Constructing A Revised Version of the Face Stimulus Assessment to Measure Formal Elements: A Pilot Study

Donald C. Mattson
CONSTRUCTING A REVISED VERSION OF THE FACE STIMULUS ASSESSMENT TO MEASURE FORMAL ELEMENTS: A PILOT STUDY

By

DONALD C. MATTSON

A Dissertation submitted to the Department of Art Education in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Degree Awarded
Spring Semester, 2011

Copyright © 2011
Donald C. Mattson
All Rights Reserve
The members of the Committee approve the dissertation of Donald C. Mattson defended on February 4, 2011.

David E. Gussak
Professor Directing Dissertation

Anuj Srivastava
University Representative

Marcia L. Rosal
Committee Member

Pat Villeneuve
Committee Member

Approved:

Dave Gussak, Chair, Art Education Department

Sally McRorie, Dean, School of Visual Arts and Dance

The Graduate School has verified and approved the named committee members.
“Everything is number.”
Pythagoras
To Dr. James Filgas, for consistently lighting a fire under my ass.
ACKNOWLEDGMENTS

First, I would like to thank Dr. Dave Gussak (not Gussack), for fighting for me. A particular thanks also goes to Dr. Marcia Rosal, for giving me a chance. Many thanks to Pat Villeneuve, my now-colleague and fellow Yooper. Thank you also, Dr. Srivastava, for representing the technical side of my dissertation process. Thanks to John C. Russ, Wayne Rasband, and John Pickle for their expert advice on image analysis. I would also like to thank all those responsible for getting me through the dissertation process.
# TABLE OF CONTENTS

LIST OF TABLES ........................................................................................................ xi

LIST OF FIGURES ................................................................................................... xii

ABSTRACT ............................................................................................................. xiii

CHAPTER 1 INTRODUCTION .................................................................................1

  Problems to Be Investigated ..............................................................................5

    Purpose of Study ...............................................................................................5

    Research Problems ..........................................................................................5

    Justification of Study .......................................................................................5

    Hypothesis ........................................................................................................8

    Main Instruments Used ....................................................................................8

    Brief Overview of Study ..................................................................................9

    Research Questions and Assumptions ............................................................11

    Definition of Terms .........................................................................................11

    Conclusion ........................................................................................................16

CHAPTER 2 LITERATURE REVIEW .....................................................................18

  Computer Programs to Rate Formal Elements .................................................18

  Historical Work ..................................................................................................21

    Art Assessment of Depression .......................................................................21

  Test Item Pool Sources .....................................................................................26

    Test Construction Considerations ..................................................................27

    Sources of Item Content ...............................................................................29

    Justification for Main Instruments Used .........................................................38
LIST OF TABLES

Table 1 Demographics of Control Group………………………………………..53
Table 2 Demographics of Experimental Group…………………………………..56
Table 3 Differences in Formal Element Percentages between Experimental and
    Control Group……………………………………………………………….....67
LIST OF FIGURES

Figure 1 Brightness histograms……………………………………………………66
Figure 2 Examples completed by experimental and healthy controls……………72
ABSTRACT
The Face Stimulus Assessment-Revised (FSA-R) is an art-based instrument constructed from elements of the Face Stimulus Assessment (FSA, Betts, 2003). The pilot test involved computerized rating of formal elements between those with Major Depressive Disorder \((n = 20)\), and controls without known diagnosis of Major Depressive Disorder \((n=20)\). Significance resulted from a multiple \(t\)-test analysis of the data. In response to the hypothesis that the formal elements of color and/or free space from the FSA-R, rated by public domain image analysis software (PDIAS), can distinguish Major Depression artwork from control group artwork, this study concluded that certain colors and free space distinguished the groups. Those with Major Depression drew less purple \((t(38) = -2.95, \ p = .05, \ d = -.96)\), less orange \((t(38) = -2.28, \ p = .05, \ d = -70)\), and more left free space \((t(38) = 2.26, \ p = .05, \ d = .73)\) than controls. As a result, it may be possible for the FSA-R to become a standardized instrument for screening Major Depression.
CHAPTER 1
INTRODUCTION

With the advent of healthcare reform, there is now more demand than ever for demonstrating the effectiveness of therapeutic modalities (Mattson, 2010). Agencies that recognize and fund therapeutic practices want concise instruments that are cost-effective and provable (American Psychological Association, 2009). The Face Stimulus Assessment-Revised (FSA-R) holds the potential to be such an instrument.

The FSA-R is an instrument intended to measure graphic distortions caused by Major Depressive Disorder (MDD). The main problem to address is whether features of MDD artwork differ in comparison to controls, as calculable by computer analysis. An investigation of both the presence of MDD pathology in artwork and the feasibility of the programs that analyze the pathology underlie the construction of the FSA-R. The research strategy for the experiment involved pilot testing groups to obtain norms while systematically tailoring the instrument for measures of reliability and validity.

In an effort to generate valid and reliable results from an art-based mental health assessment, I began research on the quantification of patient artwork through public domain image analysis programs (PDIAS), originally developed by the National Institute of Health. While generally designed for the analysis of biomedical imagery, I adapted these programs for the assessment of artwork (Mattson, 2009, 2010; Mattson & Veldorale-Brogan, 2010). These studies, along with expert system technology (Kim, S. I., Kim, Y. Lee, Lee, & Yoo, 2006), formed the basis for the current development of the FSA-R instrument. The construction of the FSA-R is largely a reaction to the need for increased reliability and validity of art-based instruments (Betts, 2005a; Brooke, 2004, Gantt & Tabone, 1998).

This study is similar to the forerunning assessment studies of art therapists Linda Gantt, Myra Levick, and Rawley Silver. These individuals conducted extensive research within their dissertations that led to the development of the assessment instruments known as the Person Picking an Apple From a Tree (PPAT, Gantt, 1990), the Levick Emotional and Cognitive Art Therapy Assessment (LECATA, Levick, 2001), and the Silver Drawing Test of Cognition and Emotion (SDT, Silver, 2003). The PPAT became a useful quantitative assessment instrument in art therapy, while the LECATA offered a comprehensive testing manual (Brooke, 2004). Silver (2003) studied assessment as far back as the 1960s. Her doctoral dissertation (1966) led to the
development of the SDT, an influential work in the development of the Face Stimulus Assessment (FSA, Betts, 2003). The instrument used in my study is a modified version of the FSA, and the rating procedure is an extension of previous computer-assisted techniques developed by S. I. Kim, Yoo, Kim, and Lee (2007). Several formal elements applied to the FSA are tested (Hamilton, 2008). This and other studies explored graphic indicators of psychiatric symptoms, of which there is much interest in the realm of art-based assessment (Gantt, 1993; Gravitz, 1969; Holmes & Wiederholt, 1982; Jamison, 1993; Williams, Agell, Gantt, & Goodman, 1996).

Similar studies delineating elements of artwork for assessment have been conducted in the past. Otherwise known as the sign method, the single element approach supposedly determines the patient’s condition. For example, a sketchy mark on a page could indicate anxiety. While assessors still debate the assessment of a single sign, one researcher found statistical promise when analyzing sign clusters (Stora, 1963). A subsequent section of this dissertation, historical work, further details this concept.

A few key studies that are especially pertinent to this dissertation are worth mentioning. One of the most notable of these is the study on the Descriptive Assessment for Psychiatric Art (DAPA, Hacking, Foreman, & Belcher, 1996). The other relevant study involves the Sheppard Pratt Art Rating Scale (SPARS, Bergland & Gonzalez, 1993).

Previous research attempted to distinguish patient and non-patient groups using only features of artwork. Hacking et al. (1996) discovered that it was possible to differentiate certain categories of mental illness based on quantified formal elements of subject artwork. Their instrument, the DAPA, revealed that orange, yellow, color intensity, and line—all distinguished diagnostic categories of affective disorder, non-affective psychosis, brain injury, and drug abuse with alpha levels set at .05. Additionally, this study revealed differences between these diagnostic categories and a control group. For instance, the drawings of those with depression featured less form, more free space, heavier black color, and more painted lines than those in control groups. Form, in these cases, was defined as any spatial attribute within a drawing; free space was any unused portion of the paper. “Heavier” use of black generally entailed the liberal quantity in which it was used. The results of the Hacking et al. study were consistent with other findings (Carstairs, 1967; Cohen, 1988; Dax, 1953; Kirk & Kertesz, 1989; Wadeson, 1980).
A similar study posed the question whether mental illness indicators in artwork were quantifiable. Bergland and Gonzalez (1993) issued the SPARS to hospitalized \( (n = 67) \) and non-hospitalized patients \( (n = 63) \) in an effort to differentiate personality disorders. Through the application of SPARS to 67 artwork portfolios, these researchers found significant differences in the scores of six individuals, and notable differences through indicators of others.

