Improving our classification system for the treatment of individuals who have experienced traumatic events: The contribution of unsupervised statistical learning to our existing methods

A Dissertation Submitted to the Graduate Division of the University of Hawai‘i at Mānoa in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

In

Psychology

By

Michelle Lopez Raab

May 2016

Dissertation Committee:
Charlene Baker, Chairperson
John Barile
David Cicero
Brad Nakamura
Les Wilson
Acknowledgements

This study would not have been possible without the work and support of many people. A big thank you goes to Lu Liang for giving me the data set. I would like to acknowledge Kentaro Hayashi, PhD for his support in the initial phases of this project, and Elaine Heiby, PhD for her feedback about the description of the psychometric properties of the measures used in the data collection of this study. I want to extend my deepest gratitude to Charlene Baker, PhD in the tremendous support in finishing this project as my chair, including making the document readable. I also extend my gratitude to Jack Barile, PhD for his help with the methods section and stepping in at the last moment, and the unwavering support of the rest of my committee: Les Wilson, PhD, Brad Nakamura, PhD, and David Cicero, PhD. Without the emotional support of my family and closest friends, I would not have been able to complete my degree. Thank you John Rolph, Cathy Rolph, and Richelle Daraban. My degree should also have your names on it. You guys have hung in there for me for the past decade. I want to also thank Todd Shayler for helping me to understand math stuff. And finally, I want to thank my husband, Drew Raab, for listening to my thoughts about statistics even though you don’t love statistics.

I thank all of you for your help and support. I would not have been able to do this without you all.
Abstract

Rationally derived theories have had a limiting effect on the advancement of psychology as a science, compared to theories born out of or tested by empirical studies. As an example, while the diagnostic system (DSM) has been informed by science, the categories have not often been empirically derived (DSM-I, 1953; DSM-II, 1968; DSM-III, 1980, DSM-IV-TR, 2000; DSM-5, 2013). There is an emerging inclusion of empirical methods in the diagnostic classification system, as seen with some diagnostic categories of the DSM-5 (2013; Krueger, Derringer, Markon, Watson, & Skodol, 2012); however, there are many criteria and categories that have gone untested (Kramer et al., 2016). And, simply using hypothesis testing may not be sufficient in generating new knowledge. To improve our methods, we add to our current research and statistical methods through the use of unsupervised statistical learning, where data are allowed to tell their own story. Two statistical learning techniques, k-means cluster analysis and finite mixture modeling (Duda, Hart, & Stork, 2012; Hastie et al., 2009; James, Witten, Hastie, & Tibshirani, 2013) were applied to a data set collected on university students who had been displaced in the aftermath of Hurricane Katrina to understand the relationship between resource loss and stress. These techniques were used to demonstrate how to explore the data so that unanticipated knowledge could be distilled from the data.

Findings showed that this data set was not easily studied using k-means cluster analysis, because the structure of the multivariate data did not contain clearly defined subgroups. Exploring the data using finite mixture modeling did, however, yielded possible areas of further research, such as the relationship of gains and losses to items on the depressive scale. However, conclusions about the relative performance of these techniques should not be made without the
use of simulated data. This research study demonstrated the importance of expanding our
techniques to explore what the data can tell us, as the findings would not have been revealed had
the data only been explored by using hypothesis testing. Future research should include
unsupervised statistical learning as a method to advance knowledge and classification in
psychological research.
Table of Contents

Acknowledgements ......................................................................................................................... 2
Abstract ............................................................................................................................................. 3
Table of Contents ............................................................................................................................ 5
Table of Figures ............................................................................................................................... 7
Table of Tables ................................................................................................................................. 8
Chapter 1. Introduction .................................................................................................................... 9
Chapter 2. Literature Review .......................................................................................................... 12
  Rationally Derived Theories in Psychology: Then and Now ......................................................... 12
  Developing a Definition for the Events that Make People Emotionally Distressed ................. 18
  Conservation of Resource Model ............................................................................................... 21
  The Importance of Accurate Classification ............................................................................... 22
  Data behavior reflects patterns of people’s actual experiences ............................................... 23
  Why look for subgroups? ............................................................................................................ 24
  Unsupervised Statistical Learning Techniques ......................................................................... 25
  Visualizing k-means cluster analysis ....................................................................................... 27
  Visualizing finite mixture modeling (FMM) ............................................................................. 28
  Characteristics of the models: Things to know for interpretation ........................................... 29
  Research Goals ......................................................................................................................... 34
Chapter 3: Methods ....................................................................................................................... 37
  Description of Data Set ............................................................................................................. 37
  Parent data set ......................................................................................................................... 37
  Raw data set ............................................................................................................................ 37
  Item-level data set .................................................................................................................... 37
  Measures ................................................................................................................................ 39
  Center for Epidemiologic Studies Depressive Scale (CES-D) .................................................. 39
  COR: Conservation of Resources ............................................................................................. 41
  Data Handling .......................................................................................................................... 42
  Missing Data ............................................................................................................................ 43
  Data Analysis ............................................................................................................................... 46
  Preprocessing the Data with Data Reduction: Principle Component Analysis ...................... 47
  Initial Values for the Parameters Being Estimated by the Algorithms of K-means and FMM ........................................................................................................................................... 47
  Finding K: The Number of Subgroups ....................................................................................... 48
Chapter 4. Results ............................................................................................................................ 50
  Data Reduction: Benefits and Costs ........................................................................................... 50
  Benefit: Solution stability ......................................................................................................... 51
  Dimension reduction: Benefits and costs summary .................................................................. 53
  Finding K: The Number of Subgroups ....................................................................................... 53
  Analyzing the k-means cluster solution for K = 2: Demonstrating the lack of separation and clear boundaries within the data ....................................................................................... 53
  Finding Outliers ....................................................................................................................... 56
# Table of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1</td>
<td>Summary of Missing Data by Variables, Values, and Cases</td>
<td>44</td>
</tr>
<tr>
<td>Figure 2</td>
<td>Missing Value Patterns</td>
<td>45</td>
</tr>
<tr>
<td>Figure 3</td>
<td>Frequency of Missing Value Patterns by Pattern</td>
<td>46</td>
</tr>
<tr>
<td>Figure 4</td>
<td>K = 2 Item Means by Each Component</td>
<td>60</td>
</tr>
<tr>
<td>Figure 5</td>
<td>Stacked FMM Two Component Model</td>
<td>61</td>
</tr>
<tr>
<td>Figure 6</td>
<td>K = 3 Item Means by Component</td>
<td>64</td>
</tr>
<tr>
<td>Figure 7</td>
<td>Stacked FMM Item-Level Three Component</td>
<td>65</td>
</tr>
<tr>
<td>Figure 8</td>
<td>K = 4 Item Means by Component</td>
<td>68</td>
</tr>
<tr>
<td>Figure 9</td>
<td>Stacked FMM Item-Level 4 Component Model</td>
<td>69</td>
</tr>
<tr>
<td>Figure 10</td>
<td>FMM 4 Component Detail of Depressive Symptoms for Components 1 and 2</td>
<td>70</td>
</tr>
<tr>
<td>Figure 11</td>
<td>Conceptual Model of 4 Component Solution</td>
<td>71</td>
</tr>
</tbody>
</table>
Table of Tables

Table 1: Participant Frequency by Year and Gender ................................................................. 34
Table 1: Participant Frequency by Year and Gender ................................................................. 38
Table 2: Participant Ethnicity Frequency .................................................................................. 38
Table 3: Proportion of Membership for K Group by Set Seed Value for Project Data Set .... 51
Table 4: Proportion of Membership for K Group by Set Seed Value for PCA 4-Component Solution .................................................................................................................................................. 52
Table 5: Proportion of Membership for K Group by Set Seed Value for PCA 9-Component Solution .................................................................................................................................................. 52
Table 6: Proportion of Membership for K Group by Set Seed Value for PCA 19-Component Solution .................................................................................................................................................. 52
Table 7. K = 2 Sum of Squares Item-level .................................................................................... 53
Table 8. K = 2 Sum of Squares PCA 4-Component ..................................................................... 55
Table 9. K = 2 Sum of Squares 9-Component .......................................................................... 55
Table 10. K = 2 Sum of Squares 19-Component ......................................................................... 56
Table 11: Loglikelihood Trials 1-5 for K = 2-5 Project Data Set .................................................. 58
Table 12. K = 2 Demographic Means by Component ................................................................. 59
Table 13. K = 3 Demographic Means by Component ................................................................. 63
Table 14. K = 4 Demographic Means by Component ................................................................. 67
Chapter 1. Introduction

Historically, rationally derived theories on the origin and treatment of mental illnesses have led to thousands of years of debates on the effects of phlegm, bile, star constellations, and demonic possession on the psyche and the use of bloodletting and exorcisms for the relief of those effects (Radden, 2002). In more recent times, therapy based on ego personality theory seemed to inhibit a precise understanding of how psychopathology developed, perhaps in part because the theory’s foundation was imprecisely developed. Ego personality theory was rationally derived, and research was done to prove the theory rather than test it (Freud, 1920). There have since been therapies that were at first conceptualized through observation and hypothesis generation, but were connected to empirically derived theories. One such therapy is cognitive behavioral therapy (CBT), which originated as trial and error but was grounded in learning theory. Learning theory was developed by empirical means (Goodwin, 2015).

In psychopathology research, there has been some interest in finding methods to test long held understandings of knowledge in psychology. For example, researchers are questioning the meta-structure of our classification system, where they suggest that the current science may indicate that the classification structure is not optimal (Carragher, Krueger, Eaton, & Slade, 2015). In psychopathology research, science can help us understand why and how we suffer and to find ways to alleviate that suffering, maybe even prevent it. To achieve that goal, science requires rigorous testing of our knowledge.

Quantitative empirical research requires the use of numerical data that is based on a testable and measurable representation of a concept that is also testable in its existence (Allen & Yen, 2001). Generating hypotheses that are testable can increase the veracity of the conclusions.
However, hypothesis testing alone may be insufficient and can be augmented through exploratory techniques, because there may be interesting patterns in the data that may not reveal themselves if the data are not fully explored.

One way to explore the data is through the use of an analytical method called unsupervised statistical learning (Duda et al., 2012). K-means cluster analysis and finite mixture modeling (FMM) are commonly used unsupervised statistical learning techniques in psychopathology research. Another method that has been used to examine the classification of groups in psychological research is exploratory factor analysis (Krueger et al., 2012). The basic difference between methods like k-means cluster analysis and FMM versus exploratory factor analysis is that the former finds covariance between participants’ response patterns and the latter finds covariance between variables (Hastie et al., 2009). With k-means and FMM, the researcher is interested in how individuals are grouped together. With exploratory factor analysis, the researcher is interested in finding latent variables, which is to say the larger construct that is defined by the variables. Since the intent of the current study was to examine commonalities of participants’ responses, k-means cluster analysis and FMM were used rather than exploratory factor analysis.

Although both k-means and FMM have been used for discovering patterns in the data, there are differences between cluster analysis and FMM. Cluster analysis partitions the data according to similarity as measured by distance between data points (Duda et al., 2012); whereas, FMM decomposes subpopulations, also known as subgroups, using probabilistic functions rather than Euclidean distance functions (McLachlan & Peel, 2000). The algorithm that was used and is the most current method for estimating subpopulations is called the expectation-maximization algorithm or EM (McLachlan & Peel, 2000). It is an iterative
algorithm that alternates between estimating the parameters, assigning members, and recalculating the parameters that maximize the likelihood evaluation function (McLachlan & Peel, 2000). Using one or the other will lead to differences in what can be said about the data, and by extension how we understand human suffering as well as how to alleviate it.

With this in mind, the purpose of this study is to provide an illustration of how unsupervised statistical learning can be used to investigate what the data may be able to tell us, in this case where the data are related to human suffering. One source of human suffering occurs after traumatic events that we have no control over, like natural disasters. People can suffer emotional distress or other adverse reactions after a natural disaster (Friedman, Resick, Bryant, & Brewin, 2011). Using an archival data set that measured the resources lost and emotional distress of university students displaced by the aftermath of Hurricane Katrina, unsupervised statistical learning techniques were applied. The process of this application demonstrates how to use unsupervised statistical learning for the purpose of letting the data speak for themselves, with the goal of advancing our methods and current practice for treating people who have experienced traumatic events.
Chapter 2. Literature Review

To provide context for the present study, three topics will be covered in this review of the literature. First, to explore the dangers of rationally derived theories, I will examine the history of defining and describing depression (one response to experiencing a traumatic event) and codification as a diagnostic class. Next, I will discuss the history of research on stress, as one of the hypothesized causes of depression. Third, I will describe what unsupervised statistical learning is and the techniques it employs. Finally, I will present research questions for the current study.

Rationally Derived Theories in Psychology: Then and Now

Prolonged experiences of sadness, sorrow, or grief (and occasionally anxiety symptoms) were called melancholy, which was thought to be caused by an imbalance in a mythical emotional hydraulic system called humors (Radden, 2002). From the Ancient Greco-Roman Period until the Romantic Period (nearly two millennia), humor imbalance was the predominant explanation of sadness, sorrow, grief, with the occasional addition of demonic possession or spiritual weakness or influences of the stars within the astrology tradition (Radden, 2002). In the two thousand years of debate and treatise exposition, there were no empirical studies that tested the existence of the humor system, demonic possession, or the influence of star constellations on temperament and mood (Radden, 2002). They were reasoned and argued into existence.

Humors, demons, and stars constellations as explanatory agents were replaced by malfunctions of pathological personalities, as purported by psychiatrists in the late nineteenth to early twentieth century (Freud, 1922; Radden, 2002). Although this explanation included the possibility that the reaction could be normal and not pathological, there were time limits as to how long someone could be sad in order for the reaction not to be a manifestation of a
pathological ego (Freud, 1922). If someone, after a loss, was sad for longer than what the late industrial revolution psychiatry luminaries believed was long enough, then the dim affective state was a result of a malfunctioning ego, a purported subcomponent to the rationally imagined personality construct, an ill-defined emergent property of the self-awareness in humans. The edge between the duration of a normal grief reaction and an abnormal malfunctioning ego reaction was based on analyzing case studies without measurement, without experimentation, but solely on the rationing of the psychiatric literati, a method that is sometimes called armchair theories. The existence of the ego and personality were reasoned into existence. The so-called proof tended to include tautologies, without any examination that was actually testable (Freud, 1920, 1922; Radden, 2002).

In the mid-century, there were psychoanalysts who began questioning the veracity and usefulness of therapy based on ego personality theory. One such psychologist was Albert Ellis. Through his clinical work, he noticed that one of the techniques in psychotherapy was to search for the origins of where belief patterns began (McMahon & Vernon, 2010). He saw a connection between that technique and learning theory (McMahon & Vernon, 2010), developed by researchers such as Ivan Pavlov (Goodwin, 2015). In his practice, he began approaching therapy as a way to uncover learned responses and to extinguish that response (McMahon & Vernon, 2010). This is an interesting example of a therapeutic technique that was developed by trial and error, so not empirical, but grounded in an empirically derived theory.

There are other techniques that have been developed through trial and error, but also were not grounded in a well-established empirically based theory. One such technique is referred to as eye movement desensitization or EMDR (Davidson & Parker, 2001). Studies have shown that in comparison with desensitization techniques, which are based on learning theory, that EMDR
does not perform as well (Davidson & Parker, 2001). In a dismantling study, it was found that
the eye movement did not contribute to relief of symptoms (Cahill, Carrigan, & Frueh, 1999).
Other studies have suggested that EMDR is no more effective than no treatment (Davidson &
Parker, 2001). However, this technique remains controversial. EMDR is an interesting example
of a therapeutic technique that is not based in a well-established and empirically based theory,
but may work. When it does, it may not work because of the mechanism purported by the
developers. In other words when this technique does work in the reduction of symptoms, it may
not be because of the eye-movement.

