (2018). SemEval-2018 Task 1: Affect in Tweets. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*.

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, *29*(3), 436–465.

Moors, A. (2017). Integration of two skeptical emotion theories: Dimensional appraisal theory and Russell's psychological construction theory. *Psychological Inquiry*, *28*(1), 1–19.

Muljono, Winarsih, N. A. S., & Supriyanto, C. (2016). Evaluation of classification methods for indonesian text emotion detection. In *2016 International Seminar on Application for Technology of Information and Communication (ISemantic)* (pp. 130–133).

Nabi, R. L. (2002). The theoretical versus the lay meaning of disgust: Implications for

emotion research. *Cognition & Emotion*, *16*(5), 695–703.

O'dea, B., Larsen, M. E., Batterham, P. J., Calear, A. L., & Christensen, H. (2017). A linguistic analysis of suicide-related Twitter posts. *Crisis*, *38*(5), 319–329.

Oaten, M., Stevenson, R. J., Williams, M. A., Rich, A. N., Butko, M., & Case, T. I. (2018). Moral violations and the experience of disgust and anger. *Frontiers in Behavioral Neuroscience*, *12*, 179.

Ophir, Y., Tikochinski, R., Asterhan, C. S., Sisso, I., & Reichart, R. (2020). Deep neural networks detect suicide risk from textual facebook posts. *Scientific Reports*, *10*(1), 1–10.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*(1), 50–64.

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of s-X2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement*, *27*(4), 289–298.

Ortony, A., & Turner, T. J. (1990). What's basic about basic emotions? *Psychological Review*, *97*(3), 315–331.

Paek, I., & Cole, K. (2020). *Using R for item response theory model applications*. Routledge.

Paul Ekman Group. (2021). *Universal emotions*. https://www.paulekman.com/universal-emotions/

Pearson, K. (1896). Mathematical contributions to the theory of evolution. — On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London*, *60*(359-367), 489–498.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. LIWC.

Pestian, J. P., Matykiewicz, P., Linn-Gust, M., South, B., Uzuner, O., Wiebe, J., Cohen, K. B., Hurdle, J., & Brew, C. (2012). Sentiment analysis of suicide notes: A shared task. *Biomedical Informatics Insights*, *5*(Suppl 1), 3–16. https://doi.org/10.4137/bii.s9042

Plutchik, R. (1962). *The emotions*. University Press of America.

Polignano, M., Basile, P., de Gemmis, M., & Semeraro, G. (2019). A Comparison of Word-Embeddings in Emotion Detection from Text Using BiLSTM, CNN and Self-Attention. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization* (pp. 63–68). Association for Computing Machinery. https://doi.org/10.1145/3314183.3324983

Poria, S., Gelbukh, A., Cambria, E., Yang, P., Hussain, A., & Durrani, T. (2012). Merging SenticNet and WordNet-Affect emotion lists for sentiment analysis. In *2012 IEEE 11th International Conference on Signal Processing* (Vol. 2, pp. 1251–1255). IEEE.

Poria, S., Gelbukh, A., Hussain, A., Howard, N., Das, D., & Bandyopadhyay, S. (2013a). Enhanced SenticNet with affective labels for concept-based opinion mining. *IEEE Intelligent Systems*, *28*(2), 31–38.

Poria, S., Gelbukh, A., Hussain, A., Howard, N., Das, D., & Bandyopadhyay, S. (2013b). Enhanced SenticNet with affective labels for concept-based opinion mining. *IEEE Intelligent Systems*, *28*(2), 31–38.

Posner, J., Russell, J. A., & Peterson, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, *17*(3), 715–734.

Quan, C., & Ren, F. (2009). Construction of a blog emotion corpus for chinese emotional expression analysis. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3* (pp. 1446–1454). Association for Computational Linguistics.

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Raghunathan, R., & Pham, M. T. (1999). All negative moods are not equal: Motivational influences of anxiety and sadness on decision making. *Organizational Behavior and Human Decision Processes*, *79*(1), 56–77.

Raji, S., & De Melo, G. (2020). What sparks joy: The AffectVec emotion database. In

Proceedings of The Web Conference 2020 (pp. 2991–2997).

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Nielsen & Lydiche.