Mental illness indicators in artwork are also evident through *formal elements* in drawing. Gantt and Tabone (1998) helped formalize these elements named through categorical scales: Prominence of Color, Color Fit, Implied Energy, Space, Integration, Logic, Realism, Problem-Solving, Developmental Level, Details of Objects and Environments, Line Quality, Person, Rotation, and Perseveration. These elements vary between diagnostic categories. The questioned documents theory section further explains this theory.

These constants and variations in artwork are assessable with some degree of reliability and validity. For instance, drawings from individuals with Major Depression generally featured less color and more free space (Gantt & Tabone, 1998; Hammer, 1958; Wadeson, 1980). This is a recurrent observation throughout art-based assessment literature (Cavézian, Danckert, Lerond, Daléry, d’Amato, & Saoud, 2007; Dawson, 1984; Gantt & Tabone, 1998). As a result, the formal elements within the Prominence of Color scale and Free Space scale theoretically distinguish between Major Depression and non-depression groups within this dissertation. Additional elements studied, such as line quality and omission of details, are ancillary to this focus, though included within the pilot test results. The sign method applied to the study of these elements.

The use of the sign method for diagnosis is highly debated. Many argue for a global approach to the assessment of drawings that considers biopsychosocial information prior to diagnosis (Kaplan, 2003). Some individual graphic analyses prove more promising than others, such as the analysis of line quality in diagnosing depression through art. Current neuropsychological research now links psychomotor retardation (PR) copying tasks to the actual diagnosis of Major Depression using alpha levels of .05 (Hegerl et al., 2004; Mergl et al., 2007; Sabbe, Hulstijn, Hoof, & Zitman, 1996; Thomassen, Meulenbroek, & Tibosch, 1991; Pier, Hulstijn, & Sabbe, 2004a; Pier, Hulstijn, & Sabbe, 2004b; Sabbe, Hulstijn, Hoof, Tuynman-Qua, & Zitman, 1999). This dissertation includes the development of art-based items adapted from these copying task templates. Development of all items within the FSA-R is justified through the literature review.
The use of a stimulus drawing, such as the FSA (Betts, 2000), modified to include additional stimulus templates in the FSA-R, may yield more distinguishable formal elements than free drawings utilized in previous studies. A contained drawing with the start of a face and a few objects likely presents fewer variables to standardize than a free drawing. The FSA-R used stimulus drawings for these reasons.

The Person Picking an Apple from a Tree (PPAT) instrument uses the Formal Elements Art Therapy Scale (FEATS), which relied on inter-rater agreement in order to generate valid assessment scores (Gantt & Tabone, 1998). In the instance of the FSA-R, the stimulus templates could contribute to reliability. The procedures section provides a complete overview of FSA-R development, while the discussion section outlines avenues for quantitative reliability and validity.

To quantify effectively a stimulus drawing such as the FSA-R, the standard drawing scans into a computer, and the National Institutes of Health (NIH) public domain image analysis software (PDIAS) known as Image Java (ImageJ) analyzes it. The other public domain software used is Measures of Vegetative Health (MVH), or MVHimage. The use of image analysis software results in more intensive and precise analysis (Mattson, 2009). These programs are capable of objectively measuring most formal elements in graphic representations. They often feature high accuracy ratings (Murakami, Turner, Van Den Berg, & Schaberg, 2005). This work incorporated previous research in computer-assisted image analysis software known as the Computer Color-Related Art Therapy Evaluation System (C-CREATES, Kim, Bae, & Lee, 2007). The study I am conducting is different in that the test and control group formal elements underwent a comparative study using publicly available software.

Unlike previous work relying solely on human raters, computer-assisted rating is capable of measuring the near-exact percentage of color, line dimension, form, intensity, hue, and value, which may ultimately yield a test with appreciable reliability and validity (Mattson, 2009). Because of the statistical and numerical procedures employed, a mixed methods post-positivist research approach provided the theoretical framework for this study, along with Test theory. Bell’s discourse comprises the philosophical standpoint. The section on philosophical and theoretical underpinnings explains the tenets of this dissertation.
Problems to Be Investigated

Purpose of Study

The purpose of this study is to pilot test an objectively rated instrument for Major Depression to determine the significance and feasibility of both the format of the test, and computer rating of the resulting artwork. This work incorporated and expanded research in both art-based assessment and computer image analysis. Developing and refining an assessment instrument will likely take years to complete (Cohen, Hammer, & Singer, 1988). Consequently, this study is ongoing. The ultimate purpose of this study will be to construct an automatically rated visual screening test for Major Depression rated by image analysis software to address shortcomings of manual rating, reliability, and validity.

Major Depression is the most common mental disorder in the United States affecting roughly 16% of individuals 18 years and older, and it is the leading cause of disability for ages 15 to 44 (National Institute of Mental Health, 2008). Early detection of depression improves short-term prognoses (Maj, Lopez-Ibor, Sartorius, Sato, & Okasha, 2005). For these reasons, Major Depression will be the main mental illness assessed through this instrument, and the construction of the instrument is the focus of this dissertation.

Research Problems

The main problem of the FSA-R study is whether features of Major Depressive Disorder (MDD) artwork differ in comparison to controls via computer analysis. The concept of using computers to score instruments is a relatively new subfield of art therapy that requires investigation (Mattson, 2009; Kim, S. I., Kim, Y. H., & Kim, 2008). Computerized rating underlies the whole construction of the FSA-R instrument.

Computerized rating of art-based instruments is still in its infancy (Kim, Bae, & Lee, 2007). If PDIAS successfully rates the drawings by those with Major Depression and those without, then the degree of accuracy will require further audit. The field of art therapy would likely benefit from an improved rating scale for existing measures of formal art elements (Amos, 1982; Brooke, 2004; Kramer & Lager, 1984; Wadeson, 1980). In time, reliability and validity studies could establish the efficacy of the PDIAS rating system.

Justification of Study

Art therapy should continue to grow through research recognizable by established fields in order to offer effective and widespread services. The fields that take center stage in services
are those with the most powerful legislators (J. A. Rubin, personal communication, May 28, 2008). These legislators require evidential research to fuel their causes. To date, computerized assessment is billable under certain health insurance codes (P. Craven, personal communication, April 27, 2009). In order to qualify for this code, however, the technique must be credible.

To become credible, art-based assessments should establish validity and reliability through standardization (Brooke, 2004). Common procedures, theoretical approaches, and observational terms are all important characteristics of valid fields (Boyd, Gasper, & Trout, 1991). Increased validity of computer-assisted art therapy assessment may result from standardizing processes of image analysis techniques. This and other studies have begun to explore objective methods of analysis applied to art assessment. Joining science and art could benefit art therapy by offering standardization and worth (Kaplan, 2003). This study incorporated practices and procedures from objective disciplines.

Art therapy will likely profit from using objective measures, though not everyone in the field agree. Critics of rigorous methods herald protests, stating that art therapy must remain qualitative in nature, because science would subtract from its allure (Brooke, 2004; Phillips, 1994). The rebuttal to these criticisms includes the importance of assessment for ongoing research, reimbursement, and recognition.

Art therapy will benefit from studies of conjoining scientific disciplines (Kaplan, 1998). Art therapy is still in its infancy; it is unrecognized in many areas. As a result, reimbursement for services becomes difficult. Managed care recently mandated the use of new instruments with varied formats (Stout, 1997). There is an increasing demand for brief, symptom-focused instruments due to cost containment and changes in documentation (Groth-Marnat, 2003). Computer rating of art-based instruments could further these qualities and foster establishment within the field.

Objective measures within established fields often feature standardization. Therefore, the FSA-R instrument, materials, and instructions are standardized and compiled into a manual. Gantt (2000) outlined important advances needed for assessments alongside existing successes with the FEATS. They included comparisons to existing psychological tests, generalizability, reliability, validity, and usability. The ultimate purpose of assessment research, however, is to discover predictive variables to assess, or even diagnose, particular disorders (Rosal, 1992). This is a major aim of the FSA-R.
Many art-based assessments exclude quantitative data. An overview of major art-based assessment tests revealed few with reasonable validity or reliability (Brooke, 2004; Seitz, 2001). To achieve credibility, art-based assessments should establish validity and reliability through standardization of techniques and assessments (Brooke, 2004). It is important to standardize measurement techniques so tests may be replicated; they should also measure what they claim (Fishman & Galguera, 2003).