As these examples show, the conceptualization of the origin of specific types of
psychopathology and their treatment continues to be a work in progress. That said, there is cause
for optimism. For example, the development of cognitive behavior therapy (CBT) is a good
illustration of how the field is trying to use empirical research combined with theory to develop
therapies. Albert Ellis, a trained psychoanalyst, observed that techniques used in psychoanalysis
were in essence examining where behaviors and believes were learned (McMahon & Vernon,
2010). Ellis hypothesized that if behaviors and beliefs are learned then they could be
extinguished (McMahon & Vernon, 2010). Learning theory, such as conditional learning theory
developed by Ivan Pavlov, are empirically derived theories (Goodwin, 2015). Ellis called his
therapy rational emotive behavior theory, but it is considered one form of CBT (McMahon &
Vernon, 2010). Therefore, CBT was based on learning theory, an empirically derived theory,
and has since been extensively tested through empirical means (McMahon & Vernon, 2010). In
the field of psychology not only has there been a movement to develop therapies that are
empirically based, but there has been a movement towards empirically based classification of
mental health disorders.
Therefore, in addition to advances made through new empirically derived therapies, such as CBT, the codification of psychopathology categories has made improvements, though it has been a long process. In America, after decades of competing classification systems, the American Psychiatric Association put forth its version of a classification system, which was the first Diagnostic and Statistical Manual or the DSM (American Psychiatric Association, 1952). At this time affective disorders were groups under Psychotic Disorders and anxiety related disorders were groups under Psychoneurotic Disorders (DSM-I, 1952). For both categories, the explanation system focused on how the so-called “personality” reacted to stressors, internal or external (DSM-I, 1952). There was no explanation as to why a “personality” would react one way or another, nor was there any explanation as to what a personality was.

Additionally, there were no empirical studies cited that explained the existence of the id-ego-superego-based personality and the existence of malfunctions leading to reactions like Psychotic or Psychoneurotic. Also, there was no explanation as to why anxiety would be grouped separately from depression, other than the categorization criteria were based on the rationally derived theory of malfunctioning personalities.

The second edition of the DSM (DSM-II, 1968) continued in the vein of the first with its psychodynamic tendencies (Mayes & Horwitz, 2005). Then, researchers found that the diagnostic categories were not reliable (Spitzer, Endicott, & Robins, 1978), which became a problem as then research results were not comparable across studies. The third edition was revolutionary in that instead of general descriptions of the disorders, there was a list of criteria, and it was atheoretical (i.e., not based on personality theory), in that the descriptions of the disorders were not based on a particular psychological theory (DSM-III, 1980; Mayes & Horwitz, 2005). The subsequent versions of the DSMs retained the criteria based format.
As a positive step, with the most recent DSM publication, there have been new developments in the creation of diagnostic categories, where some of the categories are being created and defined through quantitative empirical means. For example, Krueger et al., (2012) developed categories for personality disorders. In three phases, they used traits distilled from the DSM-5 personality disorder workgroup and proceedings from workgroup discussions, and used exploratory factor analysis to pair down the initial list of traits (Krueger et al., 2012). At the end of the three phases, they were able to identify 25 characteristics associated with maladaptive personalities, including anxiousness combined with suspiciousness. They were also able to develop a clinical inventory for assessment purposes (Krueger et al., 2012). The work of the DSM-5 workgroup is an example of using numerical methods to develop a diagnostic classification category. The current study endeavors to add to the conversation about diagnostic category development improvements.

Interestingly, even with these recent advancements, one feature that has remained constant since the inception of the DSM creation was that anxiety disorders and mood disorders were in different taxonomic groups. However, problems arise if these disorders are not, in fact, separate disorders, but are part of a larger construct. If the separation between depression and anxiety is based solely on an artifact of the personality theory of psychopathology, data collected using these categories run the risk of not actually having construct validity (i.e., that the construct actually exists) (Allen & Yen, 2001). In other words, it may not be ontologically sound.

As mentioned before, the reason that the two disorders were originally separated was because they were thought to be separate malfunctions of the ego, a theory that has since been disproven. In fact, Watson (2009) found that some of the anxiety syndromes have more in common with depression than with other anxiety syndromes. Therefore, one problem that arises
is that research based on faulty categories may not fully capture the etiology of the disorder, which then affects the treatment of the disorder. If this is the case, our knowledge would be based on inaccurate descriptions of the disorders. From a public health perspective, faulty categories may fail to identify those in need of treatment. Watson and colleagues (2009; 2008) also pointed out that heterogeneity is inherent in how the diagnostic categories are defined. This may explain high comorbidity rates (Watson, 2009; Watson, Clark, & Tellegen, 1988; Watson et al., 2008).

Much of psychopathology research is done by categories (Watson, 2009; Watson et al., 1988; Watson et al., 2008). The criteria and division of categories may not be as accurate as it could be, which is why some researchers have recently taken an interest in testing long-held beliefs in our diagnostic system (Krueger, Hopwood, Wright, & Markon, 2014a, 2014b; Mewton, Slade, Teesson, Memedovic, & Krueger, 2014). One criticism of the criteria-based DSMs is that the criteria are not dimensional (Krueger et al., 2014a; Watson, 2009; Watson et al., 1988; Watson et al., 2008). Using item-level data allows for more information in the analysis than using dichotomous variables. Using dichotomous variables does not lend itself well to an overall and integrated understanding of human suffering and the relief of that suffering. From a human standpoint, it limits our understanding of why people suffer and how to help them.

Whether or not anxiety and depression are part of larger constructs, the research suggests that when trying to understand underlying structures of psychological phenomena, analyzing data based at the item-level may yield more accurate results than using already defined constructs in the analysis (Watson, 2009; Watson et al., 1988; Watson et al., 2008). To understand the mechanism for how people react to traumatic or stressful events, using finer measurements (i.e., item-level responses) may prove to be a better approach for unlocking the how’s and why’s. To
aid in our understanding, it is also important to take a closer look at what is contemporarily labeled as stress.

**Developing a Definition for the Events that Make People Emotionally Distressed**

Historically, there has been a division of research between the psychological symptoms, as seen with the discussion of diagnostic systems, and the cause of those symptoms, specifically stress. According to modern research on stress, there are two main components in stress models: the stressor (i.e., the event causing the discomfort) and the stress (e.g., the reaction to the stressor; Hobfoll, 1988).

The term “stress” was first coined by Cannon (1932) to describe the effects of emotions on the body. He is an early user of the term homeostasis in connection with emotions and the physical body (Cannon, 1932). Some twenty years later Selye (1950a, 1950b) used the term stress, again to describe the effects on the body, but also as a process an organism goes through in defending itself. These early models emphasized the person’s response to the precipitating event (Cannon, 1932; Selye, 1950a, 1950b, 1952). Two main criticisms of these models were: 1) purported uniform response to stressors, and 2) that we can only identify if someone experienced stress once he or she was adapting to the event (Appley & Trumbull, 1986; Lazarus, 1966; Lazarus & Folkman, 1984). Because stress could not be identified until the person or organism had already gone through distress and was adapting to the precipitating event, research in this area became difficult.

Another problem is that these models do not account for the characteristics of the events. About ten years after Cannon began writing about the psychological effects of physical stress, Lindemann (1944) proposed that the distress that followed an unpleasant/taxing event is not necessarily a product of ego malfunction, but that anyone could have a similar response to a
distressing event. However, Freud did not allow for so-called natural reactions in the cases where the severity and duration were more than what was usually observed; this suggested to Freud that the individual had a malfunctioning ego (Freud, 1922). In contrast, Lindemann suggested that it was because the individual did not fully engage in what he called grief work (Lindemann, 1944), deemphasizing the pathology of the person and emphasizing what the person was not doing.

Caplan (1964) expanded on this normative viewpoint, meaning that it is normal and not pathological to be distressed after the death of a loved one, and argued that stress reactions are normal in that given a particular event anyone would also be distressed. Loss of a loved one, or bereavement, research is an offshoot of Lindemann’s and Caplan’s propositions in that everyone will experience distress after a loved one dies, but it is the variance in the reactions that needs to be researched and explained (Parkes, 1970, 1972). If characteristics of events can induce a stress reaction, then here must be something about the events that cause that reaction.

To illustrate this point, let us begin with the most extreme type of distressing event: traumatic events. There is a difference between a stressful event and a traumatic event. The current literature makes it quite clear that not every event can be considered traumatizing (Weathers & Keane, 2007). Events can be unpleasant, stressful even, but there are very specific characteristics of the precipitating event that make it traumatic in that there are specific characteristics of the event that predict a traumatic reaction, such as when the event threatens life and/or limb or causes actual injury (including injury of self and body by sexual assault) or death (e.g., the reaction known as PTSD; Friedman et al., 2011; Weathers & Keane, 2007).

In the same vein, not all unpleasant events can be considered stressful (Hobfoll, 1989). Rather there is a combination of aspects of the event that make it stressful, which is what the
Conservation of Resources (COR) model, and others like it, attempts to explain. Although stressful events and hassles can lead to deleterious effects, a recent study showed that they act independently of each other, suggesting that the event and the reaction to the event is different (Aldwin, Jeong, Igarashi, Choun, & Spiro, 2014; Aldwin, Levenson, Spiro, & Bossé, 1989). Next, researchers began looking at the events themselves for these characteristics. This began a debate over which type of event can be considered stressful (Dohrenwend, 1984; Dohrenwend, Dohrenwend, Dodson, & Shrout, 1984; Holmes & Rahe, 1967). Research teams debated over lists of events that could be considered stressful. Others began examining the characteristics of those events that would explain why an event was stressful.

Around the same time period that the characteristics of stressful events were being examined and debated, another model was being developed. Theorists began positing the notion that stress was caused by an imbalance of an organism’s response capacity (e.g., available resources and ability to use those resources) and the demands of external environmental events (Lazarus, 1966; Lazarus & Folkman, 1984; McGrath, 1970). This type of model has gone by names such as homeostatic and transaction models or appraisal-based models. The appraisal label stemmed from a cognitive component of the developing model, where the demand of the environment may not be actual, but perceived (Appley & Trumbull, 1986). The criticisms for this model were that: 1) the logic created a tautology, and 2) stress lies in perception, which alludes to the inherent pathological model of ego-psychologists (Dohrenwend, 1978, 1984; Dohrenwend et al., 1984; Kasl, 1978). The model relies on the two components defining the other: perceived demand and stress reaction. Although these theorists attempted to integrate the psychological, biological, and environmental processes, the resulting model was not supported.
Conservation of Resource Model

In contrast to the many previous models of stress, Hobfoll (1988, 1989, 2001, 2004); Hobfoll and Lilly (1993) developed a stress model that posits that resource loss is a causal mechanism for stress. The premise of the model is that “individuals strive to obtain, retain, protect, and foster those things that they value” (Hobfoll, 2001, p. 341). The valued items are referred to as resources. These resources are valued because they either have an intrinsic value stemming from meeting a survival need or are a means to maintaining or obtaining an item that would meet survival needs (Hobfoll, 1989; Hobfoll & Lilly, 1993). There are four categories of resources: 1) objects, 2) conditions, 3) personal characteristics, and 4) energies (Hobfoll, 1988, 1989; Hobfoll & Lilly, 1993). First, objects are tangible items with an intrinsic value or with an added status value, both of which can meet a survival need but with the status valued item costing more to obtain, which in turn adds to the value of the item. A car can help a person go to work, but a Bentley has the added status value because it costs more money than other cars. Second, conditions are environmental resources that aid in obtaining resources or help create stability in the maintenance/protection of resources, such as living with someone. Third, personal characteristics are personal skills (things that you can do or are a part of the individual’s make-up), like job skills, intelligence, social skills, that aid in obtaining or maintaining resources. Finally, energies are intangible items that required effort to obtain or are intangible items that a person possesses that in turn can help in obtaining or maintaining resources, such as money (effort), knowledge (effort), or time (possessions). Therefore, stress is a result of threatened loss of a resource, actual loss of a resource, and failure to gain a resource. The distressful even seems to be along a continuum where the least distressing is annoyance, the second level is stressful, and the most distressing is traumatic.

21
The Importance of Accurate Classification

Using explanation systems that are based on rationally derived theories may not be useful in the pursuit of finding effective means for alleviating suffering. Understanding precisely how psychopathology is caused can lead to more precise methods for treatment. For example, how does experiencing a traumatic event lead to the development of PTSD? Acknowledging that our current understanding of this process was incomplete, researchers have re-examined the assumption that to develop PTSD there is a specific course of development, starting with the experiencing of a traumatic event (Kramer et al., 2016). Kramer et al. argued that there is evidence to suggest that the symptoms of PTSD can emerge even without exposure to a traumatic event. To test whether or not PTSD is a discreet disorder that emerges only after a traumatic event, they posited that if it is true that the emergence of PTSD symptoms can only emerge after experiencing a traumatic event that there should be a distinct class of people who exhibit PTSD symptoms and that this class emerges after a traumatic event (Kramer et al., 2016).

Using exposure to combat among military personnel, they assessed for PTSD symptoms pre and post deployment (Kramer et al., 2016). They used a number of methods (e.g., latent trait analysis, latent class analysis, and FMM using a hybrid of latent trait and latent class analysis, and confirmatory factor analysis) to test for the latent trait of the PTSD structure and the latent trait of the Depression structure (Kramer et al., 2016). With these different methods, the authors wanted to understand the data structure in terms the number of subgroups with respect to the response patterns versus the variable covariance, outcome variables as continuous or categorical, and parametric versus non-parametric methods fit the data best. The results showed that using the covariance of the variables was better than using the response patterns and using dimensional versus categorical latent variables that describe severity of symptoms fit better (Kramer et al.,
They also found that no new classes of symptoms emerged post exposure (Kramer et al., 2016). They concluded PTSD may not be distinct from general distress and that describing the symptoms as continuous variables is better than using categorical variables (Kramer et al., 2016). They did not, however, discuss the methods in terms of performance. Even with this limitation, the study illustrated the utility of empirically driven diagnostic classification and letting the data determine the story.

Additionally, precise understanding of how treatment modifies aspects, or symptoms, of psychopathology might also help to further the science of psychopathology and its treatment. For example if EMDR is effective because of the imaginal exposure, then the eye-movement portion of the therapy could be superfluous. If so, the therapist could be more direct in performing therapy by concentrating on the portions of the techniques that are actually causing symptom change. Therefore, additional studies that use methods to examine data beyond simply hypothesis generation and testing are necessary.

Data behavior reflects patterns of people’s actual experiences

By whatever label seems most appropriate, emotional distress caused by a stressful event has characteristics that can be quantified. The quantified representation of the emotional distress and aspects of what may be causing that emotional distress is numerical data. The numbers in concert represent an event and the emotional consequences of experiencing that event. Patterns of the data represent similarities and differences in the experience and emotional reactions of the event. Similarities and differences in the patterns of the numerical data represent differences in the characteristics of the event and emotional reaction. These patterns in the data can be said to have behavior. If there are groups behaving the same way, then this can be called a subgroup
The data behaves one way, as a representation of one way of reacting to the stressful event. The data behaves in another way, as a representation of a different reaction. If there is a pattern where certain groups of resources tend to be followed by more severe emotional distress, the data are behaving according to types of resources. The type of resources could be referred to as a causal mechanism. If there is a cause and effect relationship between the loss of a type of resource and the severity of the emotional reaction, then one would expect the data to behave in a way where this relationship is evident. More loss of a type of resource should equal more severe emotional reaction. Using item-level data, the numerical representation tends to be more nuanced (Watson et al., 2008). Dimensional variables also tend to be more nuanced (Krueger et al., 2014a). Therefore, finding patterns in the numerical data requires the use of numerical analytic techniques.

Why look for subgroups?