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, *21*(1), 25–36. https://doi.org/10.1177/0146621697211002

Robinson, M. D., & Clore, G. L. (2002). Belief and feeling: Evidence for an accessibility model of emotional self-report. *Psychological Bulletin*, *128*(6), 934.

Roseman, I. J., Wiest, C., & Swartz, T. S. (1994). Phenomenology, behaviors, and goals differentiate discrete emotions. *Journal of Personality and Social Psychology*, *67*(2), 206–221.

Rosensohn, W. L. (1963). A logical method for making a classification of emotions, using Wilhelm Wundt's theory of emotion formation. *The Journal of Psychology*, *55*(1), 175–182.

Royzman, E. B., Atanasov, P., Landy, J. F., Parks, A., & Gepty, A. (2014). CAD or MAD? Anger (not disgust) as the predominant response to pathogen-free violations of the divinity code. *Emotion*, *14*(5), 892–907.

Royzman, E. B., & Kurzban, R. (2011). Minding the metaphor: The elusive character of moral disgust. *Emotion Review*, *3*(3), 269–271.

Royzman, E. B., & Sabini, J. (2001). Something it takes to be an emotion: The interesting case of disgust. *Journal for the Theory of Social Behaviour*, *31*(1), 29–59.

Rozin, P., Berman, L., & Royzman, E. B. (2010). Biases in use of positive and negative words across twenty natural languages. *Cognition and Emotion*, *24*(3), 536–548.

Rozin, P., Haidt, J., & McCauley, C. R. (2000). Disgust: The body and soul emotion. In T. Dalgleish & M. Power (Eds.), *Handbook of Cognition and Emotion* (pp. 429–445). Wiley.

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion.

Personality and Social Psychology Review, *5*(4), 296–320.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178.

Russell, J. A., & Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, *11*(3), 273–294.

Salinas, C. M. S., Fontaine, J. R., & Scherer, K. R. (2015). Surprise in the GRID. *Review of Cognitive Linguistics. Published Under the Auspices of the Spanish Cognitive Linguistics Association*, *13*(2), 436–460.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, *34(4, Pt. 2), 100*.

Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika*, *75*(2), 243–248.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177.

Scherer, K. R., & Wallbott, H. G. (1994). Evidence for universality and cultural variation of differential emotion response patterning. *Journal of Personality and Social Psychology*, *66*(2), 310–328.

Schmidt, Karen M, & Embretson, S. E. (2012). Item response theory and measuring abilities. In J. A. Schinka & W. F. Velicer (Eds.), *Handbook of Psychology* (2nd ed., pp. 451–473). John Wiley & Sons, Inc.

Schnall, S., Haidt, J., Clore, G. L., & Jordan, A. H. (2015). Landy and Goodwin confirmed most of our findings then drew the wrong conclusions. *Perspectives on Psychological Science*, *10*, 537–538.

Schrauf, R. W., & Sanchez, J. (2004). The preponderance of negative emotion words in the emotion lexicon: A cross-generational and cross-linguistic study. *Journal of Multilingual and Multicultural Development*, *25*(2-3), 266–284. https://doi.org/https://www.tandfonline.com/doi/abs/10.1080/01434630408666532

Schütze, H., Manning, C. D., & Raghavan, P. (2008). *Introduction to information retrieval*. Cambridge University Press.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

SemEval-2018 Task 1: Affect in Tweets. (2018). In *CodaLab*. SemEval-2018: International Workshop on Semantic Evaluation. https://competitions.codalab.org/competitions/17751

Sharma, S., Mukherjee, S., Kumar, A., & Dillon, W. R. (2005). A simulation study to investigate the use of cutoff values for assessing model fit in covariance structure models. *Journal of Business Research*, *58*(7), 935–943.

Shaver, P., Schwartz, J., Kirson, D., & O'connor, C. (1987). Emotion knowledge: Further exploration of a prototype approach. *Journal of Personality and Social Psychology*, *52*(6), 1061.

Shi, D., Maydeu-Olivares, A., & DiStefano, C. (2018). The relationship between the standardized root mean square residual and model misspecification in factor analysis models. *Multivariate Behavioral Research*, *53*(5), 676–694.

Silge, J., & Robinson, D. (2016). Tidytext: Text mining and analysis using tidy data principles in R. *JOSS*, *1*(3). https://doi.org/10.21105/joss.00037

Staiano, J., & Guerini, M. (2014). Depechemood: A lexicon for emotion analysis from crowd-annotated news. In *52nd Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 427–433).