Developing standardized methods in assessment rating and scoring could assist in generating reasonable levels of reliability, validity, and objectivity. There appears to be few objective solutions to problems arising from the interpretation of drawings (Kim, Bae, & Lee, 2007). Subjective scoring usually results in notable statistical error because human-scored image analysis is qualitative (Carpenter et al., 2006). A test featuring some measure of quantitative objectivity may remedy problems of inter-rater reliability (Adkins, 1960). A computerized rating scale could offer such a measure. For example, the Prominence of Color subscale from the Formal Elements of Art Therapy Scale (FEATS, Gantt, 1998) adapted to the FSA had a low level of inter-rater reliability (Hamilton, 2008). PDIAS programs are capable of measuring color prominence with high degrees of accuracy (Mattson, 2009).

Gantt (2000) developed a means of analysis using manual methods from the FEATS. She also outlined advancements that researchers can make with this scale. These included comparisons to psychological tests, generalizing, and advancements in reliability, validity, and usability. The use of computerized rating measures will likely address some of these main issues.

For instance, concerning measurement of scale number 4 (Space) in the FEATS, Gantt (1998) stated, “[w]e have plans to measure this variable more precisely by using a computer and a scanner. Thus, we could convert this scale into a powerful ratio scale and be precise in detecting differences between groups” (p. 34). This illustrated a proposal of using computers to improve rating scales. In addition to the computerized measurement of the space element, researchers can also measure color, line, and shape (Mattson, 2009).

With the eventual development of a series of valid and reliable art-based tests, assessment procedures may change. A clinician might be able to administer a barrage of art-based tests to glean an adequate picture of the client’s mental status prior to diagnosis. In short, a series of these tests could replace conventionally verbal personality inventories—a valuable approach for
largely nonverbal or resistant clients. Prior to this work, however, an initial hypothesis of this
dissertation lays the groundwork for further inquiry.

**Hypothesis**

The hypothesis states that PDIAS can rate the formal elements of color and/or free space
to distinguish artwork of those with Major Depression from the artwork of those without Major
Depression. To test this hypothesis, a series of instruments will be required.

**Main Instruments Used**

Several instruments integral to the formation of the FSA-R are worth mentioning at the
outset of this work. The FSA-R borrows components from each one. They include the Face
Stimulus Assessment (FSA), the Beck Depression Inventory (BDI-II), and the Formal Elements
Art Therapy Scale (FEATS). A brief mention of these instruments follow, and the justification
for their inclusion is present in the literature review.

The Face Stimulus Assessment Platform (FSA). The Face Stimulus Assessment-Revised
(FSA-R) is a modified version of the Face Stimulus Assessment (FSA) originally developed by
Donna Betts, which she constructed in response to nonverbal clientele (Betts, 2003). The FSA is
a series of three drawing templates measuring 21.59 cm x 27.94 cm. The first contains a pre-
drawn face. The second is an outline of the face. The third is a blank sheet of paper. In
*Developing a Projective Drawing Test* (Betts, 2003), there are references to many avenues of
future research regarding the FSA:

> To develop a valid and reliable rating method for FSA, researchers need to establish a
> method to rate those elements in the drawings that graphically demonstrate identified
> concepts, abilities, or skills—such as the role of color, figure and size differentiation,
> drawing style, compositional strategies, and the use of pictorial space. (p. 4)

The Beck Depression Inventory (BDI-II). The Beck Depression Inventory in its shortened
form (BDI-II) detects a wide range of depression symptoms (Groth-Marnat, 2003). The BDI-II is
a 21-item self-report of depressive symptoms derived from psychiatric research (Beck, Steer, &
Brown, 1996). It was widely used for screening depression amongst psychiatric patients for
many years (Steer, Ball, Ranieri, & Beck, 1999). The test is sensitive enough to detect depression
in both psychiatric populations and normal individuals (Beck et al., 1996).

Individuals completing this test receive instructions to answer the items in accordance with
how they feel within a 2-week period. They answer these items by means of an intensity
scale from 0 to 3. Typical questions include dimensions of low mood, irritability, guilt, failure, and disturbances in sleep and appetite (Groth-Marnat, 2003). After the administration of the inventory, which usually takes anywhere from 5 to 10 minutes, the scores are added and assessed with an interpretive scale. Clinically depressed individuals generally score between 14 and 28 (Beck et al., 1996).

The Formal Elements Art Therapy Scale (FEATS). Gantt and Tabone (1998) developed the *Formal Elements Art Therapy Scale* (FEATS). It attempted to address the possible presence of symptoms within artwork based on the Diagnostic and Statistical Manual categories (American Psychiatric Association, 2000). Standardization was another goal of this instrument, and the FEATS addressed this aim. It is a research tool developed for furthering credibility within the art therapy field.

The content of the FEATS assessment resulted from the initial work of Lowenfeld and Brittain (1947). The directive is to draw a person picking an apple from a tree. From this, the participant produces a single drawing. An administrator or aid rates this drawing by comparing it to 14 scales in Likert format.

These 14 scales differentiate diagnostic categories through formal elements. Ten out of 12 scales distinguish two or more diagnostic groups (Brooke, 2004). Scales that distinguish one group from another at the alpha level of .05 include: Person, Details, Developmental Level, Problem Solving, Logic, Realism, Integration, and Energy. Inter-rater reliability is .90 or higher for most scales, including Major Depression (Gantt & Tabone, 1998).

Public Domain Image Analysis Software (PDIAS). Public domain image analysis software (PDIAS) rates the FSA-R drawings. The primary rating software used is Image Java version 1.43 (ImageJ) developed by the National Institutes of Health. Its main purpose is the analysis of biomedical imagery, though it also applies to art (Hedges, 2006). Measures of Vegetative Health (MVH), or MVHimage v.8, is an ancillary program used for this study. It originally analyzed topographical imagery from satellite photography. Other research applications are pending for this software.

**Brief Overview of Study**

The basic structure of this study follows the steps to instrument design proposed by Fishman and Galguera (2003): literature review, conceptualization of components, drafting of item pool, formatting, and pilot testing initial administration. Subsequent steps for future
research include deleting, adding, and formatting the test items to prepare the instrument for a second administration. For this study, the pilot test stage is terminal. The *methodology* section outlines these stages.

The first stage is the literature review. The design of a valid test depends on its literature review. An instrument requires construction based on information from its substantive field. Research has revealed the *componentiality* that a test will likely require. A literature review also provides the building blocks for subcomponents of the test. Research is helpful to support varying aspects of the phenomena measured by the items.

The second stage is the conceptualization of components. This means determining and selecting the components or dimensions of a test. In many ways, this stage determines what respondents will likely answer to inform a valid test. One should strive to build a test from a number of separate parts comprising unitary phenomena. This requires sound judgment of the literature and methods used on the part of the researcher. Experience I gained during an internship at Professional Testing, Incorporated (PTI) informed this dissertation.

The third stage is the actual drafting of the item pool. It is prudent to start with a larger item pool than needed to account for deletions. Kline (1987) recommended twice as many items as thought necessary for the actual test. The best test constructors build items one-at-a-time (J. Hanson, personal communication, April 6, 2007). A respondent should easily understand the items. Items also need to be concise, unambiguous, relatively free of bias and *ipsativity*, and relatable to the criteria a researcher desires to test. Above all, items need to exhibit mutual independence of each other within a certain range. Fifteen items for every dimension generally attenuates response error.

The fourth stage is formatting. The format of a test varies wildly on its application. It can be a questionnaire, scale, or a host of other designs. The natures of the questions are either dichotomous or continuous. An instrument should be neither too long nor too short in length. It should be appealing to the respondent, with clear instructions. A reasonable time limit needs to be set for its completion, preferably 15-30 minutes (Hambleton, Mills, & Simon, 1983).

The fifth and final stage is the administration of the pilot test. The subjects and conditions of the pilot test should approximate the ones of the actual test. Initial pilot test time should take longer than intended time allotted. Subjects should not receive the impression that the test is a pilot, but that it is the actual test. Elements in the testing room require standardization for
experiments (Christensen, 2001; Cronbach, 1984). The tester needs to standardize and record lighting, seating arrangement, materials, ventilation, temperature, time, and other environmental factors. It is prudent to end the pilot test with an interview about the structure of the test to garner information for improving it. The results of the pilot test will address the hypothesis, and several basic research questions.

**Research Questions and Assumptions**

The research questions and assumptions are as follows:

1. Can PDIAS rate the formal elements of color and/or free space to distinguish artwork of those with Major Depression from the artwork of those without Major Depression?
2. Can the FSA-R instrument become a standardized instrument for Major Depression?
3. Do common graphic indicators exist amongst the artwork of those with Major Depression?
4. Once completed, can components of this computer-rated instrument generalize to other assessments, or populations?