If letting the data reveal itself is the goal, unsupervised statistical learning techniques can find patterns in the data by looking for how the data tend to group together and how those groups tend to be more separate from each other (Hastie et al., 2009; James et al., 2013). For example, the behavior of the data may suggest that there are two types of reactions: one where people get moderately emotionally distressed, and where another group of people tend to get severely emotionally distressed. Finding groups of data behavioral patterns is called extracting subgroups (Duda et al., 2012). Extracting subgroups can be used to identify a group of people who may need more care than other groups of people.

For example for those who develop PTSD, there are subgroups who tend to do worse in terms of symptom relief and length of time to experience that symptom compared to others who receive the same treatment (Wolf et al., 2012). An example of a subgroup is when some people
experiences a symptom called dissociation at the time of the traumatic event (Wolf et al., 2012). When people experience dissociation, they tend to have worse symptoms that are more resistant to current therapeutic methods (Wolf et al., 2012). Not only are these people suffering more, but also it is more difficult given current treatment options to help them reduce their symptoms.

Being able to identify this group means that research can be done on not only finding out why some people experience dissociation for the purpose of early identification and ideally prevention, but also could provide the opportunity to develop treatments that would work better for them. If existing data had not been examined for subgroups within the PTSD population, these dissociative groups would not have been identified, which would have slowed progress for investigating better treatments for this group. Within unsupervised statistical learning methods, there are two techniques that have different perspectives for exploring grouping patterns in the behavior of the data: k-means cluster analysis and finite mixture modeling (FMM).

Unsupervised Statistical Learning Techniques

In unsupervised statistical learning, there are a number of techniques that can be used (Duda et al., 2012; Hastie et al., 2009; James et al., 2013). Among these techniques, when the goal is to identify similarities within a data set, there are two basic ways to accomplish this, either by analyzing the similarities between variables, as in factor analysis, or identifying similarities among response patterns of individual participants, such as k-means cluster analysis and FMM (Duda et al., 2012; Hastie et al., 2009). To illustrate the contrast, grouping data by variable is found in factor analysis (Everitt & Hothorn, 2011). If a researcher is trying to determine whether there are commonalities between the variables (e.g., finding distinct, subcategories within a broad diagnostic category), then factor analysis could be a good method to use (Everitt & Hothorn, 2011). If, however, the focus of interest is comparing and contrasting
people’s experiences (and in particular, identifying subgroups within a diagnostic category that may have worse symptoms compared to other subgroups), then using techniques that focus on the groupings of individual patterns would be more appropriate (Everitt & Hothorn, 2011). This type of analysis would also work for identifying groups of outliers (Aggarwal, 2013).

I am interested in the groupings of individual patterns, and therefore, the two types of unsupervised learning techniques that can be used to explore these groupings are cluster analysis and finite mixture modeling (FMM; Hastie et al., 2009; James et al., 2013). There are many ways to conduct a cluster analysis. One of the most common is k-means cluster analysis (Jain, 2010). K-means cluster analysis has the added bonus of working in an analogous way to FMM (Hastie et al., 2009; James et al., 2013). They both can uncover subgroups (Hastie et al., 2009; James et al., 2013), meaning Jane’s pattern of symptoms fits within subgroup 1, whereas Jill’s pattern of symptoms fits within subgroup 2. Under certain circumstances, they are virtually the same (Hastie et al., 2009; James et al., 2013). Because of this similarity, comparisons between how k-means cluster analysis and FMM determine subgroups can lead to some insights into the nature of subgrouping.

While both k-means cluster analysis and FMM can both uncover subgroups, there are two main ways that they are different: 1) how they determine sameness and differentness, and 2) how they assign members to the subgroups. Therefore, to reiterate, the goal of this study is to determine if there are subgroups of people that vary according to their patterns of symptoms and experiences. Once subgroups are found, other analyses can be conducted to determine why they might be different and what that difference means. With the dissociative subgroup (mentioned above), an initial question is if such a subgroup exists. Then, the why’s, what’s, and how’s can be determined. In particular, unsupervised statistical learning can help determine if further
research is warranted and suggest what sorts of questions could be asked. To better understand the two techniques used in the current study, it is helpful to visualize the output from each one. 

*Visualizing k-means cluster analysis*

Sameness and differentness is represented by distance (Hastie et al., 2009; James et al., 2013). If we were representing two variables on a piece of paper, each person’s quantified representation of their symptoms would be a dot on a Cartesian graph. The x-axis is one symptom, and the y-axis is the other symptom. If the dots are close together, they are more similar to those dots than the dots that are farther away on the paper. If there are two clumps of dots, then there are two clusters. If the two clumps are close together, then the clusters are thought to be more similar to each other than if the clumps were farther apart. This is important for a couple of reasons: 1) some analytical methods will treat these two clusters as one cluster, in other words they will be seen as one cluster, and 2) the differences in clusters may not be meaningful. The meaning piece would have to be analyzed separately from finding the clusters.

Where the dots are in terms of placement on the page will give the magnitude of both symptoms. If the dot is at (1,3), then the symptom represented by the x-axis has a severity score of 1, and the symptom represented by the y-axis has a severity score of 3. One dot. Two symptoms. Each symptom is a dimension. If you wanted to visualize three symptoms, then the clusters would look like galaxies in space. You are now looking at three symptoms in three-dimensional space. If you wanted to look at four or more variables, then one has to use numbers to describe the clusters, rather than pictures to understand the data behavior, because visualizing more than three dimensions is next to impossible for most people. Looking at the graphic representation of the data, one can get a sense of the story that the data are telling, such as if there are subgroups within the data and how much the participants are suffering.
By contrast, sameness and differentness is represented by density distributions in FMM. What is a density distribution? The most familiar one is the normal distribution. The x-axis represents one variable. The y-axis represents how many people have the same x-value. For example if 10 people had a symptom score of five, then at the x-value of five the curve would be at 10. The highest point of distribution represents the mode for the distribution, which is the value that most often occurred in the sample or population. If you have two variables, then the distribution curve becomes three-dimensional. There are two dimensions for the variables (i.e., the symptoms) and one dimension for the frequency. If you have subgroups, then you will see humps, like camel humps, for each subgroup. Subgroups in FMM are called components (McLachlan & Peel, 2000). The mixing components are thought to have different generators (McLachlan & Peel, 2000), which is to say that the causal mechanism for the different subgroups are different. In the dissociative subtype of PTSD, the causal mechanism that made the person experiencing that symptom at the time of the trauma event has something different about it that would: 1) make the person experience the symptoms, and 2) cause the other symptoms to manifest differently from those who did not experience the symptom.

For FMM, sameness and differentness is a function of how close the observed value, the symptom score, is to the mean. So if the symptom score happens to be the same value as the mean of component 1, then that person’s symptom pattern belongs to component 1 (McLachlan & Peel, 2000). If the value of the symptom severity is in between the mean for two components, then that person’s symptom pattern could be in either component.

This brings up one difference between k-means and FMM. Membership in a k-means cluster is absolute. It either is in a cluster, or it is not. By contrast, FMM can have one person’s
patterns belonging to more than one subgroup. In FMM the probability of membership in one or the other component is calculated, as a function of how close the score is to each of the component means. If the score is closer to the component 1 mean, then it will have a higher probability of belonging to component 1 versus belonging to component 2. If the probabilities are close to 50%/50%, then means are relatively close together. If none of the scores is less than 100%/0% probability, then the two distribution means are very far apart and the tails of the distributions do not overlap.

Between the two methods, k-means cluster analysis has the advantage of being easier to visualize and to interpret. K-means is generally called a distance model, because the data are evaluated with distance between data points. FMM is generally called a probabilistic model, because the model is evaluated by probabilities.

*Characteristics of the models: Things to know for interpretation*

In addition to visualizing the output for k-means cluster analyses and FMM, it is also helpful to consider some characteristics of these models, which can provide us with more of the story and help to better organize what the data can tell us. These characteristics include: 1) the size of the subgroups, 2) the importance of membership assignment, 3) multivariate data and the curse of dimensionality, 4) the number of groups, and 5) the importance of initial values and the algorithm.

**Size.** To get a sense of the behavior of the data, which of course reflects the experiences of those people who were surveyed, one characteristic is the size of the subgroups. The size of the subgroups can vary. The data may contain two groups with equal size (Hastie et al., 2009; Pham, Dimov, & Nguyen, 2005; Tibshirani, Walther, & Hastie, 2001). The data may contain one big group and one small group. The data may contain three groups. Group size is reflected
in the results of the analysis. For k-means cluster in R, the results will actually give you a number for each cluster (Hastie et al., 2009; Maechler et al., 2015). In FMM, the results will be given in terms of a percentage (Benaglia, Chauveau, Hunter, & Young, 2009; Hastie et al., 2009). This percentage is called the mixing proportions in FMM (Hastie et al., 2009; McLachlan & Peel, 2000).

**Membership assignment.** The difference in how participants’ responses are assigned to a specific cluster has ramifications on the final subgroup solution, meaning that given the same data set the group size and members of the subgroup could be different between the k-means and FMM methods (B. S. Everitt, S. Landau, M. Leese, & D. Stahl, 2011; Hastie et al., 2009; James et al., 2013; McLachlan & Peel, 2000). This means that the story the data tell could be different solely because of the method used and not necessarily reflective of the information contained in the data.

For k-means, a participant’s response can be in one cluster only (Duda et al., 2012; Hastie et al., 2009; James et al., 2013). In fact every response must belong to at least one cluster (Duda et al., 2012; Hastie et al., 2009; James et al., 2013). This is called hard assignment. By contrast, FMM uses a soft assignment for its component members. It is called soft, because one participant’s response can be in more than one component.

For interpretation, the hard versus soft assignment can have a profound effect on the results (Hastie et al., 2009). In situations where the subgroups may overlap, creating a hard division between the subgroups can lead to wildly different results, which of course would influence the story that emerges from the data. It could make the subgroups appear uneven in membership, if the overlapping data went to one side versus the other. It also means that k-
means cluster analysis is sensitive to outliers, where the centroid of the cluster and shape of the cluster can be greatly influenced by an extreme outlier (Hastie et al., 2009).

**Multivariate data.** It is relatively easy to visualize clusters from k-means cluster analysis and components from the FMM if the number of variables is one or two. Beyond that, it is difficult to visualize. At that point, one must rely on numbers to get a sense of how the data are grouping together. Both k-means and FMM can find subgroups within multivariate data (Hastie et al., 2009; James et al., 2013). There are some limitations with each of the models when there are too many variables (Aggarwal, 2013; Hastie et al., 2009; James et al., 2013; Pham et al., 2005; Tibshirani et al., 2001). This is a situation called highly dimensional data (Aggarwal, 2013; Hastie et al., 2009; James et al., 2013). Data points within highly dimensional data tend to become more equidistant (Aggarwal, 2013), making it difficult to find patterns in the data. If the subgroups are distinct and well separated, then high dimensionality becomes less of a problem (Aggarwal, 2013; Tibshirani et al., 2001).

A solution to what is termed the curse of dimensionality is to employ techniques that reduce the number of dimensions (Aggarwal, 2013; Duda et al., 2012; Hastie et al., 2009; James et al., 2013). A common technique is to use principal component analysis (PCA), which is similar to factor analysis in that the PCA will extract variables based on correlations between the variables (in contrast with techniques that will extract variables based on the characteristics found when viewing the data as a matrix called eigenvalues) (Aggarwal, 2013). It is important to retain as much of the variance as possible, because that variance is the “information” of the data set (James et al., 2013).

**Number of groups.** The number of groups has to be preset for k-means cluster analysis as well as for FMM (i.e., when you are asking the program to estimate the FMM component
means, variance, and proportion, though there is more flexibility in the pre-settings when using
FMM) (Hastie et al., 2009; James et al., 2013). To find the number of subgroups before running
the analyses, there are k-means cluster analysis based functions in R that will automatically find
the most likely number of subgroups (Maechler et al., 2015; Pham et al., 2005; Rodriguez, 2015;
Tibshirani et al., 2001). The gap statistic was created by Tibshirani et al. (2001). It compares
the data set with a simulated data set.

The simulated data set has one cluster in it (no subgroups). The user gives the program a
range of possible clusters. The gap statistic then determines if it is more likely that the data set
has one cluster or each of the given range of clusters. For example, if the user tells the gap
statistic to look for two to 20 clusters, the program will determine if it is more likely that the data
has one or two, one or three, and so on up to 20. A result of one does not necessarily mean that
there are no subgroups (usεr11852, 2015). What it does mean is that the function was unable to
reject the null (Tibshirani et al., 2001; usεr11852, 2015).

The other method for finding the likely number of groups using k-means cluster is the k
selection function developed by Pham et al. (2005). It analyzes the data and uses a weighted
function to determine the most likely value for the number of clusters. Neither function works
perfectly. There are situations where the functions will not be able to find clusters, even if there
are clusters in the data, including if the clusters are very close together or if there is a big cluster
and a very small cluster.

**Initial values and algorithm.** The algorithms, the iterative steps for calculating
solutions, for the k-means and FMM are similar. They both start with initial values that are
generated by random number generators. For k-means, the initial values are for the clusters’
center, called the centroid. Those centroids are called local minimizers. It is called a minimizer,
because the within sum of squares as a measure of distance is minimized (B. Everitt, S. Landau, M. Leese, & D. Stahl, 2011). The subsequent calculations are made around the centroid, analogous to finding the variance around a mean (Hastie et al., 2009; James et al., 2013). It is also an important concept when interpreting the results. Eventually, the k-means algorithm will find centroids around which have the smallest within sum of squares distances between data points. This is the final solution of the analysis. The mean for the whole data set is called the global maximize, and it is analogous to the grand mean in ANOVA (Hastie et al., 2009; James et al., 2013). Using the global maximizer helps in the interpretation of the results and evaluation of the results (Desgraupes, 2013; Gurrutxaga, Muguerza, Arbelaitz, Pérez, & Martín, 2011). For the full explanation of the algorithm, please see Appendix A.

For FMM, the current method uses an algorithm called the expectation-maximization algorithm. Originally, FMM was calculated with a cumbersome equation (see Appendix B; McLachlan & Peel, 2000). With current computing power, the ability to use an iterative approach has become possible. The initial values for the parameters being estimated, such as the component means, component variance, the mixing proportions, and the number of components, are generated randomly (McLachlan & Peel, 2000). At least one parameter must be given, for use in unsupervised statistical learning; the number of components must be given a priori. This step is called the expectation step. The next step is called the maximization step, where the evaluation criterion is calculated. Then, the membership is reassigned based on the criterion value, and the parameters are recalculated, and so on. For the full explanation of the algorithm, please see Appendix B.
Research Goals

Rationally derived theories on the origin and treatment of psychopathological states seem to be problematic. The problem can occur when theories are not based on carefully and rigorously tested assumptions. Additionally, the current taxonomic system is informed by science, but not created by science. It is possible that this system is inaccurate. Using dichotomous variables in the diagnostic criteria or in research studies, such as the presence of a diagnosis studied as present or not present, may lead to the loss of information.

I have suggested that using item-level data can provide a solution to this problem. I have also suggested that if we use only hypothesis testing we may be losing the opportunity to find out what the data could be telling us. We can begin to fill this gap by using unsupervised statistical learning techniques. In fact, others in the field are beginning to use these techniques, with important findings that are improving our understanding of and treatment for specific types of disorders (Kramer et al., 2016; Krueger et al., 2012).

Therefore, the aim of the current study is to continue this conversation by demonstrating, as an example, how to use unsupervised statistical learning on a data set, given frequently used methods (i.e., K means cluster analysis and FMM). These methods are exploratory in nature. They also emphasize response patterns, rather than covariation among variables. The data set used in the current study is comprised of responses from university students in Louisiana and Mississippi about their experiences after Hurricane Katrina, with a specific focus on the relationship between depression and loss of resources. If it is true that loss of resources is related to distress, then there should be a discernable relationship between the two. Using the same data set, Liang, Hayashi, Bennett, Johnson, and Aten (2015) found a significant relationship between finance loss and depressive symptoms. It is possible then that those who suffered financial
losses could be classified as a subgroup of the data set created for this study (see Chapter 2. Methods for how the current study’s data set was extracted from the larger data set used in the Liang et al. paper). Interestingly, other resource losses were not found to be significantly related to depressive symptoms (Liang et al., 2015).