Strapparava, C., & Valitutti, A. (2004). WordNet-Affect: An affective extension of WordNet. In *Fourth International Conference on Language Resources and Evaluation*. European Language Resources Association.

Sykora, M. D., Jackson, T., O'Brien, A., & Elayan, S. (2013). Emotive ontology: Extracting fine-grained emotions from terse, informal messages. *Proceedings of the IADIS International Conference Intelligent Systems and Agents, Prague*, 19–26.

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2007). *Using multivariate statistics* (5th ed.). Pearson.

Tabak, F. S., & Evrim, V. (2016). Comparison of emotion lexicons. *IEEE HONET-ICT 2016*, 154–158.

Thomas, S. L., Schmidt, K. M., Erbacher, M. K., & Bergeman, C. S. (2016). What you don't know can hurt you: Missing data and partial credit model estimates. *Journal of Applied Measurement*, *17*(1), 14–34.

Tompkins, S. S. (1962). *Affect, imagery, consciousness: Vol I. The positive affects*. Springer.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*(1), 1–10.

Vicario, C. M., Rafal, R. D., Pellegrino, G. di, Lucifora, C., Salehinejad, M. A., Nitsche, M. A., & Avenanti, A. (2020). Indignation for moral violations suppresses the tongue motor cortex: Preliminary TMS evidence. *Social Cognitive and Affective Neuroscience*.

Vorakitphan, V., Cabrio, E., & Villata, S. (2021). "Don't discuss": Investigating semantic and argumentative features for supervised propagandist message detection and classification. In *Recent Advances in Natural Language Processing (RANLP 2021)*. Varna/Virtual, Bulgaria.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, *57*(2), 307–333.

Wang, W., Chen, L., Thirunarayan, K., & Sheth, A. P. (2012). Harnessing Twitter "big data" for automatic emotion identification. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing* (pp. 587–592). IEEE.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191–1207.

Wellman, H. M., Harris, P. L., Banerjee, M., & Sinclair, A. (1995). Early understanding of emotion: Evidence from natural language. *Cognition & Emotion*, *9*(2-3), 117–149.

West, S. G., Taylor, A. B., Wu, W., et al. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 209–231). The Guilford Press.

Wijffels, J. (2022). *udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit*. https://CRAN.R-project.org/package=udpipe

Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist*, *34*(6), 806–838.

Xia, Y., & Yang, Y. (2019). RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. *Behavior Research Methods*, *51*(1), 409–428.

Yang, M., Zhu, D., & Chow, K.-P. (2014). A topic model for building fine-grained domain-specific emotion lexicon. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 421–426).

Zad, S., Jimenez, J., & Finlayson, M. (2021). Hell hath no fury? Correcting bias in the NRC emotion lexicon. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)* (pp. 102–113).

Zhang, B., & Walker, C. M. (2008). Impact of missing data on person—model fit and person trait estimation. *Applied Psychological Measurement*, *32*(6), 466–479.

Zimmerman, C., Stein, M.-K., Hardt, D., & Vatrapu, R. (2015). Emergence of things felt: Harnessing the semantic space of Facebook feeling tags. In *Thirty Sixth International Conference on Information Systems*.

Appendix A

Exploratory Oblimin Model: The Thirty Highest and Lowest Scoring Words Per Factor