These questions and assumptions were developed to address the hypothesis that the formal elements of color and/or free space from the FSA-R, rated by PDIAS, can distinguish artwork of those with Major Depression from the artwork of those without Major Depression. The answers to these questions will likely contribute to the specialized body of knowledge surrounding the topic of image analysis and art-based assessment. The results of this research are directly useful to the subfield of assessment in art therapy. The definitions of terms used in this dissertation are in the next section.

**Definition of Terms**

The following terms are included in this dissertation or relevant to the topic investigated.

**Acquisition**: The initial step in image analysis (Russ, 1997). It involves acquiring an image through sensors, cameras, range calculators, or scanners.

**Archaic transition**: A stage within a proposed spectrum of art therapy where the artist struggles to draw or paint within object boundaries (Simon, 1992).

**Algorithm**: A set of computerized instructions (Puntambekar, 2009).

**Artificial Intelligence**: The machine capability of performing functions normally associated with human intelligence, such as reasoning and judgment. Artificial Intelligence (AI) is a branch of
computer science that approximates human reasoning by using factual and heuristic knowledge (Ginsberg, 1993).

Art for Life: A teaching model of art that encompasses branching areas of the human existence (Anderson & Milbrandt, 2004).


Barnum effect: The inclination to accept vague or generalized results as true (Marks, 2000).

Binary conversion: An image analysis technique that reduces images to distinct black and white points, leading to easier measurement. This also allows for greater discernment between figure and ground relationships (Gonzalez & Woods, 2002).

Brightness value: The average luminance of a scene. Brightness value (BV) is a typical measurement of digital media ranging from 0 to 250, 250 being the lightest value (Russ, 1997).

Class characteristics: Shared writing traits and styles amongst commonly socialized groups (Hackney, 1999).

Classical Measurement Theory: Classical test theory assumes that each participant has a true score obtainable if there were no errors in measurement. The resulting model is \( X = T + E \), where \( X \) = observed score, \( T \) = true score, and \( E \) = error (Allen & Yen, 2002).

Clipping: Distortion of a digital signal in which high amplitude peaks are cut off, usually caused by electronic overload (Watkinson, 2008).

Coloremes: Graphic equivalents to alphabetic letters (Saint-Martin, 1990).

Componentiality: The individual parts comprising a test, synonymous with categorical dimensions (Fishman & Galguera, 2003).

Convenience sampling: Selecting subjects for research in a non-probabilistic manner (Christensen, 2001).

Data processing reliability: The level of consistency in handling data (Robertson & Williams, 2009).

Discriminability: The prime goal in the analysis of test items. It is the subtraction of a particular item response across subjects from the total score of its contribution (Fishman & Galguera, 2003, p. 64)
**Discrimination**: The second step in image analysis (Russ, 1997). It is a feature that separates imagery through basic color scaling.

**Distractors**: All multiple-choice answers that are not the true answer (Downing & Haladyna, 2006).

**Dysthymia**: An overwhelming yet chronic state of depression, exhibited by a depressed mood for most of the days, for more days than not, for at least 2 years, without a Major Depression mood (American Psychiatric Association, 2000).

**Edge detection**: A means to determine significant changes in value in a computerized image. These significant changes mark the edges of colors or objects. Using this tool, a researcher can separate these colors or objects for further study (Allen & Mills, 2004).

**Experimenter effect**: When an experimenter consciously or unconsciously delivers cues to the test subjects (Ary, Jacobs, Razavieh, & Sorenson, 2009).

**Expert system**: Software that attempts to reproduce the performance of human experts in a specific area to solve problems. It is an application of artificial intelligence that uses knowledge bases and subject expert information to solve problems (Jackson, 1998).

**Filtering**: A pre-processing method used to reduce image noise (Russ, 1997).

**Fitt’s task**: A psychomotor retardation task where patients connect a series of circles through drawn lines (Sabbe et al., 1996).

**Formal elements**: Categories of drawing features. May include: Prominence of Color, Color Fit, Implied Energy, Space, Integration, Logic, Realism, Problem-Solving, Developmental Level, Details of Objects and Environment, Line Quality, Person, Rotation, and Perseveration (Gantt, 1998).

**Gaussian blur**: An image processing filter that reduces pixel brightness. It works by blurring an image through concentric brightness attenuation of image pixels (Shapiro & Stockman, 2001).

**Geons**: Mental representation of basic shapes derived from common objects (Biederman, 1987). This is a cognitive function used for organisms to make sense of complex figures.

**Greyscale**: A continuous monotone image in shades of black and white (Salomen, 2004).

**Hawthorne effect**: When participants increase performance upon learning that they are being tested (McCarney, Warner, Iliffe, Haselen, Griffin, & Fisher, 2007).
**Hemineglect**: This is a disabling condition following insult to the brain (Houde, 2003). Patients with hemineglect are unaware of objects to one side of their peripheral view. For whatever side of the brain is affected, the opposite side of perception becomes vacuous.

**History effect**: Events occurring before or between testing that affect the outcome (Christensen, 2001).

**Horror vacui**: Filling nearly or all usable space within an artwork (Prinzhorn, 1972).

**Ill-structured paradigm (ISP)**: A vague domain where no clear objective solutions for analysis exist (Kim, Bae, & Lee, 2007).

**Image processing**: A general term for the acquisition and manipulation of digital images (Russ, 1997).

**Instrumentation**: Any change in equipment or materials leading to inconsistent experimental outcomes (Ary, Jacobs, Razavieh, & Sorensen, 2009).

**Instrument**: A tool used in a standardized way to support, amplify, or specifically direct our senses in investigating human thought, experience, and behavior and to those means of research that are used to evaluate or analyze data or even to support theoretical explanations (Sturm & Ash, 2005).

**Instrument reliability**: The degree to which responses in an instrument are uniform (Black, 1999).

**Interaction of personal variables**: Any individual difference substantial enough to affect experimental outcomes (Houser, 2009).

**Internal variation**: Significant variables in a writer that contribute to a style vastly different from the writer’s original style (Osborn, 1978).

**Ipsativity**: A funneling effect of test items where later items answered by the respondent become constrained by previously answered items. The influence of one test item on a subsequent item that tends to skew testing results (Fishman & Galguera, 2003).

**JPEG**: Joint Photographic Experts Group image file format. A format with little data loss, resulting in images that are faithful to the original image (Salomen, 2004).

**Mandala**: Refers to a circular drawing used in history by many cultures to represent the self or outside world. It is a reflection of the inner conscious (Jung, 1968).

**Mask**: An image separate from an original that is usually the result of digital filtering (Gonzalez & Woods, 2002).
**Measurement**: The fourth step in image analysis (Russ, 1997). It is any observation expressed by a number, according to rules (Glass & Hopkins, 1996).

**Median filter**: A common tool used in image analysis (Gonzalez & Woods, 2002); it is effective for most image analysis applications (Russ, 1997). The median filter works by ranking neighboring pixels in a square, selecting the brightest pixel, and positioning it in the center of this square.

**Mortality**: Participant attrition or death, often affecting experimental outcomes (Ary et al., 2009).

**Multiple treatment interference**: When previous treatment modalities affect current testing performance (Houser, 2009).

**Noise**: Any unwanted information in the image that may detract from analysis (Russ, 1997).

**Novelty effect**: An increase in test performance based on novel stimuli (Clark, 1994).

**Participant effect**: When subjects improve an aspect of their measure behavior in response to the fact of being studied (Christensen, 2001).

**PDias**: Public domain image analysis software. Downloadable software usually free of charge and used to analyze digital images (Mattson, 2009).

**Pixel**: A single point of color in a digitized image (Salomen, 2004).

**Projective instrument**: An assessment that presents ambiguous stimuli so that the respondent projects personal meaning onto its interpretation (Hammer, 1958).

**Psychomotor retardation**: A key symptom of Major Depressive Disorder manifested through the slowing of cognitive and motor processes (Pier et al., 2004b).

**Region of interest (ROI) tool**: A shape drawn around the area that requires measurement (Russ, 1997).

**Reliability**: Refers to the degree that an instrument is stable, consistent, predictable, and accurate (Groth-Marnat, 2003).

**Segmentation**: The third step in image analysis (Russ, 1997). It refers to partitioning regions of pixels (Shapiro & Stockman, 2001).