Therefore, using K-means cluster analysis and FMM, the research goal for the current study is to determine if there are subgroups related to depression and resource loss in the extracted Hurricane Katrina data set. Finding these subgroups will be done using the following steps:

1) Determine how to handle missing data. Since the focus of this study is on the methods and not the content of the measures, the criteria for handling missing data is to provide a clean data set without introducing confounding variables.

2) Determine if preprocessing the survey data should be performed using Principle Component Analysis (PCA). The literature suggests using the components that have eigenvalues above 1 and retaining as much variance as possible (James et al., 2013).

3) Determine how to handle initial values of the k-means and FMM parameters. It is the starting point of the algorithm for the cluster center in k-means and the mean, variance, and mixing proportion for FMM (i.e., the percentage of cases assigned to each subgroup).

4) Determine benefits and costs of dimension reduction.

5) Find the value or possible values for k, where k is the number of subgroups.

6) Determine how to find outliers.

7) Determine strengths and weaknesses of each model for analyzing subgroups, given this data set.
8) Given this demonstration, evaluate the usefulness of unsupervised statistical learning in the acquisition of knowledge in psychopathology.
Chapter 3: Methods

Description of Data Set

Parent data set

The raw data set for this project was extracted from an IRB approved study. The original data set from which this project’s data set was extracted will be referred to as the parent study. The parent study was comprised of university students who were forced to leave their home after Hurricane Katrina, but who were able to remain enrolled in their current universities. Participants from this parent study included 654 college students at two universities affected by Hurricane Katrina, the University of New Orleans (UNO) and the University of Southern Mississippi (USM).

Raw data set

The raw data set was created from the parent data set, and included demographic variables and the measurement variables. From this data set, the project data set was created. The University of Hawai‘i Institutional Review board granted approval to analyze the extracted data for this project.

Item-level data set

A data set of complete cases was extracted from the raw dataset. This will be referred to as the project data set. Missing data were addressed by listwise deletion. (Please see the sections on missing data and data handling for further explanation). The project data set had a total of

---

1 As part of being an IRB approved study, each participant was informed that the participation was voluntary and they had the right to withdraw from the research.
2 On Monday, August 29, 2005, hurricane Katrina made landfall as a Category 3 hurricane (3 out of 5 categories; Beven et al., 2008). It was one of the deadliest hurricanes in United States history, killing indirectly and directly 1833 people and costing approximately $81 billion in damages (Beven et al., 2008).
539 participants, 408 women, 124 men, and 7 people who did not report their gender (see Table 1 for a breakdown of participants’ year in school and gender). There was only one graduate student (gender will not be given to maintain confidentiality). Two people did not give information on their year in school. Table 2 shows participants’ reported ethnicity.

Table 1: Participant Frequency by Year and Gender

<table>
<thead>
<tr>
<th>Year</th>
<th>Women</th>
<th>Men</th>
<th>Did not answer</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>62</td>
<td>14</td>
<td>7</td>
<td>77*</td>
</tr>
<tr>
<td>2</td>
<td>69</td>
<td>29</td>
<td></td>
<td>98</td>
</tr>
<tr>
<td>3</td>
<td>132</td>
<td>37</td>
<td></td>
<td>174*</td>
</tr>
<tr>
<td>4</td>
<td>143</td>
<td>43</td>
<td></td>
<td>187*</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

*Note:* Totals may be different from gender split total due to missing data.

**2 did not answer the year in school question.

Table 2: Participant Ethnicity Frequency

<table>
<thead>
<tr>
<th>Ethnicity Category</th>
<th>Total</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>African-American</td>
<td>184</td>
<td>35.1%</td>
</tr>
<tr>
<td>African</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Asian-American</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Asian</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Caucasian</td>
<td>338</td>
<td>62.7%</td>
</tr>
<tr>
<td>Native American or Alaskan</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Hispanic</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Pacific Islander</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Other</td>
<td>*13</td>
<td>0.02%</td>
</tr>
<tr>
<td>Missing</td>
<td>4</td>
<td>0.007%</td>
</tr>
</tbody>
</table>

*Note:* Due to low frequency, included in “other” category. No one actually answered “other.”

In terms of religious affiliation, 488 participants reported being Baptist, Roman Catholic, Protestant, Mormon, or other type of Christian, with the majority (n = 275) being Baptist. There were 20 participants who reported being atheist or agnostic with most reported being agnostic.

There were 25 participants who did not answer the question. The remaining 6 participants were
Buddhist, Muslim, Jewish, or Unitarian Universalist. Regarding age, the median age was 21, the mean was 22.27, and the range was 18 to 55 years of age. There were 11 people who did not answer the question about age.

**Measures**

In addition to demographic variables, there are two measures that comprise the project data set. See Appendix C for the items used in each measure. The psychometric properties of each measure will be discussed.

*Center for Epidemiologic Studies Depressive Scale (CES-D)*

This measure was developed to detect depressive symptoms in epidemiological studies where a clinical diagnosis may not be present (Breslau, 1985; Radloff, 1977). It was developed during the time that the DSM-III was the prevailing diagnostic standard (Carleton et al., 2013). The items were not written to be reflective of the DSM-III diagnostic criteria (Carleton et al., 2013; Radloff, 1977). It was intended to reflect a four-factor structure: depressed affect, somatic activity (or lack thereof), and interpersonal challenges (Carleton et al., 2013; Radloff, 1977). The instrument asks for symptoms from the past week, which has been labeled as current symptoms (Radloff, 1977). It is a 20-item self-report measure with a Likert-type scale with four levels measuring frequency of days in the past week: rarely or none (0-1 days), some or a little, occasionally or a moderate amount of time, and most or all of the time (Radloff, 1977). The non-depressive items are reverse scored (Radloff, 1977). A score of 15 or below is often considered normal or not pathological (Breslau, 1985; Carleton et al., 2013; Comstock & Helsing, 1977). However, there is some debate on the accuracy of this cut-off point (Carleton et al., 2013).
With respect to reliability, Cronbach’s alpha is considered an estimate that all of the items are measuring the same construct (Allen & Yen, 2001). One study reported the instrument had a Cronbach’s alpha of .85 for the general population and .90 for the psychiatric population (Radloff, 1977). Another study reported a Cronbach’s alpha of .91 for a community sample, .85 for a clinical sample, .83 for the NHANES³ sample (Carleton et al., 2013; Miller, 1973). Using the listwise deleted data set in the R package psych (Revelle, 2014), the standard Cronbach’s alpha based on correlations was .93 for all items. It is important to note that depressive symptoms cannot be assumed to be stable over time, so stability measures might be measuring the construct’s stability, rather than the stability of the measure itself.

With respect to sensitivity and specificity, one study suggests that the measure is sensitive to detecting symptoms in common with the emotional distress, but not necessarily just depressive symptoms. In comparison with those who have current MDD, there is sensitivity of 87.5% and specificity of 73% (Breslau, 1985). For those who meet DSM-III criteria for current MDD, the predictive value of the CES-D measure is that 15% of those had a score above 16 (Breslau, 1985). For those who met criteria for generalized anxiety disorder as defined by the DSM-III, the sensitivity was 80%, and the specificity was 73% (Breslau, 1985). The predictive value of this assessment in predicting GAD is 17% (Breslau, 1985). This assessment performs best in identifying those with both MDD and GAD (Breslau, 1985). Thus, this evidence suggests that the measure detects a more loosely defined construct of emotional distress, rather than a specific diagnosis.

³ National Health and Nutrition Examination Survey. The data was collected from 1971-1975. Participants completed the CES-D as part of the nationwide study of the National Center for Health Statistics (Miller, 1973).
With respect to validity, there has been evidence that the intended four-factor structure may not be valid. Numerous studies have investigated the four-factor solution, but found solutions ranging from one to four factors (Carleton et al., 2013). Most of these studies used a PCA with an orthogonal rotation (Carleton et al., 2013). The validity of several items have been called into question (Carleton et al., 2013). For example, the somatic items may be elevated with elderly or chronic pain patients (Carleton et al., 2013). The social impairment items may confound with social anxiety and self-perceived social impairment (Carleton et al., 2013). Culturally based social norms of emotional expression may confound the “crying spells” item (Carleton et al., 2013). In terms of validity, the measure does well with screening for emotional distress with some cautions like the somatic items, but should not be used as a diagnostic tool.

**COR: Conservation of Resources**

The Conservation of Resources Evaluation (COR-Evaluation) (Hobfoll & Lilly, 1993) is a 60-item survey that evaluates the gain and loss of resources. The responses are the degree of loss or gain from -1 (little loss) to -4 (very great loss) from 1 (little gain) to 4 (very great gain) within the past month. If there was no loss or gain, the participants report “0”. Total score of the scale reflects the whole resource loss and resource gain. The survey has been used in numerous studies, including a study on the emotional effects of natural disasters and workplace burnout (Benight et al., 1999; Freedy, Shaw, Jarrell, & Masters, 1992; Halbesleben, 2006).

The COR survey has been found to have good reliability. In one study, internal reliability (no specific method stated) was reported as .94 with a test-retest was .67 over nine months (Benight et al., 1999). However, in the Benight et al. (1999) study, only 42 items were used. In another study, using the raw data in the R package *psych*, the standard Cronbach’s alpha based on the correlations was .93 for all items (Revelle, 2014).
In terms of the constructs that COR is measuring, Liang et al. (2015) conducted a principal components factor analysis and extracted four components, which corresponded with the following categories of resource loss: 1) social relationships, 2) household items, 3) finance (i.e., money), and 4) time. Although these studies show some promise for the reliability and validity of the COR survey, in general, there has been very little reporting of the psychometric properties of the measure.

**Data Handling**

Data handling was performed mostly in the R environment with the exception of the three data sets created (see below in the Data Reduction section for more information on the creation of these data sets) by using a principal component analysis (PCA) for the purpose of dimension reduction (a.k.a., data reduction). This was done using SPSS 22.0. A data set object was created in the R environment that included the following variables: the participant identification number (which was a sequential number), the items from the depression measure (CES-D), and the items from the conservation of resource measure (COR). Another data set was extracted to only include complete data, using the na.omit function in R on the measurement data only. Using only the complete cases of the measurement data, this left 539 participant data from 634. A data object was created with just these participants that included demographic information to calculate summary statistics. This was done by merging the demographic variables with the listwise deletion data. Since the demographic data were used only for data descriptions and were

---

4 Data reduction was also done in the R environment, but the results were different. R uses singular value decomposition (Urbanek, Bibko, & Iacus, 2014). SPSS performs PCAs as a factor extraction method (IBM, Released 2013.).
not part of the analysis, no data imputation was performed on the missing demographic data (see below).

**Missing Data**

Prior to extracting the project data, using SPSS 22.0, the raw data were analyzed for missing data patterns. Since there were no variables with more than 5% missing data, the mismatch table was not produced (IBM, Released 2013). A mismatch table is derived from examining each pair of variables (IBM, Released 2013). For every pair, there could be a value missing in one variable and not in the other (IBM, Released 2013). Five percent is the default minimum for producing a mismatch table (IBM, Released 2013). It is a descriptive indicator of missingness (IBM, Released 2013).

The Little’s MCAR test (i.e., missing completely at random) determines if the missing data is missing completely at random (IBM, Released 2013). The null hypothesis is that it is missing completely at random (IBM, Released 2013). This test does not evaluate for missing at random (MAR) or missing not at random (MNAR). The result of this test was $\chi^2 = 8027.215$, DF = 5824, Sig. = .000, which indicates that the missing data was not MCAR.

The overall summary of missing data is broken down into three perspectives: the variables, the cases, and the values (see Figure 1). The left-most pie chart shows that there is at least one missing data point in each of the variables. The middle pie chart shows that there are 539 complete cases, or the number of cases that have no missing data. Nearly all of the values are represented in the dataset.
The missing data pattern showed that there was one case where someone only answered the demographic information, but did not answer the CES-D and COR measures. There were a few cases where half of one measure was complete, but the other half was not. Otherwise, the missingness is relatively scattered (see Figure 2). The vast majority of the data (see Figure 3) showed no missing data. According to Enders (2010), missingness of less than 5% of the data is negligible.

With the missingness being less than 5% and the data most likely not MNAR, using listwise deletion is relatively safe (Enders, 2010). Additionally, since the aim of the project is not to draw conclusions about the content of the data, but about the methods of analysis, using listwise deletion seems appropriate. The main concern for the data set is that there is one and only one. If I were to impute the data before running the analyses, the process of imputing the data could introduce slight variations in the values that replaced the missing data, thus producing several data sets, which would then have to be pooled together before running the analysis. Therefore, to reduce the chances that variance would be introduced into the data set by the process of the imputing the data, I took steps to ensure that the data set was stable before the
analysis. With just the measurement variables, listwise deletion was performed. Later, the
demographic data was added to the measurement variables.

Figure 2: Missing Value Patterns
Data Analysis

To review, the research goal is to determine if there are subgroups in the data related to depression and resource loss, which includes the following steps: 1) determine how to handle the missing data (see Missing Data section above), 2) determine if dimension reduction is necessary while maximizing variance; 3) determine how to handle initial values; 4) determine the benefits and costs of data reduction; 5) determine the number for k (i.e., number of clusters and components), 6) find outliers, and 7) demonstrate the use of k-means cluster analysis and FMM. Most of the analysis was done in the R environment (Urbanek et al., 2014). SPSS 22.0 was used for dimension reduction using the factor extraction method principal component analysis (PCA) and the missing data analysis.
Preprocessing the Data with Data Reduction: Principle Component Analysis

When determining how many dimensions to keep, the literature suggests that the components should have eigenvalues above 1 (B. Everitt et al., 2011). The components should represent a large preponderance of the explained variance, but there are no hard and fast rules on this (James et al., 2013). The project data set\(^5\) was used to reduce the dimensions. The PCA was performed with three conditions. The first condition was running the PCA using no rotation, inspecting the scree plot, and then choosing the number of components to keep. Then, the PCA was run again with no rotation, keeping the chosen number of components. The component coefficients were then read into R for further analysis and were saved as a data object. The second condition was running the PCA with a varimax rotation, inspecting the screen plot, and choosing the number of components from the elbow of the scree plot that would maximize the variance. The component coefficients were read into R and saved as a data object. The third condition was choosing all of the components with eigenvalues above 1. The component coefficients were read into R and saved as a data object. The benefits of dimension reduction will be evaluated on the basis of stabilizing the results and adding to the quality of identifying subgroups. The cost of the data reduction will be evaluated in terms of information retained or lost.

Initial Values for the Parameters Being Estimated by the Algorithms of K-means and FMM

As part of the algorithm for both k-means cluster analysis and FMM, the initial values for the parameters, such as the center of the cluster in k-means or the mean for the subgroup distribution in FMM, are generated by random number generators. Both k-means cluster

---

\(^5\) The data set of complete cases for the measurement data.
analysis and FMM are sensitive to initial values (Jain, 2010). To make comparisons between the four data conditions easier (i.e., the project data set and three data reduction conditions), the initial values were generated by setting the seed using the R function `set.seed(value)` with the values of 0, 50, 100, 200, and 300. By setting the seed, the random number generator is constrained. This has the benefit that each time the algorithm is run given a particular value for the seed, the same results will occur. For example without setting the seed, the initial values for the parameters will be randomly generated, making it so that the final parameters, such as the subgroup means, may be different. Each time the algorithm is run with a `set.seed(50)`, the mean will be exactly the same as each other time the `set.seed(50)`. Setting the seed simulates the random number generator in that there are different starting values, but does so in a controlled manner in order to make meaningful comparisons of the results.