F1 ("Anger")		F2 ("Positive")		F3 ("Sadness")		F4 ("Fear")	
Highest	Lowest	Highest	Lowest	Highest	Lowest	Highest	Lowest
abuse	adore	advance	agony	abandon	admirable	agony	admirable
attacking	bless	cash	bastard	abandoned	admiration	anxiety	admiration
bitch	blessing	celebrity	compression	abandonment	adoration	avalanche	adoration
catastrophe	brilliant	destination	destroyer	crushed	adventurer	bloodshed	affection
criticize	champion	excite	discouragement	depressed	appreciation	bomber	appreciation
cruel	charitable	feeling	dowry	deprivation	bounty	cadaver	approve
cruelty	comfort	graduation	dumb	despair	champion	catastrophe	benevolence
destructive	courtship	highest	emptiness	devastate	confident	confine	confident
hate	delightful	infant	empty	devastating	credit	dictatorship	darling
hateful	eagerness	intense	fatigue	doomsday	encourage	epidemic	encourage
hatred	elegance	liberate	feudalism	failure	equality	insecurity	engaging
hell	engaged	lovely	filth	grievous	freedom	invade	excellent
hellish	enjoy	marry	fume	grim	generous	lawsuit	favorite
insulting	enjoying	money	humiliating	hopelessness	hardy	mortification	freedom
jealousy	faith	morals	humiliation	hurt	harmony	nervous	freely
mad	glory	musical	liar	hurtful	heal	nervousness	generous
murder	happy	nurture	miserable	hurting	helpful	overwhelm	harmony
murderer	healing	opera	miss	isolated	hug	rape	heal
murderous	heavenly	powerful	moron	lonely	improvement	risky	helpful
nasty	improve	pray	narcotic	lose	kind	rot	hug
offender	inspire	retirement	neglect	loss	praise	rubble	improvement
offense	lover	romance	overwhelm	misery	promise	scold	kind
poison	magnificence	supremacy	rogue	mournful	proud	terrorism	praise
poisoned	passion	thrill	sigh	rejection	safe	terrorist	promise
poisonous	passionate	treat	sterile	rejects	save	terrorize	safe
prick	pastor	unexpected	suffer	resign	silly	traitor	save
rape	peace	vote	tithe	ruin	tranquility	treachery	tranquility
revenge	perfect	weight	useless	ruined	true	tremor	true
violent	pretty	winning	whine	ruinous	vitality	worrying	vitality
violently	treasure	youth	yearn	tragedy	wealth	worse	wealth

Appendix B

Exploratory Bifactor Model: The Thirty Highest and Lowest Scoring Words Per Factor

General Factor		Group 1		Group 2		Group 3	
Highest	Lowest	Highest	Lowest	Highest	Lowest	Highest	Lowest
agony	admirable	cash	compression	bereavement	admit	agitated	anxiousness
alcoholism	admiration	celebrity	discouraged	cry	adventurer	agitation	avalanche
catastrophe	adoration	destination	discouragement	dull	annex	annoy	confuse
confine	affection	endless	dowry	isolate	arrange	antagonist	confused
criticize	appreciation	excite	dumb	isolation	atrium	antagonistic	confusing
deadly	approve	feeling	empty	longing	ballot	argue	dislocation
destroyer	benevolence	graduation	fatigue	lost	believer	argument	distressingly
feudalism	confident	highest	fume	lower	binoculars	asshole	dragon
hell	darling	honest	gravel	lowest	blend	battle	elements
homicide	encourage	infant	graver	lowly	bounty	cheat	emotional
incarceration	engaging	intense	humiliating	melancholic	capsule	contemptible	encounter
misery	favorite	liberate	humiliation	melancholy	cinder	damn	forceps
murder	freedom	lovely	inadequacy	mourning	coal	despise	handkerchief
murderer	freely	marry	inferiority	pity	corrupt	furious	indecision
nasty	generous	money	liar	regret	diffusion	grating	insecurity
offender	harmoniously	morals	lowering	regrettable	distribute	greed	microbe
opium	harmony	musical	miss	regretted	dugout	harass	nervous
poison	heal	nurture	moron	regretting	enumeration	hostility	nervousness
poisonous	helpful	opera	narcotic	remorse	episode	idiotic	oil
rape	hug	powerful	neglect	resignation	equality	jealous	phobia
rot	improvement	pray	neglecting	resigned	fleet	lying	proxy
schizophrenia	kind	retirement	overwhelm	sadly	invigorate	nag	risk
scold	praise	romance	rejected	sadness	jot	offend	risky
terrorism	promise	supremacy	ruins	sob	knot	offensive	shake
terrorist	safe	treat	sigh	sorrowful	misbehavior	outrage	stress
terrorize	save	unexpected	suffer	unhappiness	recreational	playful	tense
threatening	tranquility	vote	tithe	unimportant	reinforcement	prejudice	uneasiness
traitor	true	weight	tragic	unsuccessful	revenge	shit	uptight
treachery	vitality	winning	useless	weeping	twilight	smother	worried
violently	wealth	youth	yearn	woeful	vicar	vicious	worrying