**Selection bias**: When systematic error in sample selection affects the outcome of the study (Szklo & Nieto, 2006).

**Sigma value**: The standard deviation of a filtering operation (Russ, 1997).

**Sign method**: Delineating components of artwork for the purpose of assessment as opposed to viewing the work holistically (Brooke, 2004).
Stem: The test question prompt (Osterlind, 1998).
Subject effect: When a test subject expects a certain outcome, thereby changing the result of the test (Gray, 2002).
Skeletonization: A common processing tool that reduces the foreground pixels of a binary image to its basic midline (Russ, 1997).
Stimulus drawing: A standard drawing with visual prompts intended to encourage the completion of a task (Silver, 2001).
Striatum: The inner region of the brain consisting of a pair of masses forming the basal ganglia (Finger, 2001).
Subject/observer reliability: The ability of experimenter and subjects to conduct an experiment in a consistent manner over a series of trials (Robertson & Williams, 2009).
Test: An evaluative device or procedure in which a sample of an examinee's behavior in a specified domain is obtained, evaluated, and scored using a standardized process (American Educational Research Association (AERA) et al., 1999).
Threading: Horizontal graphic movement with little vertical commitment (Thomassen et al., 1991).
Thresholding: An image processing technique that separates elements by means of pixel brightness discrimination (Gonzalez & Woods, 2002).
TIFF: Tagged Image File Format is a common type of image file format that contains a variety of descriptive information as well as the image data itself (Salomen, 2004).
t-test: A test of a statistical hypothesis about the population means (Glass & Hopkins, 1996).
Validity: The degree to which an instrument measures what it intends (Groth-Marnat, 2003).
Zaner-Bloser: The primary system of handwriting taught within the school systems of the United States (Hackney, 1999).

**Conclusion**

This chapter included the research problem, rationale, and justification of the study with the selected instrument. Previous work noted differences in formal elements between groups. The FSA-R instrument is used to test the hypothesis that formal elements of color and/or free space will significantly differentiate people with Major Depression from those without Major Depression through PDIAS rating capability. The results come from controlling extraneous variables, offering means of standardization, conducting image analysis, and incorporating
statistical procedures. The research questions, assumptions, brief overview, and definition of terms were included in this chapter. The literature review is in the next chapter.
CHAPTER 2
LITERATURE REVIEW

This literature review begins with emergent research in computerized art-based assessment. It then segues into a brief account of existing assessment that this technology seeks to improve. Subsequent sections illustrate various art-based assessments developed through time.

The topics of this literature review are important for several reasons. For instance, many issues of protocol arise with the use of computers in art therapy (Peterson, Stovall, Elkins, & Parker-Bell, 2005). Art-based instruments are widely used by mental health professionals to formulate effective treatment plans (Betts, 2005; Brooke, 2004). Despite the popularity of usage, art-based assessment research is poor and generally obsolete; little research points to the reliability and validity of such assessments (Seitz, 2001). Research into the computerized assessment of art is relatively new. Therefore, this literature review will explore topics relevant to these domains. The first section of this chapter explains some of the issues involved with art therapy and technology, particularly with the use of computer technology for assessment.

**Computer Programs to Rate Formal Elements**

Since the emergence of computer technology, new opportunities arose for art therapists (Gussak & Nyce, 1999). Art therapists use this technology throughout their practice to produce images, to teach directives, to communicate with clients, and even to conduct therapy (Peterson et al., 2005). Approximately 83% of all art therapists surveyed used computer technology in one form or another (Orr, 2005). Of these, a large percentage never received training in the use of digital media, owing to a definite need for expansion in this realm. Throughout all this research, there is hardly any mention of using computers for assessment.

Existing work in computer rating of formal elements began with an article titled *An Expert System Approach to Art Psychotherapy* (Kim, S. I., Ryu, Hwang, & Kim, 2006). This article stated that art psychotherapy offered numerous challenges to machine mimicry of human constructs. The researchers proposed an expert system to model the human decision making process, which, they stated, would be particularly useful to the diagnosis of children’s drawings, as children readily exhibit psychical information in artwork (Dileo, 1970). Regardless of the age assessed, cultural and personal variables tend to skew standardization (Wegmann & Vusenbrink, 2000). Additionally, art psychotherapy to these researchers is an ill-structured paradigm (ISP), which means multitudinous variables, such as artistic ability and cultural context, complicate
representative standardization. To address the problems in mimicking the human decision process, these researchers proposed a remedial model. This model underpinned the Expert System for Diagnosis in Art Psychotherapy (ESDAP) present in the follow-up article.

In subsequent work by Kim, Ryu et al. (2006), an expert system theoretically supplanted human reasoning in the art assessment process. Early researchers of artificial intelligence presented initial models of human reasoning and compared them to artificial intelligence (Giarratano & Riley, 2005). The idea of interactive relations between elements within a given artwork complicates machine diagnosis. The human mind is capable of higher order processing to judge and make decisions, and this is a difficult model for a machine to simulate. These researchers proposed several conditions for simulating human diagnosis, forming the eventual framework for art knowledge in subsequent publications.

Kim, S. I., Yoo, Kim, & Lee (2007) explained the ISP of art therapy assessment. It required complex models in order to work as an expert system. Such a system, however, was useful to address client resistance. Additionally, there was the question of making a program for non-experts. A solution for this ISP would be to separate all extraneous variables into classificatory systems, and attempt analysis by convergence of all of them in a logical manner to arrive at a diagnosis. Even with this system, subjective knowledge is required to judge the decision of the system, reinforcing the notion of the ISP.

This, and other initial models served as a springboard for future articles published by this group, and became a forerunning philosophical approach to their future work. Take, for instance, A Computer System to rate the Color-Related Formal Elements in Art Therapy Assessments (Kim, Bae, & Lee, 2007). In this article, the researchers established the importance of color in assessment (Hollins, Horrocks, & Sinason, 1998; Kreitler & Kreitler, 1980; Lowenfeld & Brittain, 1970; Malchiodi, 1998; Milne & Greenway, 1999; Steinhardt, 1977). They borrowed color elements from Gantt and Tabone (1998), and used an expert system to analyze these elements through two case studies of children completing the Diagnostic Drawing Series (DDS, Cohen, 1988).

Aside from color, these researchers studied DDS placement in the context of computerized analysis (Cohen, 1985; Kim, S. I., Kang, & Kim, 2008). They supported the initial tenets of the DDS that placement was indicative of emotional and psychological states, and figure placements were rated as unusual or normal. For this study, the researchers rated drawings
through grids on a page; they analyzed drawings completed by elementary aged children ($n = 50$). Human raters analyzed the drawings alongside the computer system, and the researchers discovered that the human raters presented with less accuracy in determining placement categories than the computerized rating system.

Kim, S. I., Kim, Y. H., & Kim (2008) then designed a computer program to track the progress of client color usage within mandala drawings. This system relied on theoretical connections between personal color preference, and mood states (Fincher, 1991; Gantt, 1998; Lüscher, 1971; Rorschach, 1951; Wadeson, 1980). While largely in its prototypical stage, these researchers promised to offer automatic mandala rating to the public in the near future.

In later work, Kim, S. I., Kang, & Kim (2009) delineated elements of the structured mandala. Using mandala examples from individuals with dementia ($n = 58$), they correlated Mini-Mental Status Examination-Korean (MMSE-K) scores with color clusters, edges, prominence, and order, and discovered strong relationships between level of dementia and scores.

A single color was also important for analysis (Kim, 2008). Assessment often neglects the main color used, and human perception of main color varies greatly. A computer algorithm determined the main colors used with case examples.

More recently, Kim, S. I., Betts, Kim, & Kang (2009) correlated the MMSE-K with structured mandala scores rated by their expert system. They conducted a multiple regression test, and discovered that light green and brown correlated positively with dementia. Additionally, they discovered that the mandala scores and the MMSE-K scores correlated significantly.

Mattson (2009) outlined several image analysis techniques intended to complement interpretive analysis through demonstrating public domain image analysis software (PDIAS) capabilities adapted for measuring formal elements (Gantt & Tabone, 1998). An FSA drawing completed by an individual formally diagnosed with schizophrenia served as the model. This study also extended manual and computer-assisted ratings through this model by using readily available PDIAS. Case study results indicated that PDIAS is able to analyze accurately formal elements of an FSA drawing. Research implications included improved inter-rater reliability in the adapted Prominence of Color Scale, and the development of highly accurate art therapy assessment rating scales for those without extensive computer analysis backgrounds.
Mattson and Veldorale-Brogan (2010) explored placement and count of standardized sand tray figures based on a sample of adults \((n = 5)\) completing the Erica Method in an effort to address the need for higher accuracy in sand tray assessment (Danielson, 1998). The study demonstrated image analysis capabilities and limitations in assessing both 3-dimensional object count and placement. Public domain image analysis software (PDIAS, Mattson, 2009) objectively determined total sand tray figure count and total sand tray figure area used. Human raters \((n = 4)\) assessed the same criteria. Human-PDIAS inter-rater reliability was higher than human inter-rater reliability. As a result, computerized rating may hold potential for more efficient rating based on this study.