**Finding K: The Number of Subgroups**

For both the k-means cluster analysis and FMM, the number of subgroups has to be set before running the algorithm. One way to determine the number of clusters for k-means cluster analysis is to use the gap statistic and kselection functions in R (Pham et al., 2005; Tibshirani et al., 2001). Using the R function `mvnormalmixEM`, the algorithm cannot automatically compare a range of possible subgroups (Benaglia et al., 2009).

**K-means and the functions: the gap statistic and kselection.** The gap statistic and the kselection package were applied to each of the four data conditions for each of the seed values, which would give the value for k for the k-means cluster analysis.

**Finite mixture modeling: Using log likelihood as quality measure.** For FMM, comparing the goodness-of-fit for proposed models, such as $k = 2$ versus $k = 3$, the information criterion AIC and BIC are often used (James et al., 2013). For both AIC and BIC, the log
likelihood is used and adjusted by penalty functions (Akaike, 1973; Schwarz, 1978). Since the goal is not to find the best solution, but only a good possibility, the penalty function is unnecessary. Therefore just using the log likelihood to evaluate the models, where $k = 2+$ is sufficient.
Chapter 4. Results

Data Reduction: Benefits and Costs

As described in the data analysis section, using PCA with no rotation, varimax rotation, and keeping all components with an eigenvalue above 1 with no rotation resulted in project data that were reduced to four dimensions (no rotation), nine dimensions (varimax rotation), and 19 dimensions (eigenvalue above 1). The PCA four-component solution with no rotation explained 35% of the variance. The PCA nine-component solution with varimax rotation explained 48% of the variance. The PCA 19-component solution with no rotation explained 64% of the variance, where all of the components had an eigenvalue of at least 1.

According to the literature, there is no hard and fast rule about what percentage of variance should be explained by the model, but there is some agreement that good models should account for at least 80% of the variance (James et al., 2013). Less than 80%, some authors state that the reduced dimensions have lost too much information to be considered reflective of the original data set (James et al., 2013). None of these conditions met the criteria for a good solution for data reduction. The most adequate dimension solution was the 19-component solution, where 64% of the variance was explained by the components, but it is unclear if this is an ideal solution.

The literature also suggests that highly-dimensional data is when the number of variables is much greater than the number of cases or $p >> n$ (B. Everitt et al., 2011). The project data set is 80 by 539. Therefore, the best solution of the three is the PCA 19-component solution, because it has the most explained variance of the three. The scree plots for each condition are presented in Appendix D.
Benefit: Solution stability

Using the four dimension reduction data sets, FMM and k-means cluster analysis were performed at each seed value of 0, 50, 100, 200, 300. One of the evaluation criteria for this project was to ascertain if the membership of the subgroups remained the same between FMM and k-means cluster analysis. Thus, it can be considered a rough estimate of membership composition.

For the item-level data, the explained variance is, of course, 100%. For the FMM, there was some shift in membership between trial 1 and trial 2 as shown with the proportion of membership for \( k = 2 \) 55.1/44.9% (trial 1) and 48.1/52.9% (trial 2). The FMM trial 3 was quite different from the first two trials with 75.8/21.5%. For trials 4 and 5, the solution did not converge. For the k-means cluster analysis, the first two trials were the same solutions with 68.3/31.7% for both. For trial 3, there was a flip in membership and it was slightly different than the first two with 37.1/62.9%. The last trial was 26.9/73.1%. The results are summarized in Table 3.

<table>
<thead>
<tr>
<th>Trial</th>
<th>FMM K1</th>
<th>FMM K2</th>
<th>K-means K1</th>
<th>K-means K2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (0)</td>
<td>55.1%</td>
<td>44.9%</td>
<td>68.3%</td>
<td>31.7%</td>
</tr>
<tr>
<td>2(50)</td>
<td>48.1%</td>
<td>52.9%</td>
<td>68.3%</td>
<td>31.7%</td>
</tr>
<tr>
<td>3(100)</td>
<td>78.5%</td>
<td>21.5%</td>
<td>37.1%</td>
<td>62.9%</td>
</tr>
<tr>
<td>4(200)</td>
<td>*</td>
<td>*</td>
<td>64.4%</td>
<td>35.6%</td>
</tr>
<tr>
<td>5(300)</td>
<td>*</td>
<td>*</td>
<td>26.9%</td>
<td>73.1%</td>
</tr>
</tbody>
</table>

*did not converge

For the PCA four-component solution data, the FMM showed more stable membership assignments. For the k-means solution, there was some membership assignment shift. For the FMM, the proportion of membership for all of the trials was 43.8/56.2% or flipped, which indicates the same membership composition for each component but with the assignment
flipping. The first trial of the k-means was 70.1%/29.9%. For trial 2, the proportion of membership was 26.7/73.3%. For trial 3 and 4, the proportion of membership was 72.9/27.1% with no flipping. The fifth trial had a proportion of membership of 31.4/68.6%. The results are summarized in Table 4.

Table 4: Proportion of Membership for K Group by Set Seed Value for PCA 4-Component Solution

<table>
<thead>
<tr>
<th>Trial</th>
<th>FMM K1</th>
<th>FMM K2</th>
<th>K-means K1</th>
<th>K-means K2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (0)</td>
<td>43.8%</td>
<td>56.2%</td>
<td>70.1%</td>
<td>29.9%</td>
</tr>
<tr>
<td>2(50)</td>
<td>43.8%</td>
<td>56.2%</td>
<td>26.7%</td>
<td>73.3%</td>
</tr>
<tr>
<td>3(100)</td>
<td>56.2%</td>
<td>43.8%</td>
<td>72.9%</td>
<td>27.1%</td>
</tr>
<tr>
<td>4(200)</td>
<td>43.8%</td>
<td>56.2%</td>
<td>72.9%</td>
<td>27.1%</td>
</tr>
<tr>
<td>5(300)</td>
<td>56.2%</td>
<td>43.8%</td>
<td>31.4%</td>
<td>68.6%</td>
</tr>
</tbody>
</table>

With the PCA 9-component with varimax rotation, the proportion of membership in the FMM for each trial was the same, just flipped with 43.7/56.2%. The k-means proportion of membership for trials 1 and 2 were 67.2/32.8%, but flipped. The proportion of membership for Trials 3 and 4 was 71.8%/28.2% with no flipping. Finally for trial 5, the proportion of membership for the k-means was 26.9/73.0%. The results are summarized in Table 5.

Table 5: Proportion of Membership for K Group by Set Seed Value for PCA 9-Component Solution

<table>
<thead>
<tr>
<th>Trial</th>
<th>FMM K1</th>
<th>FMM K2</th>
<th>K-means K1</th>
<th>K-means K2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (0)</td>
<td>43.7%</td>
<td>56.2%</td>
<td>67.2%</td>
<td>32.8%</td>
</tr>
<tr>
<td>2(50)</td>
<td>43.7%</td>
<td>56.2%</td>
<td>37.3%</td>
<td>62.7%</td>
</tr>
<tr>
<td>3(100)</td>
<td>56.2%</td>
<td>43.7%</td>
<td>71.8%</td>
<td>28.2%</td>
</tr>
<tr>
<td>4(200)</td>
<td>43.7%</td>
<td>56.2%</td>
<td>71.8%</td>
<td>28.2%</td>
</tr>
<tr>
<td>5(300)</td>
<td>56.2%</td>
<td>43.7%</td>
<td>26.9%</td>
<td>73.0%</td>
</tr>
</tbody>
</table>

With the PCA 19-component with no rotation solution, the proportion of membership was 56/44%, but flipped. The k-means proportion of membership was 31.1/69% for the first trial. For the second trial, the proportion of membership was 64/36.1%. For the third trial, the
proportion of membership was 31.2/68.0%. The fourth trial has a proportion of membership of 71.1/28.9%. Finally, the proportion of membership for the fifth trial was 53.8/46.2%. There was very little overlap, as demonstrated in the posterior probability results. The results are summarized in Table 6.

Table 6: Proportion of Membership for K Group by Set Seed Value for PCA 19-Component Solution

<table>
<thead>
<tr>
<th>Trial</th>
<th>FMM K1</th>
<th>FMM K2</th>
<th>K-means K1</th>
<th>K-means K2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (0)</td>
<td>56%</td>
<td>44%</td>
<td>31.1%</td>
<td>69.0%</td>
</tr>
<tr>
<td>2(50)</td>
<td>44%</td>
<td>56%</td>
<td>64.0%</td>
<td>36.1%</td>
</tr>
<tr>
<td>3(100)</td>
<td>56%</td>
<td>44%</td>
<td>31.2%</td>
<td>68.0%</td>
</tr>
<tr>
<td>4(200)</td>
<td>56%</td>
<td>44%</td>
<td>71.1%</td>
<td>28.9%</td>
</tr>
<tr>
<td>5(300)</td>
<td>56%</td>
<td>44%</td>
<td>53.8%</td>
<td>46.2%</td>
</tr>
</tbody>
</table>

Dimension reduction: Benefits and costs summary

Although, the dimension reduction did increase the stability and distinctiveness of the clusters, there was a cost. The cost for reducing the dimensions was in the reduction of explained variance, which makes the data sets based on the dimension reduction suspect in terms of whether they accurately reflect the information in the item-level data.

Finding K: The Number of Subgroups

Analyzing the k-means cluster solution for K = 2: Demonstrating the lack of separation and clear boundaries within the data

The best k-means clustering solution for subgroups that are distinctive and well-separated will be found with the k-means functions of the gap statistic functions (Pham et al., 2005; Tibshirani et al., 2001). When these functions were applied to the item-level data set and the three PCA derived data sets, the result was k = 1. The dimension reduction did little to help make the clusters more identifiable. The results of k = 1 does not necessarily mean that there is
only one group with no subgroups. There are a number of situations where both functions will fail, including when there are not very distinctive clusters and when there are size differences between the clusters (Pham et al., 2005; Tibshirani et al., 2001). Since clustering algorithms will find as many clusters as is specified even if the clusters are not really present (Jain, 2010), the k-means algorithm did find two clusters, but with very little difference between the clusters, as measured by the ratio of the between sum of squares and the total sum of squares. If the value of k is set for k = 2, the ratio of the between sum of squares and total sum of squares will approach zero if the clusters found are very close together.

Using k = 2 for the k-means cluster analysis, the algorithm was applied to the four data conditions: item-level and three versions of dimension reduction. This was done to ascertain if the dimension reduction would help with creating more distinct clusters. The highest ratios were found in the PCA four-component data set, ranging from 15.7% to 17.6%. This data set only represented 35% of the variance, as previously mentioned. Using the project data with no dimension reduction had ratios ranging from 9.2% to 9.5%. The ratios found with the PCA 9-component solution were from 7.2% to 8.0%. The lowest ratios were found in the 19-component solution with an explained variance of 64%, ranging from 3.6% to 3.8% (see Tables 7-10). The Tables demonstrate that the data do not have well defined clusters. These results show that using k-means as an exploratory method for finding subgroups that may overlap will likely lead to inconclusive results. Because of this, k-means may not be the best method for finding subgroups with this data set. Therefore, the better method for finding subgroups that may overlap is FMM.
Table 7. K = 2 Sum of Squares Item-level

<table>
<thead>
<tr>
<th>Trial</th>
<th>Project Data Set</th>
<th>Within SS Cluster 1</th>
<th>Within SS Cluster 2</th>
<th>Between/Total SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (0)</td>
<td></td>
<td>30,287.67</td>
<td>19,313.99</td>
<td>9.2%</td>
</tr>
<tr>
<td>2 (50)</td>
<td></td>
<td>30,287.67</td>
<td>19,313.99</td>
<td>9.2%</td>
</tr>
<tr>
<td>3 (100)</td>
<td></td>
<td>21,417.08</td>
<td>28,022.40</td>
<td>9.5%</td>
</tr>
<tr>
<td>4 (200)</td>
<td></td>
<td>28,235.6</td>
<td>21,209.2</td>
<td>9.4%</td>
</tr>
<tr>
<td>5 (300)</td>
<td></td>
<td>17,158.62</td>
<td>32,439.02</td>
<td>9.2%</td>
</tr>
</tbody>
</table>

Table 8. K = 2 Sum of Squares PCA 4-Component

<table>
<thead>
<tr>
<th>Trial</th>
<th>PCA 4-component</th>
<th>Within SS Cluster 1</th>
<th>Within SS Cluster 2</th>
<th>Between/Total SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (0)</td>
<td></td>
<td>1039.34</td>
<td>774.0017</td>
<td>15.7%</td>
</tr>
<tr>
<td>2 (50)</td>
<td></td>
<td>720.0449</td>
<td>1071.6619</td>
<td>16.7%</td>
</tr>
<tr>
<td>3 (100)</td>
<td></td>
<td>1037.458</td>
<td>735.779</td>
<td>17.6%</td>
</tr>
<tr>
<td>4 (200)</td>
<td></td>
<td>1037.450</td>
<td>735.779</td>
<td>17.6%</td>
</tr>
<tr>
<td>5 (300)</td>
<td></td>
<td>872.6499</td>
<td>836.0217</td>
<td>16.0%</td>
</tr>
</tbody>
</table>

Table 9. K = 2 Sum of Squares 9-Component

<table>
<thead>
<tr>
<th>Trial</th>
<th>PCA 9-component with Varimax</th>
<th>Within SS Cluster 1</th>
<th>Within SS Cluster 2</th>
<th>Between/Total SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (0)</td>
<td></td>
<td>2359.034</td>
<td>2113.279</td>
<td>7.6%</td>
</tr>
<tr>
<td>2 (50)</td>
<td></td>
<td>2134.732</td>
<td>2360.90</td>
<td>7.2%</td>
</tr>
<tr>
<td>3 (100)</td>
<td></td>
<td>2640.939</td>
<td>1814.639</td>
<td>8.0%</td>
</tr>
<tr>
<td>4 (200)</td>
<td></td>
<td>2640.39</td>
<td>1814.639</td>
<td>8.0%</td>
</tr>
<tr>
<td>5 (300)</td>
<td></td>
<td>21,982.754</td>
<td>2497.772</td>
<td>7.5%</td>
</tr>
</tbody>
</table>
Table 10. K = 2 Sum of Squares 19-Component

<table>
<thead>
<tr>
<th>Trial</th>
<th>PCA 19 Component</th>
<th>Within SS Cluster 1</th>
<th>Within SS Cluster 2</th>
<th>Between/Total SS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 (0)</td>
<td></td>
<td>4993.497</td>
<td>4842.986</td>
<td>3.8 %</td>
</tr>
<tr>
<td>2 (50)</td>
<td></td>
<td>4874.303</td>
<td>4979.393</td>
<td>3.6%</td>
</tr>
<tr>
<td>3 (100)</td>
<td></td>
<td>4682.078</td>
<td>5159.707</td>
<td>3.7%</td>
</tr>
<tr>
<td>4 (200)</td>
<td></td>
<td>5069.749</td>
<td>4759.428</td>
<td>3.8%</td>
</tr>
<tr>
<td>5 (300)</td>
<td></td>
<td>4205.730</td>
<td>5622.959</td>
<td>3.8%</td>
</tr>
</tbody>
</table>

The results from the FMM give the posterior probability of the individual member, jth member, observed value \( y_j \) belongs to the \( i \)th component of the mixture component (McLachlan & Peel, 2000). This gives an indication of how much the subgroups overlap. Since FMM does not require subgroup membership to be mutually exclusive, this model is useful in situations where there are no distinctive subgroups. By contrast, k-means works well when the subgroups are distinctive.