Mattson (2010) explored some of the main issues springing from the computerized assessment of art-based instruments (CAABI). The results indicated that there were opposing viewpoints, limitations, and solutions regarding the limitations. Despite opposition and limitation, the author concluded that there is a need for increased research in CAABI based on the advantages it offered, such as ease-of-use, early detection, less scoring time, reduction of subjective human error, improved statistical measures, and resiliency in the face of healthcare reform (Kim, Ryu, et al., 2006; Kim, et al., 2006; Kim, S. I., Kim, Y. H., & Kim, 2008; Kim, Betts et al., 2009).

Im, Oh, Chung, Yu, and Lee (2010) applied the DataBase Management System (DBMS) to the Kinetic Family Drawing (KFD). This entailed storing retrievable KFD objects in a computerized database so that the therapist could assess drawings more efficiently. The purpose of the tool is to aid clinicians scoring the KFD.

This was the extent of research featuring computer rating of art-based assessments. The next section is a historic account of sources linking depression to artwork. It begins with the early 19th century writings of psychiatrists attempting to make sense of “insane art” in early period writings. Accounts of actual assessments made about depression through art in the 20th century follows. Last, this section outlined sources on depression and art in the current era.

**Historical Work**

**Art Assessment of Depression**

Betts (2005) noted that research in assessment of patient artwork spanned across 100 years of history. European writers commented on the art of patients in mental wards as far back as 1843 (MacGregor, 1978). The French psychiatrist Pinél mentioned the “art of the insane” in
the early 19th century (as cited in Anastasi & Foley, 1941). Another French psychiatrist named Tardieu published a reproduction of a mental patients’ artwork in 1872 (MacGregor, 1978). Just a short while later, Lombroso (1891/2010) examined the artwork of those within mental wards, and arrived at brief impressions about what they meant.

The use of art as diagnosis followed in the early 20th century (Klepsch & Logie, 1982). This included the work of Fritz Möhr, the German researcher who sought standardized methods for the assessment of art. Prinzhorn (1972) assembled a collection of patient artwork in the early 20th century, around the time of Möhr. He gave impressions of artwork from those with schizophrenia and bipolar disorder, but did not engage in diagnostic practices. This was to come later in history. All of this became the forerunning work of mid-twentieth century assessment instruments, such as the House-Tree-Person (Buck, 1948), Thematic Apperception Test (Murray, 1943), and others (Betts, 2005).

From the 1940s onward, there was tremendous interest in art as a means of personality assessment throughout the psychoanalytic fields. These often took the form of projective art techniques. Buck (1948) developed the House-Tree-Person (H-T-P) as a projective means of assessment for adult and children’s drawings. It offered the client a means to depict the self and relations to family. Additionally, it determined developmental levels (Burns, 1987). The H-T-P drawing directive aided researchers in determining depression with nonverbal patients (Mölner, 2008).

Around the same time, the Human Figure Drawing Test (HFD) focused on emotional indicators in lieu of developmental drawing aspects (Koppitz, 1968). Koppitz discovered that those with depression \( n = 14 \) frequently omitted eyes, nose, and mouth, and tended to draw smaller figures. Numerous mentions of indicators of depression of the HFD occurred throughout literature. An overview of literature in the HFD revealed increased instrument reliability and validity over time in figure size and its relation to depression (Clifford, 1968). For instance, the drawings of 50 patients with depression were significantly shorter \( p < .01 \) than those of 50 non-depressed patients (Lewinsohn, 1964).

One of the forerunning studies of categorizing and analyzing content in art assessment focused on artwork from Manic-Depressive patients (Wadeson & Bunney, 1970). Distinguishing characteristics recorded were color, line, configuration, organization, and affect between phases of bipolar II. Presence or absence of mania configured on a scale of 1 to 15, 15 being the most
present. Based on their explorations, these scales proved to be about 50% accurate in distinguishing these bipolar phases.

The Kinetic Family Drawing (KFD) also used constructed scales (Burns & Kaufman, 1972). This series of family drawings coincided with a rating scale known as the Family Drawing Depression Scale (FDDS). The FDDS paired well with the Zung Self-Rating Depression Scale (SDS), making this a useful and reliable measure for depression (Wright & McIntyre, 1982). The FDDS scale was again tested by researchers using pre- and post-test measures of participants ($n = 19$) at a pain clinic, and results coincided with reports of depression from hospital staff (Sawyer, 1987).

Throughout the clinics of this period, the Rorschach test of personality became increasingly popular (Holley, 1973). The Rorschach boasted objective and multivariate statistical procedures, and it discriminated psychiatric categories, such as depression. Vegelius (1976) spearheaded this analysis. Additionally, Hampson and Kline (1977) further supported this study through the analysis of criminal personalities using the Rorschach.

The Mandala Assessment Research Instrument (MARI) also tests personality through artistic components (Kellog, 1978). The originator based its design on thousands of mandala drawings from psychiatric patients. It rooted itself in the theory that the drawing of the mandala changed through developmental stages (Cox, 2002). Bruscia, Shultis, and Denner (2007) found that it held potential to assess depression stemming from inpatient cancer and patients of cardiac disease. Inconsistencies in card choice and mandala drawing color components may be indicative of depression (Cox, 2002).

Henderson, Rosen, and Mascaro (2007) conducted an experiment using the mandala where participants in the experimental ($n = 19$) and control condition ($n = 17$) were tested with the mandala paired with the BDI-II pre- and posttest. They found that those with depression used dark colors. Further investigation into the differences between the drawings of depressed and non-depressed adults occurred later (Dawson, 1984). Dawson hypothesized that the drawings of those with depression scoring on certain spectrums of the BDI would exhibit less color, more free space, smaller figures, more omitted details, more shading, and less extra details than those of the control participants. The group with depression left more free space in their drawings and excluded more details than the non-depressed group. The difference between the group means differed, but not significantly.
The Kinetic School Drawing (KSD), which is a variant of the KFD, was capable of analyzing similar elements (Knoff & Prout, 1985). Some indicators of depression in this assessment included figures drawn small in relation to peers, or figures that appeared asleep (Klepsch, 1988). Like the other kinetic drawings, this was mainly a projective instrument. However, Andrews and Janzen (1988) developed a rating scale, scoring sheet, and a reference list of depression indicators. Using their methods revealed appreciable inter-rater agreement above .70. More learning disabled than non-learning disabled children drew themselves in a way that scored highly on depression indicators.

The Diagnostic Drawing Series (DDS) also used clinical information from clients, and featured several sources on its use with depression (Cohen, 1985). The structured and semi-structured drawing tests used criteria from early versions of the Diagnostic and Statistical Manual of Mental Disorder (DSM) (Brooks, 2004). Cohen stated that the DDS was useful in assessing mental dysfunction. Later research explored its potential for assessing depression (Brudenell, 1989; Cohen et al., 1988; Gulbro-Leavitt, 1988; Gulbro-Leavitt & Schimmel, 1991; Johnson, 2004; McHugh, 1997). According to this research, the DDS drawings of samples correlated with instruments such as the Center for Epidemiologic Studies Depression Scale (CES-D), and the Millon Clinical Multiaxial Inventory in its second form (MCMI-II). Some inter-rater reliability statistics were also studied. Additionally, these references found trends in the use of less than one third of paper, less idiosyncratic colors, higher shading, geometric shapes, and animals depicted in place of people for drawings by those with depression. The results of most of this research, however, are fraught with both methodological and validity problems (Betts, 2005).

To improve H-T-P methodological limitations, Burns (1987) developed the Kinetic House-Tree-Person Test (K-H-T-P) to address shortcomings with the H-T-P, as non-kinetic techniques failed to address dynamics and interactions (Knoff & Prout, 1985). This test combined all three images on one sheet of paper. Burns presented an appendix in his manual that included indices of depression from the sign method of assessment. An example would be placement of the figure on the lower portion of the page, which signified depression (Buck, 1948; Machover, 1949).

Burns (1990) also created the Family-Centered Circle Drawings (FCCD) to assess symbolic representations of feelings towards clients’ parents. Projective material stemmed from
the notion of a client's inner and outer parents, and such symbolism could bring a client closer to the center. Within this test, Burns (1990) included a list of symbolic references. The presence of moons within this drawing signified depression (Burns, 1987).

Other tests used heavy client symbolism around this time. The Draw-A-Person Test: Screening Procedure for Emotional Disturbance (DAP-SPED) identified children and adolescents who may have been mentally disturbed (Naglieri, McNeish, & Bardos, 1991). The 20-item D-A-P differentiated dementia from depression with cutoff scores (Clément, Marchan, Boyon, Monti, Léger, & Derouesné, 1996). The D-A-P paired with the BDI revealed slight differences in drawn figure size and depression (Holmes & Wiederholt, 1982).

Silver (2001) discovered that a large amount of adolescents with depression presented with negative themes in their artwork as opposed to any other group she studied; this was significant at $p<.005$ to $p<.0005$. In response, she developed the Silver Drawing Test of Cognition and Emotion (SDT, 1983, 1990, 1996, 2002), a precursor to the FSA. It is a stimulus drawing with scores extracted from self-image and emotional content, analyzed as $t$-scores and percentile ranks (Silver, 2002). Silver claimed the test contained validity and reliability. However, Rosal (2006) reviewed material on the SDT and purported that it has yet to use established tests to implement concurrent validity; it did not account for the complexity of the mental illness constructs it intended to assess. Despite evidence for the utility of stimulus drawings, some research revealed its shortcomings. Children tended to draw larger figures with more preferred colors in the absence of a stimulus-drawing template (Burkitt, Tala, & Low, 2007). Children also drew more spontaneously when asked to depict a positively characterized topic versus a standard directive. Regardless, work with depression was evident with this assessment (Silver, 1988, 1993). At least three studies claimed a connection between depression and trends in themes with this test (Silver, 1996). Silver also formed the Draw-A-Story Test (DAS, 1988, 1993, 2002). Its primary purpose is to screen for depression through emotional content (Brooke, 2004). It demonstrated good overall inter-rater reliability for assessing depression in children and adolescents ($r = .81$), but not adults (Silver, 2002).

Levick (2001) constructed the Levick Emotional and Cognitive Art Therapy Assessment (LECATA) to measure normal cognitive and emotional development within children. Little research documents the validity of this instrument. However, this assessment labeled depression through absence of color. There are five drawing tasks, and unless these criteria were present in
all five, the client would likely not be depressed (M. Levick, personal communication, May 21, 2009).

Hacking and Foreman (2001) developed a coding system for the assessment of depression and other conditions in psychiatric art in an effort to quantify the data. Seven individuals with brief training in art psychotherapy rated color, color intensity, line quality, space, and subjectively judged emotional tone. The inter-rater reliability was satisfactorily high ($r > .91$). Four groups of diagnostic category included depression, psychosis, brain injury, and drug abuse. The diagnostic group differed on 4 of 13 variables: yellow, orange, color intensity, and line quality ($p < .02$).

The Art Therapy-Projective Imagery Assessment (AT-PIA), which is a series of projective drawing tasks collected under standardized conditions, is also a relatively new instrument (Bernier, 2004). Though the manual is in its infancy, many students and professionals of art therapy are currently working to expand research with this instrument. The AT-PIA is supposed to coincide with psychological instruments through personality and diagnostic indicators in artwork, so inter-rater studies are pending. Despite this, the creator cautioned against using a single sign in the assessment of depression. To date, art therapists report the ability to assess depression through the projective process of the client and some elements of artwork, though specific work in diagnostic indicators is also pending (M. Bernier, personal communication, May 28, 2009).

Aside from deriving research from clients, it is sometimes helpful to look to archival information on well-known artists to gain insight into the connection between art and depression (Jamison, 1993; Keshavan, 2006). A simple rating system conducted Keshavan’s study of 24 raters (12 psychiatrists, 12 laypersons) examined randomly selected paintings of four major artists with bipolar disorder. Psychiatrists reliably identified paintings before and after bipolar onset for individual ($F = 31.42, df = 1, p = .001$) as well as sets of paintings ($F = 17.34, df = 1, p < .001$), whereas the laypersons could only make these distinctions from sets of paintings ($F = 4.32, df = 1, p < .05$). This relatively small study concluded that additional work was required in this realm.

**Test Item Pool Sources**

Every valid and reliable test requires sources to support its individual items (Groth-Marnat, 2003). This section presents important considerations of test development. A link
between verbal and visual items follows. Last, the actual pool of items used within the instrument, complete with sources, support their inclusion into the FSA-R instrument.

**Test Construction Considerations**

The 12 considerations for effective test development. Downing and Haladyna (2006) proposed 12 main considerations to follow in the construction of a truly valid test. These are present within the framework of Fishman and Galguera (2003). The methodology section of this dissertation explains the application of these models. They are: (a) overall plan, (b) content definition, (c) test specifications, (d) item development, (e) test design and assembly, (f) test production, (g) test administration, (h) scoring test responses, (i) passing scores, (j) reposting of test results, (k) item banking, and (l) test technical report. Of these, only the first eight steps are relevant for this work, as they end with pilot testing. Future work includes the remaining steps. These steps are important because they adhere to the testing standards of the American Educational Research Association (AERA), the American Psychiatric Association, and the National Council on Measurement in Education (NCME) (AERA et al., 1999). These standards are in Appendix E.

The first consideration through this model includes the overall plan of test construction (Mislevy, Steinberg, & Almond, 2003). The desired outcome, test format, sources of validity evidence, purpose, desired inferences, psychometric model, timeliness, security, and quality control are all components of this step. The second consideration is an important part of this model.

The second consideration includes content definition (Kane, Crooks, & Cohen, 1999). This involves ruling out alternative explanations of test answers so that the core justification of the test remains. To address this, the examiner needs to pay close attention to the target domain that the test measures, and ensure that the sample derives from the target population.

The third consideration involves test specifications (Raymond, 2002). Test specifications should list behaviors and topics a test purports to measure, and outline their possible subfields. Format, difficulty, and features of test stimuli are generally included here (Osterlind, 1998).

Welch and Hoover (1993) offered advice on the fourth consideration of item development. Content, logistical, and technical considerations for test item development are crucial. The ideal item writer is one who understands the intended audience, and tailors the
difficulty level to them. The items require professional review from colleagues. Content that is
followed closely and grouped tightly produces valid items.

Guidelines for the fifth consideration include the actual test development (Roid &
Haladyna, 1982). Test development requires planning, research, psychometric considerations,
and pragmatics. Computerized data results of the pilot are crucial for the planning of a test.
Theory and rationale should underpin the test through research. The test requires a clear
definition of the psychometric scale used. The pragmatics includes the consideration of which
audience should undergo testing. An actual pilot test of the instrument follows.

Linking verbal test items to visual test items. The foremost consideration of constructing
a visual test is to equate verbal items comprising a test to visual items. Both share similar
qualities. The definitions and properties of a test item, along with the graphic equivalent and
sources, are in Appendix A.

Osterlind (1990) offered a straightforward definition of a test item; it is an examination of
mental attributes. When constructing a test, there are two theories from which to work from
(Downing & Haladyna, 2006). One is Classical Measurement theory; the other is Item Response
theory. As the FSA-R does not measure ability, but rather constructs of mental traits, Classical
Measurement theory is the theory of choice. Most existing art-based assessment instruments
measure mental constructs (Gantt & Tabone, 1998).

A test item should elicit a response from an examinee from which an assumed trait
becomes a factor. A stimulus drawing clearly yields a stimulus and response from the examinee
(Silver, 2002). In personal and informal research with the Face Stimulus Assessment (FSA), the
stimulus face format prompted the client to draw within the confines of the template; similar
phenomena occurred within the MARI test (Cox, 2002).

An item should exhibit discriminability, difficulty, and consistency (Fishman &
Galguera, 2003). From past research, most formal elements within a drawing were translatable
into quantitative data (Mattson, 2009). With this quantitative data, item discriminability results
from calculating Ferguson’s delta, while difficulty and consistency results from Pearson- \( r \)
correlation coefficients.

In addition, a valid item requires individualized construction (Fishman & Galguera,
2003). Such an item comes together through literature review, expert statements, and archival
information. The FSA-R used an extensive literature review, statements of other test constructors, and archival information on the link between depression and graphic indicators.

Items typically exhibit one true answer (Kehoe, 1995). From a variety of possible answers stems one objective choice. To address this in visual format, the test developer takes the average of the control group graphic element patterns and uses them as the criterion for the experimental group; this is also known as the criterion-keyed format (Kline, 1987).

Osterlind (1998) stated that items within a test should fit into a certain format. These range from multiple-choice, true/false, matching, interpretive, supplied response, essay, or performance. The format of the FSA-R most likely fits the supplied response category; it is similar in structure to that of the Bender-Gestalt test (Bender, 1938). Osterlind also stated that an item contained at least some of the following: directions, stem, text, distractors, graphics, and one correct answer. A stimulus template becomes a referent to elicit a response from the examinee. This also partially serves the role of the test directions.