**Finding Outliers**

Along with the exploration of whether there were distinctive subgroups, the question of whether there were outliers in the data was also considered. Outliers can be defined as an individual data point or subgroup which has values that are very different from the other groups (Aggarwal, 2013). Because the project data set is multivariate, just looking at extreme values separately is insufficient (Aggarwal, 2013). The project data set has item-level and multivariate data. This means that there could be a group of individuals whose multivariate responses are considered outliers, or where there is one individual whose response can be considered an
outlier. In the case where there is a group of individuals whose responses are outliers, using methods that find subgroups is appropriate (Aggarwal, 2013). Therefore, the results of the k-means cluster analysis and the FMM in aggregate were used to examine the question of a subgroup of outliers and to determine if individual participants as outliers might be a problem given this data set.

The presence of a group of outliers is subjective (Aggarwal, 2013). None of the subgroups found with FMM acted contrary to what would be expected. Some did act differently, but not so much so as to suggest that the members of that group were behaving as if their responses were determined by a completely different causal mechanism. To find a single participant who could be considered an outlier would require methods that are still being developed in the field, and is a known problem in the field (Aggarwal, 2013). This is because the participant’s responses are multivariate, which adds to the complication of finding outliers; thus, the process of finding outliers within multivariate data is difficult and requires methods that are still being developed (Aggarwal, 2013).

**Finite Mixture Modeling: Solutions for K = 2-5.**

After determining that there likely were no outliers in the data (though this determination is not foolproof given the limitations in the process), and given that FMM is a better technique (compared to k-means cluster analysis) for finding overlapping subgroups, additional FMM models were run where K = 2-5. FMM was run using only the item-level data, since it was already determined that dimension reduction was not necessary or even helpful. If the model converges at all, then it is reasonable to assume that this model is a possibility. If the model does not converge, then it is reasonable to assume that the model is most likely not a possibility. The models k = 2 to k = 4 converged. The model with the lowest log likelihood, and therefore the
best, is the k = 4 model (see Table 11). At this point the question becomes, what stories do the three models tell? Based on those stories, is it possible that the data contains subgroups?

Table 11: Loglikelihood Trials 1-5 for K = 2-5 Project Data Set

<table>
<thead>
<tr>
<th>Trial</th>
<th>K=2</th>
<th>K=3</th>
<th>K=4</th>
<th>K=5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (0)</td>
<td>-45906.79</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>2 (50)</td>
<td>-47369.73</td>
<td>-40998.4</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>3 (100)</td>
<td>-47504.6</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>4 (200)</td>
<td>*</td>
<td>-42746.6</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>5 (300)</td>
<td>*</td>
<td>*</td>
<td>-39772.75</td>
<td>*</td>
</tr>
</tbody>
</table>

*did not converge

Taking a closer look at the k = 2 to k = 4 solution.

**K = 2 model.** To be able to see the response patterns for each model, a bar graph was created for each of the models. The data used were the item-level measures, CES-D and COR. The first was for the FMM, k = 2 model. Except for two participants, the membership probability was 99-100% versus 0%. There were three that were 98% in mixing component 1 versus 1% in mixing component 2, which is less than 1% of the total number of participants. The components are relatively equal in size, with 47% and 53% for each mixing component. Looking at the demographic characteristics, there are no striking differences between the components. The only exception is that women were disproportionately represented, which is consistent with the entire sample (see Table 12, Figure 4, and Figure 5).
Table 12. K = 2 Demographic Means by Component

<table>
<thead>
<tr>
<th></th>
<th>Mixing Component 1</th>
<th>Mixing Component 2</th>
<th>Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>252 (47%)</td>
<td>284 (53%)</td>
<td>3 (98% probability in component 1; &lt;1%)</td>
</tr>
<tr>
<td>Mean Age</td>
<td>22.95161</td>
<td>21.68592</td>
<td></td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>198 (79%)</td>
<td>208 (73%)</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>50 (20%)</td>
<td>78 (28%)</td>
<td></td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>African-American</td>
<td>93 (37%)</td>
<td>90 (32%)</td>
<td></td>
</tr>
<tr>
<td>African</td>
<td>2 (&lt;1%)</td>
<td>1 (&lt;1%)</td>
<td></td>
</tr>
<tr>
<td>Asian-American</td>
<td></td>
<td>1 (&lt;1%)</td>
<td></td>
</tr>
<tr>
<td>Asian</td>
<td></td>
<td>1 (&lt;1%)</td>
<td></td>
</tr>
<tr>
<td>Caucasian</td>
<td>152 (60%)</td>
<td>184 (65%)</td>
<td></td>
</tr>
<tr>
<td>Native American</td>
<td></td>
<td>1 (&lt;1%)</td>
<td></td>
</tr>
<tr>
<td>or Alaskan</td>
<td></td>
<td>2 (&lt;1%)</td>
<td></td>
</tr>
<tr>
<td>Hispanic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pacific Islander</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>3 (&lt;1%)</td>
<td>3 (&lt;1%)</td>
<td></td>
</tr>
<tr>
<td>Missing</td>
<td>1 (&lt;1%)</td>
<td>3 (&lt;1%)</td>
<td></td>
</tr>
<tr>
<td>Year</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First Year</td>
<td>31 (12%)</td>
<td>46 (16%)</td>
<td></td>
</tr>
<tr>
<td>Sophomore</td>
<td>44 (17%)</td>
<td>54 (18%)</td>
<td></td>
</tr>
<tr>
<td>Junior</td>
<td>75 (30%)</td>
<td>97 (34%)</td>
<td></td>
</tr>
<tr>
<td>Senior</td>
<td>101 (40%)</td>
<td>85 (30%)</td>
<td></td>
</tr>
<tr>
<td>Graduate</td>
<td>1 (&lt;1%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missing</td>
<td></td>
<td>2 (&lt;1%)</td>
<td></td>
</tr>
</tbody>
</table>
Figure 4. K = 2 Item Means by Each Component
Variables 1 to 60 represent the COR measure. The variables 61 to 80 represent the items from the CES-D measure. For the COR items, the items below zero represent a loss of some sort of resource. For the CES-D items, those members of component 1 who had generally the most losses and least gains reported had the least severe reported symptoms of the depressive items, except for variables 64, 68, 72, and 76. The items that were unusually high for both components were: #64) I felt as good as other people, #68, I felt hopeful about the future, #72) I was happy, and #76) I enjoyed life. These items were reverse scored. This is interesting, because it would be expected that the group with the most reported losses would have the most severe depressive
symptoms. This subgroup is primarily defined as those who lost the most. The mixing component #2 had the worse symptoms of the two groups, but had the least loss and the most gain. This group is primarily characterized as those who have gained the most, but who do not have enough time or spending money. This pattern may indicate a separate causal mechanism for these four items that are not based on resource loss. Looking at this graph, one can see a general trend. The items from mixing component 1 represent a group where there was more loss. The most extreme losses endorsed were found in the subgroup mixing component 1 in the items “I have enough time to get things done” and “I have enough money for living expenses”. The general trend of loss consistent with worse depressive symptoms is what would be expected in the Hobfoll’s COR theory of stress (1988, 1989, 2001, 2004; 1993).

**K = 3 model.** The second model was k = 3. The membership probability for all components was 99-100% versus 0%. This indicates that these subgroups were distinct from each other. The demographic characteristics were relatively the same. The members of the first component generally experienced a moderate amount of loss with the least amount of gain. The responses from members of mixing components 2 and 3 had generally the same pattern of losses and gains and generally the same response patterns in the depressive symptoms. Groups 2 and 3 are characterized by more extreme responses. The losses and gains in Groups 2 and 3 were generally greater in magnitude than those of Group 1, where Group 2 had even more extreme gains than Group 3. The extreme values in the lost resources for all mixing components was found in the “I have enough time to get things done.” The items that were unexpectedly high for all three components were: #64) I felt as good as other people, #68, I felt hopeful about the future, #72) I was happy, and #76) I enjoyed life, which mirrors the patterns in the two component model (see Table 13, Figure 6, and Figure 7).
Table 13. K = 3 Demographic Means by Component

<table>
<thead>
<tr>
<th></th>
<th>Mixing Component 1</th>
<th>Mixing Component 2</th>
<th>Mixing Component 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>242 (45%)</td>
<td>174 (33%)</td>
<td>123 (23%)</td>
</tr>
<tr>
<td>Mean Age</td>
<td>22.27897</td>
<td>22.47701</td>
<td>21.97521</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>185 (76%)</td>
<td>128 (74%)</td>
<td>95 (77%)</td>
</tr>
<tr>
<td>Male</td>
<td>54 (22%)</td>
<td>43 (25%)</td>
<td>27 (20%)</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>African-American</td>
<td>79 (33%)</td>
<td>69 (40%)</td>
<td>36 (29%)</td>
</tr>
<tr>
<td>African Asian-American</td>
<td>1 (&lt;1%)</td>
<td>1 (&lt;1%)</td>
<td>1 (&lt;2%)</td>
</tr>
<tr>
<td>Asian</td>
<td>1 (&lt;1%)</td>
<td>101 (58%)</td>
<td>83 (67%)</td>
</tr>
<tr>
<td>Caucasian</td>
<td>154 (64%)</td>
<td>101 (58%)</td>
<td>83 (67%)</td>
</tr>
<tr>
<td>Native American or Alaskan</td>
<td>1 (&lt;1%)</td>
<td>1 (&lt;1%)</td>
<td>1 (&lt;1%)</td>
</tr>
<tr>
<td>Hispanic</td>
<td>1 (&lt;1%)</td>
<td>1 (&lt;1%)</td>
<td>1 (&lt;1%)</td>
</tr>
<tr>
<td>Pacific Islander</td>
<td>2 (&lt;1%)</td>
<td>3 (&lt;1%)</td>
<td>2 (&lt;1%)</td>
</tr>
<tr>
<td>Other</td>
<td>2 (&lt;1%)</td>
<td></td>
<td>1 (&lt;1%)</td>
</tr>
<tr>
<td>Missing</td>
<td>2 (&lt;1%)</td>
<td>3 (&lt;1%)</td>
<td>2 (&lt;1%)</td>
</tr>
<tr>
<td>Year</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First Year</td>
<td>34 (14%)</td>
<td>27 (16%)</td>
<td>16 (13%)</td>
</tr>
<tr>
<td>Sophomore</td>
<td>44 (18%)</td>
<td>38 (22%)</td>
<td>16 (13%)</td>
</tr>
<tr>
<td>Junior</td>
<td>71 (29%)</td>
<td>49 (28%)</td>
<td>54 (44%)</td>
</tr>
<tr>
<td>Senior</td>
<td>92 (38%)</td>
<td>59 (22%)</td>
<td>36 (29%)</td>
</tr>
<tr>
<td>Graduate</td>
<td></td>
<td>1 (&lt;1%)</td>
<td></td>
</tr>
<tr>
<td>Missing</td>
<td>1 (&lt;1%)</td>
<td>1 (&lt;1%)</td>
<td></td>
</tr>
</tbody>
</table>
Figure 6. K = 3 Item Means by Component
The third model was $k = 4$. There was one member of component 2 that was 97% versus 2%. When doing the demographics calculation, this individual’s scores were not included in any of the components. The demographic characteristics were relatively the same. Women were, again, over-represented within each cluster. There were interesting results in the COR and CES-D results (see Table 14, Figure 8, Figure 9, and Figure 10). Members of the first two mixing components endorsed the least amount of loss with the least amount of gain.
Group 1 had the least amount of depressive symptoms endorsed. Group 2 had a few resource items that indicated more loss than Group 1, and generally reported having more severe depressive symptoms than Group 1. This is what would be expected based on the theory posited by Hobfoll (1988, 1989, 2001, 2004); Hobfoll and Lilly (1993). Group 3 had more extreme responses to both gains and losses than Groups 1 and 2, but Group 4 reported having the most extreme losses and gains. Interestingly, Group 4 had the most severe depressive symptoms and responded with the most severity of feeling like they had no time and no money. It might be expected that gains would offset losses, but in Group 4 this did not seemed to be the case.

These findings suggest that the first two groups represent students who have slightly different patterns of losses and gains, but both show expected depressive symptoms, where less loss was associated with fewer severe depressive symptoms. Group 3 and Group 4 showed similar patterns of more severe losses and gains, as well as more depressive symptoms than seen with Groups 1 and 2. This may indicate a general pattern among these participants in responding in the extreme. It may be that there is something demographically different about these participants, though the data available for the current study was not able to determine the reason for these extreme responses.
Table 14. K = 4 Demographic Means by Component

<table>
<thead>
<tr>
<th></th>
<th>Mixing Component 1</th>
<th>Mixing Component 2</th>
<th>Mixing Component 3</th>
<th>Mixing Component 4</th>
<th>Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size (Percentage of total sample)</td>
<td>105 (19%)</td>
<td>113 (21%)</td>
<td>206 (38%)</td>
<td>114 (21%)</td>
<td>1 (97% probability in component 3)</td>
</tr>
<tr>
<td>Mean Age</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>79 (75%)</td>
<td>76 (67%)</td>
<td>162 (79%)</td>
<td>90 (79%)</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>25 (24%)</td>
<td>32 (28%)</td>
<td>44 (21%)</td>
<td>23 (20%)</td>
<td></td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>African-American</td>
<td>34 (32%)</td>
<td>37 (33%)</td>
<td>81 (39%)</td>
<td>32 (28%)</td>
<td></td>
</tr>
<tr>
<td>African</td>
<td>2 (&lt;1%)</td>
<td></td>
<td>1 (&lt;1%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asian-American</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asian</td>
<td>1 (&lt;1%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caucasian</td>
<td>71 (68%)</td>
<td>70 (62%)</td>
<td>117 (57%)</td>
<td>79 (69%)</td>
<td></td>
</tr>
<tr>
<td>Native American or Alaskan</td>
<td>1 (&lt;1%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hispanic</td>
<td></td>
<td></td>
<td></td>
<td>1 (&lt;1%)</td>
<td></td>
</tr>
<tr>
<td>Pacific Islander</td>
<td></td>
<td></td>
<td></td>
<td>1 (&lt;1%)</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Year</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First Year</td>
<td>11 (10%)</td>
<td>10 (88%)</td>
<td>37 (18%)</td>
<td>18 (16%)</td>
<td></td>
</tr>
<tr>
<td>Sophomore</td>
<td>15 (14%)</td>
<td>29 (26%)</td>
<td>36 (17%)</td>
<td>18 (16%)</td>
<td></td>
</tr>
<tr>
<td>Junior</td>
<td>27 (26%)</td>
<td>29 (26%)</td>
<td>68 (33%)</td>
<td>50 (44%)</td>
<td></td>
</tr>
<tr>
<td>Senior</td>
<td>52 (50%)</td>
<td>45 (40%)</td>
<td>62 (30%)</td>
<td>28 (25%)</td>
<td></td>
</tr>
<tr>
<td>Graduate</td>
<td></td>
<td></td>
<td>1 (&lt;1%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missing</td>
<td></td>
<td></td>
<td>2 (&lt;1%)</td>
<td>1 (&lt;1%)</td>
<td></td>
</tr>
</tbody>
</table>
Figure 8. K = 4 Item Means by Component
Figure 9: Stacked FMM Item-Level 4 Component Model
In summary, the pattern that emerged seemed to be two subgroups with two sub-subgroups (see Figure 11). In support of this conceptualization, the behavior of Groups 1 and 2 were similar, and Groups 3 and 4 were similar. For the first overall subgroup, the behavior was that the level of resource loss or gain corresponded to the levels of depressive symptoms (for both Groups 1 and 2). The second overall subgroup seemed to represent participants who tended
to respond in extreme ways. So the ones who reported moderate gains and losses (Group 3) had less severe depressive symptoms than Group 4, which showed the most extreme responding patterns for both resources gained and lost, and reported depressive symptoms. Another possible explanation for response differences between group 1 with components 1 and 2 versus group 2 with components 3 and 4 is a different in social support from their respective social networks. Some university students may have better social support and better quality social network, which could in turn act as a buffer for stress.

Figure 11: Conceptual Model of 4 Component Solution
Chapter 5. Discussion

Rationally derived theories are not necessarily productive in guiding our attempts at classification and intervention. Our methods have slowly improved; though, there is still room for additional progress. For example, our classification system is still anchored to one of the unsupported theories of psychopathology (i.e., the separation of anxiety and depression). While the description and refinement of the categories may be informed by science, to date they mostly have not been created using objective methods (Radden, 2002; Watson, 2009). Additionally, our methods for hypothesis generation have not changed. Although hypothesis testing may move psychological science forward, it may inadvertently miss interesting findings. Such an approach may not allow other findings to emerge that were not a part of a study’s a priori hypotheses (Duda et al., 2012). On a positive note, there have been advances recently by researchers who are beginning to use quantitative methods and techniques for determining categories in the DSM-5 (Krueger et al., 2012). The current study is intended to continue this trend toward empirically derived classification and treatment in psychology, with an emphasis on understanding the experiences of people who experience traumatic events.

The Hurricane Katrina data from the current study may not look interesting at first glance. It is yet another study about undergraduates. The demographic breakdown is not particularly remarkable. In fact, using proven methods for finding subgroups (the goal of this study), as in the gap statistic and k selection functions, yielded nothing. The results were $k = 1$. This, however, does not mean that there is only one group. These methods, including the often-used k-means cluster analysis, do not work under certain circumstances, where the subgroups are not well defined and clearly separated. This is a problem given that the human experience is messy and the data would necessarily reflect that messiness. Psychopathology is not known for
being clean and well-defined. It has many mitigating factors. There are subclinical syndromes where people are suffering, but have not reached the threshold to say that they are in a diagnostic category. Psychopathology does not manifest itself in a vacuum. There are nuances to the context, causal factors, and presentation of psychopathology. Therefore, our analytic methods should help us understand this messiness. On the surface, the Hurricane Katrina data set did not look messy; in fact, it looked clean and homogeneous. But, on closer inspection, with more probing, the data began to reveal themselves, like breaking into a geode. Boring on the outside, beautiful on the inside. I contend that taking a look inside the data can aid in the advancement of our understanding of psychopathology, and by extension, our development of therapeutic interventions.

How does data reflect the human experience?

There are two ways to describe subgroups. One is to say that they are naturally occurring, which is usually within the context of observed data. The other way to describe subgroups is to discuss them in terms of their generative mechanism, which is usually within the context of simulated data. The underlying theme in both of these scenarios is that there is some cause that gave rise to the difference in values beyond expected individual differences. This is how the data reflect the human experience.

For example, with a naturally occurring subgroup, the subgroup membership represents differences in the causal factors. A person is exposed to a traumatic event. Based on the presence or absence of some unnamed factors, that person develops PTSD. If in examining a group of people, they all share these causal factors, except for a few who experience an additional event (e.g., dissociation), then the fact that the individual experiences dissociation makes the causal mechanism slightly different from those who do not have this experience. This
additional event or experience makes them different than the others who experience the traumatic event and develop PTSD without other complicating factors. There is something about experiencing dissociation during the event that makes the people who experience it have worse symptoms and, therefore, have a harder time getting better. An outlier in this situation would be if a person experienced a traumatic event and experienced dissociation during the event, but never developed PTSD. In fact, this person went on to do better after the event. In subgroups, the measured responses are similar to each other, but dissimilar to the other subgroups. The distinction between a subgroup and an outlier is if one can say that the causal mechanism is the same or not. In naturally occurring data structures, the structures are unknown. This is different from simulated data.

In simulated data, the real life event is replaced by a number generator. The number generator would be given specifications for the parameters of the subgroup. Using simulated data allows the researcher to control how the data are generated. For example, if a group of data represented one cluster, but another cluster was inside the data (like a donut hole in a donut), then one could test different techniques to find that donut hole cluster. There could potentially be a time when naturally occurring events created a similar type of data structure. Using the information gathered from the simulated data, researchers would now have techniques available for finding such a data structure.

This project used naturally occurring data, or observed data. Thus, the parameters were unknown. Using the techniques available and within the confines of this project, the analyses were used to test possible structures in the data. The first was to ascertain whether there were naturally occurring subgroups.
Why might the gap statistic and k selection functions not work?

The first step in this project was to test for the most likely value for k within the context of k-means cluster analysis. The gap statistic and k selection functions were the techniques recommended in the literature (Hastie et al., 2009; McLachlan & Peel, 2000; Pham et al., 2005; Tibshirani & Walther, 2005). The idea is that once the value for k is established for the k-means cluster analysis, the value for k for the FMM would be set at the same value. Then, the results from both analyses could be compared to further our understanding of how these models function on real world data.

However, the results of the gap statistic and k selection functions for finding the value for k proved to be more challenging than expected, because the value that was obtained was k = 1. In fact, this was the value for k under all of the conditions of the data from the item-level data to the three versions of the aggregated data (i.e., the component coefficients from the principal component analysis). Although it may seem like the analyses should stop there, both the gap statistic and the k selection functions do not work perfectly under all possible data structures, and a result of 1 does not necessarily mean that there is only one group (James et al., 2013).

One weakness of methods like the gap statistic is that global procedures are undefined for k = 1, and therefore, there is no way to determine if there are really no subgroups (James et al., 2013). Hidden structures can remain hidden, even when looking for them (Aggarwal, 2013). The gap statistic does not work well when the clusters are not clearly defined, such as having cluster overlap (Tibshirani & Walther, 2005). The k selection function also does not work well if there is one large group and a small group that are close together (Pham et al., 2005). In addition, the k selection function does not work well if the subgroups are close together, because the function will “see” the clusters as one cluster with noise (Pham et al., 2005). Both the gap
statistic and the k selection function work best when there are subgroups that are distinct and far apart (Pham et al., 2005; Tibshirani & Walther, 2005). Therefore, with the results from the present study, FMM was a better choice for finding the possible k solutions, because this model can find subgroups that are overlapping (McLachlan & Krishnan, 2007; McLachlan & Peel, 2000).

As further explanation for this choice it is helpful to revisit how subgroups are found using these two techniques. Similarity and dissimilarity within the context of k-means is measured by spatial proximity, where the spatial distance (i.e., Euclidean distance) represents differences in the value of the observed variable. Probability models, like FMM, assume that characteristics or patterns of characteristics of a population occur with varying frequency. The frequency of the patterns or single characteristics is called its density. The characteristics, in combination with the density, are a description of not only the population, but also the causal mechanism. Using the PTSD example, one would see that there would be a bimodal density function for those who developed PTSD. One mode would be most people who developed PTSD. The other mode would be those who experienced dissociation. Whatever it is that causes someone to experience dissociation is the causal mechanism for that mode.

However, not everyone who experiences dissociation does so in the same way. There are individual differences. For example, the average height of women is less than the average height of men. However, it is possible that a woman could be taller than a man. The distributions of men’s heights and women’s heights overlap. This overlap is reflected in the probabilities of membership. Going back to dissociation, there may be different magnitudes of this experience, which will make the probability of being in the dissociative group more or less. So if you are measuring PTSD symptoms and that person is not recovering quickly and has severe symptoms,
we can estimate the probability that this person is in the dissociation subgroup, based on the characteristics of the symptoms.

Conversely, a person may have more severe symptoms than the average person with PTSD and may be slow in recovering, but may not have experienced dissociation. This person may not look like the average person who has experienced dissociation. Based on individual differences, this individual may look like less severe individuals who experienced disassociation. The person may have a 60% chance of being in the PTSD no dissociation subgroup and have a 40% chance of being in the PTSD dissociation subgroup. There could also be a person with the same probabilities and same symptoms, but who actually did experience dissociation. In k-means, both people would most likely be in the no dissociation cluster. In FMM, the results would show the probability of membership in both groups. In this way, mixture modeling provides more information about the characteristics of those in each subgroup, because it allows for partial membership in a subgroup.

In the first round of analyses using the item-level data set, the posterior probability of membership, as found by the FMM, was either 99 to 100 or 0 for the given component. The assignment of membership is by the individual, and not the variable. Membership is reported by participant number, and not by the answer to question 2 on survey 1, for example. For the FMM, the proportion of membership was not stable between the solutions found for each seed value. This means that at each of the seed values the results in the proportion of membership had differences. The probability of membership was basically 1 or 0, which mimics the hard assignment of k-means cluster analysis. Even though the gap statistics and k selection functions were unable to find distinct subgroups, using both FMM and k-means cluster analysis, distinct clusters were found. Once the value for k has been specified, the algorithm will find that number
of subgroups whether they are present or not (Jain, Murty, & Flynn, 1999). Then, the question becomes: Are there any circumstances where the program could find distinct, as in no overlap, subgroups, when they really do not exist. The answer to this question is yes.

One scenario would be if there are truly no subgroups. When examining the item-level project data, the FMM suggested that there was no overlap between the subgroups. This could conceivably happen by dividing relatively homogeneous data into two parts. In k-means if subgroups are distinct from each other, then this would be reflected in the results by the ratio of the between sum of squares and total sum of squares. The ratio of the between sum of squares and the total sum of squares should be larger than the ratio of the within sum of squares to the total sum of squares. Therefore, the ratio of the between sum of squares should approach 1 by definition. In none of the trials for the item-level data analysis was this even remotely true. Most of the sum of squares was taken up by the within sum of squares. So to visualize these results there are two clusters that are relatively close together (otherwise the between sum of squares would be higher) and where there are various sizes of clusters. This could occur by splitting a homogeneous data set into two clusters, based on randomly selecting a place for division.

Another scenario can arise from highly dimensional data, the data points tend to be equidistant from each other (Aggarwal, 2013). Problems arising from high dimensional data would be more expected for a proximity-based algorithm, such as k-means cluster analysis, but not necessarily for a probabilistic model, like FMM (Aggarwal, 2013). It is difficult to imagine a situation where probability distributions are distinct solely because the data were highly dimensional. But, there are three scenarios where the gap statistic and k selection functions do not perform well, thereby leading to inaccurate results.
The first scenario is if one large subgroup and a very small subgroup are close together.
A second scenario is if there are two relatively equal in size subgroups, where the subgroups are close together but not overlapping. The third possibility is that there are no subgroups and that the model is “finding” subgroups because it was told to do so. If the data were relatively equidistant, the data could be divided in half at any random point, like the k-means. As an example, if a data set were comprised of sequential numbers, the data could be divided in half at any point. The means would be different and distinct, and there would be no overlap between the two distributions.

**Finite Mixture Modeling Allowed the Data to Speak**

In the present study, I tested a number of different solutions using FMM, given the limitations of the k means cluster analysis described above. When the FMM was applied to the item-level data for \( k = 2 \) to 5, the models \( k = 2-4 \) converged. The log likelihood for the \( k = 3 \) and \( k = 4 \) were the lowest, suggesting that these models were possibly the best fit. Of these two models, differences were not really apparent in the demographics, though to be sure further testing is suggested. But an interesting pattern showed up in the \( k = 4 \) model that did not show up in the \( k = 3 \) model, which was that there were two mixing components that had similar patterns in the depressive symptoms, but somewhat different patterns in the loss of resources. In \( k = 4 \) component 1, the resource losses were more pronounced than for those participants represented in \( k = 4 \) component 2. The gains, on the other hand, were sporadically more for those in component 2 than in component 1. By using FMM where \( k = 4 \), a general difference in reporting patterns emerged. One group’s depressive symptoms seemed to be connected with levels of loss/gain, where two levels of loss and depressive symptoms were represented. One level was very little or no losses and the lowest level of depressive symptoms. The other level
was more losses with more severe depressive symptoms. Therefore this overall subgroup with two levels reflected the predicted relationship between resource loss and depressive symptoms. The other overall subgroup, on the other hand, reflected perhaps a general tendency toward extreme reporting, again with two levels. One level for this second subgroup included moderate losses and gains, which seemed to be associated with moderate depressive symptoms. The second level included extreme losses and gains, which were associated with the most severe depressive symptoms. The tendency for extreme reporting across the different measures would need to be examined in future studies. Initial examination did not find any demographic characteristics that would lend insight into the reasons for these moderate to extreme responses across all measures included in the current study. Of course, the restricted range of demographic characteristics for the current sample (e.g., due to the sample being comprised of university students in the southeast U.S.) are likely limiting the ability to draw any conclusions. Another possibility is that there is an unmeasured causal mechanism that is driving the difference. If we had not looked for possible patterns, this pattern would have remained hidden. So, even though additional questions were generated from the current study, this is one example of how unsupervised statistical learning can lead to new research questions by letting the data speak for themselves.

**Comparing the Two Methods: K-Means Cluster Analysis versus FMM**

*Instability of results*

In the preliminary analyses, the results were wildly different, which was not surprising as it was expected that both methods would have different results each time the analysis is run (Hastie et al., 2009). The literature describes this outcome as being sensitive to the initial values (Hastie et al., 2009; James et al., 2013). These initial values refer to the initial parameter values,
such as the centroid for k-means and the mean for FMM (Hastie et al., 2009; James et al., 2013). These initial parameter values can make a huge difference in the final solution for both models (Hastie et al., 2009; James et al., 2013). K-means tends to be more sensitive than FMM, which is an artifact of the algorithm (Hastie et al., 2009). The reason that k-means is more sensitive than FMM to the initial values is because k-means uses hard assignment to assign cluster membership to each participant’s responses (Hastie et al., 2009). Hard assignment refers to the membership assignment being all or nothing for k-means. This is in contrast to FMM, where membership assignment can be partial (Hastie et al., 2009). Therefore, as an example, given that k-means hard assignment is all or nothing, if the initial parameter value happened to be on or near an outlier, then the whole cluster could get pulled toward that outlier, which does not happen as drastically in FMM because the membership assignment is partial or “soft” (Hastie et al., 2009). Any local pathology (i.e., outliers) can have an effect on the final solutions for both methods, but the effect is greater in k-means than in FMM (Aggarwal, 2013; Hastie et al., 2009).

Therefore, it is expected that both models will have unstable results, which is why the literature suggests running each model several times (Duda et al., 2012; B. Everitt et al., 2011; Hastie et al., 2009; James et al., 2013; McLachlan & Peel, 2000). After which, the final solutions can be evaluated for validity, if necessary. For this project instead of running the models several times and validating the final solution, the initial values were “seeded,” which means that the random generator did not generate the initial parameter values, rather, they were user-specified by the seed value. This was done for two reasons: 1) comparisons between each run for each model would not be comparable, otherwise, and 2) controlling the initial values made it easier to evaluate the stability of these models based on this data set.
To control the randomness of the random number generator, the set.seed function was used to simulate, but in a controlled way, random assignment of values. That way there would be different starting values, but since these values would be based on the same seed for all data conditions and analytic methods, cross comparisons of the results would make more sense.

*Instability in the proportion of membership*

The proportion of membership is a reflection of which participants have been assigned to which subgroups. There was instability of membership observed in the preprocessed data, but less so than in the data sets that were created by reducing the dimension. The four conditions of the data are discussed below.

**Item-level project data.** For the FMM, there was variation in the mixing proportions (i.e., the percentage of participants assigned to either component). Although this is expected, it demonstrates that the data may have some local pathology (i.e., there may be some widely different behavior of the data when just looking at one small portion of the data). Additionally, the algorithm failed to converge with two of the seed values. It is expected for convergence to slow under certain conditions, but failure is not generally expected, unless there is local pathology in the data (McLachlan & Krishnan, 2007). From this alone, it is suspected that there are some data localities that are not behaving as one would expect.

For the k-means cluster analyses, there was some variation in the membership proportions. This is, of course, expected. It was interesting to note that the variations did not seem as large for this data set as with the FMM results. This was unexpected, and the reason for this is unclear.

**PCA four-component with no rotation solution.** This data set had only 35% of the explained variance. Under this condition, the FMM proportion of membership stabilized, and
the mixing components did show some overlap between the components. It is highly suggestive that this portion of the data really did not have more than one subgroup. The proportion of membership for the k-means analysis was more stabilized, but not nearly as much as the FMM. The most likely explanation for this is the difference in soft versus hard assignment of data points.

**PCA nine-component with varimax solution.** This data set had only 48% of the explained variance, which is more than the previous data set. The proportion of membership results from the FMM was exactly the same as the previous data set. However, the results from the k-means were different from the previous data set. There seemed to be more shifts in the membership from the previous data set. This is most likely due to the hard assignment found in the k-means algorithm, which is expected because it is known that the hard assignment of the k-means can add to drastic shifts in membership when running the algorithm on the same data set several times (James et al., 2013).

**PCA 19-component with no rotation solution.** This data set had 64% of the explained variance. The FMM results were virtually the same as the previous two data sets. There was some instability in the results from the k-means analysis. The most likely explanation for the difference between the FMM results and the k-means results, again, is likely the effect of soft vs. hard assignment.

**Observations Based on the Demonstration**

*Curse of dimensionality further discussed*

There are a number of problems in finding accurate results when there is highly dimensional data, both in terms of d subgroups and outliers. There is a notion called sparsity of data, which describes what happens when data becomes more and more dimensional (Aggarwal,
The data point pairs tend to become more equidistant (Aggarwal, 2013; Duda et al., 2012). Because of this, subgroups and outliers can become more difficult to find. In some cases the literature suggests using principal component analysis to reduce the dimensions of the data set for k-means cluster analysis, which was the method used in this project (Dash, Mishra, Rath, & Acharya, 2010; Duda et al., 2012; Napoleon & Pavalakodi, 2011). However, using PCA to reduce the dimensions can introduce new problems in finding subgroups and outliers (Aggarwal, 2013). This procedure can produce a loss of information, which could make it more difficult to find subgroups or outliers (Aggarwal, 2013). There are methods that are currently being developed for finding outliers, especially if those outliers are few in number, not behaving the same way, item-level, and a result of highly dimensional data (Aggarwal, 2013).

Importance of feature selection

It should be noted that the failure for the gap statistic and the k selection functions to find subgroups in this archival data set may also be because the two measures (i.e., depression and COR) are not actually related. Looking at items in concert with each other assumed a relationship. But, people may be depressed regardless of their resource loss or gain. The university students could be depressed for other reasons than experiencing a natural disaster. If students were experiencing emotional distress, there could be a number of reasons for this, including dealing with a break-up or being stressed about a particularly difficult assignment. Another problem could be that any emotional distress was actually better labeled as a trauma reaction, rather than depressive symptoms. PTSD does have a depressive symptom component (Friedman et al., 2011). There was no measure for trauma-related symptoms, nor was there a measure of other sources of stress beyond the conservation of resource loss. This speaks to the importance of feature selection (Duda et al., 2012), which is to say selecting what to measure.
New Questions

On the surface, the Hurricane Katrina data set did not seem that interesting. Given the weaknesses of some of the techniques that were used, it is unexpected that some interesting questions arose. This is all the more noteworthy as this question would not have arisen had I not continued to test the data. The data were structured in such a way that there were no clearly and easily identifiable groups (at least not initially), but after repeated attempts the FMM for $k = 4$ revealed an unanticipated pattern. There were two groups with very similar depressive symptom patterns, but their resource patterns were dissimilar. One had very pronounced losses with small to moderate gains. The other had moderate losses with small if any gains. One would expect that moderate losses would lead to less severe depressive symptoms. This pattern did not show up in $k = 2$ or $k = 3$ models. It was only when the data was divided into four groups that this was apparent.

This finding suggests that the items that reflect specific resource losses and gains may work in concert. It may also suggest that certain resource items function differently with the depressive symptoms than other resource items. It may also suggest that there is an unmeasured causal mechanism that is occurring. Do some kinds of gains serve to mitigate the effects of certain losses more than others? How do these items interact with each other? One way to begin exploring this is to do an item analysis, such as a Graded Response Model. These results support three notions. One, letting the data speak for themselves leads to the discovery of unexpected knowledge. Two, finer measurements, such as item-level data, can lead to uncovering knowledge more than grosser measurements. Third, the only way to make sure that a data set does not have hidden gems is to actually go digging.
Limitations

There are a number of limitations to this study. First, no conclusions should be drawn about the population or the theories that support the measures used. The study was not conceived to draw conclusions about the data. The study was designed to demonstrate how to use unsupervised statistical learning techniques. For example, the handling of missing data was predicated on the focus of the present study being an examination of statistical learning techniques and not the actual data.

That said, it is important to point out some of the flaws of the data, as they could have influenced the results. The data were collected from undergraduate students, a population not easily generalizable to the general public. There are a number of confounding variables specific to that population that would not be applicable to those outside of that population. The losses may be moderated by the fact that at least some of them are financially dependent on their parents, so they may not feel the loss of resources as strongly since the resources were provided by their parents, and potentially could be replaced by their parents. Furthermore, feeling a lack of time and like one does not have enough money for extras, two items of the COR survey, may be attributable to their being undergraduates, rather than losses felt by the aftermath of the hurricane. It is entirely possible that given a more diverse population and more participants, that the subgroups may be different or may be more pronounced. It is difficult to say.

Another limitation of this study is grounded in the limitations of the statistical learning techniques themselves. For example, k-means cluster analysis does not perform well when the clusters are not distinct. Since the data are multivariate, the curse of dimensionality might have made finding subgroups more difficult. This can occur because the higher dimensionality can make the data more equidistant, thus making any subgroup less distinct. This is a continuing
area of research in the data science field. Another example is that there is no clear method to test for a single subgroup, other than to determine the possibility that there is more than one subgroup. Both the gap statistics and the k selection functions gave the results of k = 1, but these results do not actually say that k = 1, only that the functions were unable to fail to reject the null. There are a number of situations where there could be more than one subgroup, but these subgroups would not necessarily be detected by those functions. One of those situations, that was clearly the case in this data set, is that if there were subgroups that these subgroups would be close together and not very distinct from each other. This weakness is reflective of the model on which these functions are based, which is k-means. They were designed to find k-mean clusters. A superior technique for a situation such as this data set is where the clusters are allowed to overlap, like FMM. Another technique would be fuzzy cluster analysis, which was not examined in this study. This is an additional limitation of this study; the techniques examined were not comprehensive.

A limitation in the interpretation of the subgroups, especially in the 4-component model, is the available data used in the current study. The data set was archival and a subset of the original study data, so the available variables were limited. Thus, it is difficult to make definitive conclusions about the patterns found in the 4-component model, which in turn can only offer tentative hypotheses. In future research, it would be important to include a sample with more diversity (e.g., age, education, race/ethnicity) and measures on social networks and social support. In particular, including data on social support may prove interesting in terms of a modifier for the expression of depressive symptoms.

Finally since this was a demonstration project, precise conclusions about the techniques themselves cannot be drawn. Without the aid of simulated data, comparative performance
evaluations cannot be made. For example, comparing methods on known structures of data, where the structures were systematically manipulated, would give better information about the comparative strengths and weaknesses of the techniques. To my knowledge, this has not been done. Given the results of the present study, I believe that further research into the comparative performance of statistical learning techniques is warranted.

**Concluding Remarks**

Subgroups were not found using the k-means functions of the gap statistic and k selection for this data set. There are a number of reasons why the gap statistic and k selection functions did not find subgroups. These reasons include possibly poor choice of feature selection, local pathologies of the data (i.e., hidden outliers), the highly dimensional nature of the data, artifacts of the data reduction techniques used, and finally, the possibility that there were no subgroups. Using k-means under any of these situations makes it less likely that hidden structures would be identified. With these situations, FMM is the better choice as FMM allows for partial membership.

Using FMM, the k = 4 model most likely has the best fit, both in terms of the log likelihood result and the information provided by the FMM analysis. There was a little overlap between the subgroups. There were three items that had 97% probability of membership. The rest of the membership was split between 99% and 100%, which indicates groupings that are not clearly defined. K-means works best with clearly defined clusters (James et al., 2013).

Applying FMM to k = 2+, comparing the log likelihood, and examining the patterns revealed an interesting pattern. This pattern included two components with similar depressive symptoms but different patterns of loss. If we consider solely the Hobfoll COR theory of stress (1988, 1989, 2001, 2004; 1993), then there should not be similar patterns of depressive symptoms with
different patterns of loss. This means that there is a possible moderating variable. It is possible that this moderating variable could be used as a target for interventions. If resource loss causes symptoms in a predictable pattern, more loss should equal more depressive symptoms. By contrast, in the present study, different losses equaled similar depressive symptoms.

In light of these findings, there is something interesting going on in the data that raises a question: is there something not measured that has contributed to that difference? This question in a nutshell shows why we might consider letting the data speak for themselves. Although the results from the present study are not definitive on drawing conclusions about the population, measurements, or even the techniques, we need to continue to advance our methods and classification systems in order to understand more fully psychopathology and responses to traumatic events, with the hope of developing treatment approaches that help alleviate human suffering.
References


Cannon, W. B. (1932). The wisdom of the body.


IBM. (Released 2013). IBM SPSS Missing Values 22. Armonk, NY.

IBM. (Released 2013.). IBM SPSS Statistics Base 22. Armonk, NY.


Leisch, F. (2004). FlexMix: A general framework for finite mixture models and latent glass regression in R.


McMahon, J. E., & Vernon, A. E. (2010). *Selections from the Writings of Albert Ellis, Ph.D.*


Appendix A. K-means Algorithm

The K-means algorithm is as follows (James et al., 2013, p. 388):

1. Initially randomly assign an observation $j$ to a cluster 1 to K.

2. Iterate until cluster assignments of the observations stop changing:
   2.1. For each cluster $k$, compute the cluster centroid (e.g., cluster mean), where
   \[
   \bar{x}_{jk} = \frac{1}{N} \sum_{i \in k} x_{ij}
   \]
   is the cluster centroid for observed variable $j$ in cluster $k$.

   2.2. Assign each observation to the cluster whose centroid is closest, as defined by the
   squared Euclidean. Then repeat step 2.1 until there are no longer observation
   reassignments.

   Because the initial step is randomly assigned, it is recommended to run the algorithm
   multiple times (James et al., 2013). The best solution, where the assignment of the observations
   assignments best minimize the within cluster distortion is selected, as defined by minimizing the
   squared Euclidean distance between the observed data and the cluster centroid (James et al.,
   2013).
Appendix B. Finite Mixture Modeling EM Algorithm Description

Mixture modeling is a completely different approach to finding subgroups than K-means cluster analysis. In K-means cluster analysis, data is segmented as a function of Euclidean distance. In mixture modeling, a probabilistic approach is used to find subgroups (McLachlan & Peel, 2000). The goal of the mixture modeling analysis is to find subgroups of data in terms of probability distributions (McLachlan & Peel, 2000). This is done by using the algorithm called expectation-maximization or EM (McLachlan & Peel, 2000). Suppose there is a population that has been measured and has an associated probability distribution. Let us suppose that the probability distribution is bimodal. This was the case when Karl Pearson was first developing mixture modeling (Pearson, 1894a, 1894b). Using a dataset given to him by a colleague on the ratio of the forehead to body length of crabs (Weldon, 1892, 1893), Pearson found that the ratio of the forehead to body length formed a bimodal distribution (Pearson, 1894a, 1894b). Weldon and Pearson hypothesized that the bimodal distribution may represent two subspecies of the crabs (Pearson, 1894a, 1894b; Weldon, 1892, 1893). Pearson, then, set out to develop a statistical technique that could find subgroups within a population, which was termed moments-based mixture modeling where the analysis was limited to normally distributed data (McLachlan & Peel, 2000; Pearson, 1894a, 1894b). Pearson’s solution had extremely complicated calculations that included nonic equations (e.g., to the power of nine; McLachlan & Peel, 2000).

Given the complex calculations required for Pearson’s solution, mixture modeling as a statistical technique was impractical until two major advancements: maximum-likelihood applications (ML) in the 1960s (McLachlan & Peel, 2000) and the expectation-maximization algorithm (EM) first proposed by Dempster, Laird, and Rubin (1977). It is in the expansion of
the originally proposed EM algorithm that mixture modeling techniques have advanced (Benaglia et al., 2009; Leisch, 2004).

As EM algorithm has been developed further in terms of what can be estimated and what assumptions have to be met, the EM algorithm has become a more general term for the two-step procedure (Duda et al., 2012). Originally, the algorithm assumed that there is missing data in the observed dataset, that the mixing components were normally distributed, but that the proportion was unknown and that missing data needed to be estimated using ML (Dempster et al., 1977).

Today, there are various functions available in the statistical program R to estimate the number of components (given a known distribution family), the parameters of the mixing components, and various other scenarios. The iterative two-step procedure remains relatively the same in that assigning the membership of the observed variables to a mixture component is part of the expectation step and in that the maximization step calculates the parameters and mixing proportions of the components based on the assigned members. The methods for extracting subgroups from a data set are based on a different formal definition for subgroups.
Appendix C. Items of CES-D and COR

The CES-D scale.

In the past week:

1.00 = "rarely or none of the time (less than 1 day)"
2.00 = "some or a little of the time (1-2 days)"
3.00 = "occasionally or a moderate amount of time (3-4 days)"
4.00 = "most or all of the time (5-7 days)"

CES-D Items

1. I was bothered by things that usually don't bother me
2. I did not feel like eating
3. I felt like I could not shake off the blues
4. I felt I was just as good as other people
5. I had trouble keeping mind on what I was doing
6. I felt depressed
7. I felt like everything I did was an effort
8. I felt hopeful about the future
9. I thought my life had been a failure
10. I felt fearful
11. My sleep was restless
12. I was happy
13. I talked less than usual
14. I felt lonely
15. People were unfriendly
16. I enjoyed life

17. I had crying spells

18. I felt sad

19. I felt that people dislike me

20. I could not get "going"

The COR Evaluation Items Scale.

-4.00 = "very great loss"
-3.00 = "great loss"
-2.00 = "moderate loss"
-1.00 = "little loss"
.00 = "no change"
1.00 = "little gain"
2.00 = "moderate gain"
3.00 = "great gain"
4.00 = "very great gain"

COR Items

1. sentimental possessions
2. personal transportation
3. time to get things done
4. money for living expenses
5. adequate food
6. feelings that I am accomplishing my goals
7. adequate clothing
8. support from coworkers of friends at school
9. animal or family pet
10. furniture, appliances, and household contents
11. closeness with at least one friend
12. opportunity to attend church or religious services
13. free time
14. family stability
15. closeness with one or more family members
16. companionship
17. a good relationship w/ my family or children
18. necessary tools or supplies for work or school
19. time with loved ones
20. necessary appliances for home
21. children's or family's health
22. feeling valuable to others
23. time for adequate sleep
24. feeling independent
25. housing that suits my needs
26. personal health
27. status/seniority at work
28. water or something to drink
29. a positively challenging routine
30. home that is more than what is needed
31. clothing that is more than what I need
32. stable employment
33. intimacy with spouse or partner
34. adequate furnishing for home
35. intimacy with at least one friend
36. a role as a leader
37. essentials for children
38. acknowledgment for accomplishment
39. "extras" for my children
40. understanding from my employer/boss
41. savings or emergency money
42. spouse/partner's health
43. support from co-workers or friends at school
44. adequate income
45. money for "extras"
46. companionship
47. adequate credit (financial)
48. financial assets (stocks, property, etc.)
49. affection from others
50. financial stability
51. people I can learn from
52. money for transportation
53. help w/ tasks at work or school
54. help w/ tasks at home
55. loyalty of friends
56. help w/ children
57. involvement in organizations w/ others who have similar interests
58. advancement in my education or training

59. health of family/close friends
Appendix D. Scree Plots

Scree Plots