Items should clearly communicate intent to the responder (Irvine & Kyllonen, 2002). This enables the test-taker to make a decision about the possible answer. The graphic remedy for this is a clearly designated stimulus template. Additionally, the test item should rely on a mental model as a base.

An item changes through trials (Groth-Marnat, 2003). After a pilot study, it is subject to deletion, addition, or alteration according to the responses. Just as the constructor builds the items one-at-a-time, they are alterable in a similar fashion (Fishman & Galguera, 2003).

A test item must have logical or verbal syntax and grammar. Visual figures exhibit 36 identifiable geons, similar to 44 phonemes and 26 letters of the alphabet (Biederman, 1987). Coloremes are graphic equivalents to alphabetic letters (Saint-Martin, 1990). Letters are actually lines and angles; math symbols are like graphics on tests (Osterlind, 1998). An existing test, known as the Color Trails Test (CTT), substituted colors for letters (D’Elia, Satz, Uchiyama, & White, 1996). In short, visual items can be similar to written items.

Sources of Item Content

Appendix B lists the FEATS properties and corresponding DSM-IV-TR depression symptoms as they relate to item choice for the FSA-R. This shows the dimensions that the items of the FSA-R intend on measuring. Usually, if an item number exhibits a high response rate, it is
a common knowledge question, in which case a supporting reference is not necessary (D. M. Clark, personal communication, August 3, 2009).

Dark colors/lack of colors is synonymous with color percentage measured by this study, while constricted use of space is the same as free space. The other dimensions of sketchy/hesitant line and lack of detail are for future studies, but they are included in the literature search because, (a) they comprise the complete instrument, and (b) they illustrate dimensions that PDIAS can measure. The preliminary version of the FSA-R instrument, with the original item numbers corresponding to references within this text, is in Appendix F. These item numbers are in parentheses near each dimension signifying their relation to that particular dimension.

Dark colors/lack of colors (items 5, 6, 25, 26, and 33). In one study, patients with depression tended to use darker colors. These individuals generally were not productive in art making. They often did not complete basic drawing tasks, and used few and dark colors (Anastasi & Foley, 1940).

Dark and somber colors in the paintings of depressed individuals are prevalent (Dax, 1953). Conversely, those presenting with mania often used vibrant and numerous colors. This is a testimony to the colors used in accordance with emotional states. Some researchers attempted to construct tests of color and personality, based on such observations.

Scott (1969) translated the Lüscher Color Test manual to English, and included chapters on color interpretation. In general, well-adjusted individuals judiciously used basic primary colors. Neutral or dark colors, such as brown and black, are hallmark choices of those experiencing sensory deprivation or denial of available colors.

Some researchers studied the colors within the drawings of young individuals with depression (Leo, 1973). Many well-adjusted individuals brightened human figure drawings by adding radiant sun figures. It was therefore unusual to note darkened colors of the sun and clouds (items 1 and 5). Such features are indicative of depression.

Wadeson (1980) further reported on this dichotomy of color use in those with mood disorders. She noted darker and duller colors in artistic productions of those with depression. Overall, she noted less color use with her clients who were clinically depressed. She cautioned that a debate occurred in art therapy over whether a depressed individual drew less color or darker color than an individual without depression did.
Around the 1980s, pharmacology drew the attention of art therapists. The longitudinal effects of pharmacotherapy on depression patients in relation to color associations on the LCT were examined (Demeter, Rihmer, & Frecska, 1985). Results indicated that this method was suitable for tracking changes in symptoms. This also gave credence to the claim that a definite interplay existed between depression and the general associations of color.

Oster and Gould (1987) informed their readers that color usage was often subjective. Therefore, they stressed contexts of color usage. However, for the most part, they maintained that the continued use of dark colors signaled depression.

Many individuals with depression used neutral colors or few colors in paintings (Simon, 1992). Furthermore, their paintings took on a muddled impressionistic style. The paintings studied often reflected the depressed mood of the artist. The coloring-outside-the-line phenomena noted in individuals with depression is a basis for including the stimulus templates within the FSA-R. Having a definite boundary to detect such blurring is therefore invaluable to assessment.

Some researchers observed the actual drawings of students, and discovered through interview that these drawings were closely associated with current emotional states (Kaplan, 1994). Of these, depression and its manifestations became evident based on color choice, which they claimed was predominantly red and black. The students completed a graphic form of the Rosebush gestalt therapeutic exercise (Oaklander, 1978), the rationale for its inclusion into the FSA-R as item 26.

Nolan, Dai, and Stanley (1995) studied a group of undergraduate students ($n = 72$). A portion of this sample presented with depression. Using the BDI, they discovered that those scoring above a 10 preferred the colors black or brown.

Twenty drawings of those with Major Depression from the DDS tree directive underwent examination (Morris, 1995). Rating commenced with the Tree Rating Scale (TRS, Creekmore, 1989). This researcher discovered that 65% of those with Major Depression studied used only two to three colors.

Gantt and Tabone (1998) agreed that color and affect correlated. They stated that groups of individuals with depression would use less color in drawings. The colors used tended to be monotone or dark.
Depression affects the areas in the brain that manage color perception, and this might play a role in color choice (Barrick, Taylor, & Correa, 2002). Researchers used a cross-sectional sample of 120 individuals and an unspecified self-report scale to correlate with the depression. From this, they were able to pinpoint which areas of the brain were responsible for determining color choice, and that those with depression perceived color differently from norms.

Through the Lüscher Color Test (LCT), Naoto, Takeshi, and Shigetō (2004) compared 100 freshmen and 21 students diagnosed with Major Depression and found that a significant difference arose between depression and the control group in the seventh and eighth position of color preference within the LCT scale. The students diagnosed with depression generally avoided bright colors.

Of 120 freshmen college students sampled, one-third of which presented with some form of depression, a correlation resulted between color and depression (Wu, Chang, & Chen, 2009). The students completed mosaic drawings. They used dark colors to express their depression.

Constricted space (items 2, 3, 4, 10, 14, 20, 31, 32, 34, and 35). Several researchers explained the phenomena of Horror vacui. Prinzhorn was one of the first clinicians to note this effect. It involved filling large areas of paper on behalf of psychiatric clients (Anastasi & Foley, 1941; Prinzhorn, 1972).

Anastasi and Foley (1940) told of historic accounts of the artwork of those with Manic-Depressive symptoms where clear mood transitions occurred. The art would shift from Horror vacui in mania (filling all available space) to relatively empty free space in depression. This became especially true of paintings.

Later in history, intense interest developed around projective drawings of the house, tree, and person. Tree drawing space usage derived from a multitude of subjects, and categorical statements about the link between depression and tree size resulted (Koch, 1952). When the tree figure was very small, leaving much free space, it was nearly always indicative of depression, according to some researchers.

Small figure size was often a result of lowered productivity in art making. Rapaport, Gill, and Schafer (1946) explained reduced productivity in those with depression and artwork. This resulted from behavioral and scientific observation. While most non-depressed individuals exerted more time and effort on artwork, individuals with depression exerted notably less of both these variables.
Concerning figure size, patients diagnosed with depression by a physician in a clinical setting generally drew smaller human figures than those without depression (Lewinshohn, 1964). The figure drawings of 50 patients with depression were significantly shorter than 50 non-depressed patients at an alpha level of .01 ($F = 8.16, df = 1, 92$). Female patients with depression drew smaller figures than females without depression (Roback & Webersinn, 1966).

Smaller figures correlated with depression (Deabler, 1969). An experimental study supported this claim. The test group received the H-T-P as the selected instrument for figure drawing. Figure drawing sizes fell into the category of constricted space in that the negative space left from the drawing size revealed mostly unused and hence constricted space on behalf the depressed individual.

Wadeson (1971) also recorded the content of numerous drawings from patients with depression. She discovered significant amounts of free space and emptiness. This was true for both thematic and graphic components of these drawings.

Johnson (1973) studied various drawn forms from the Bender Gestalt for validity of constriction as an indicator of depression. Prior to this study, few articles featured the hypothesis that constriction and depression related. In the comparison study, two matched groups ($n = 25, n = 25$) differed significantly in D scores ($p < .05$) for constriction, supporting the hypothesis that those with depression drew in a constricted manner.

The KFD Family Drawing Depression Scale (FDDS), paired with depression patients, contrasted with controls ($n = 71$). In conjunction with the Zune Self-Rating Scale for Depression (SDS), most items in question, including size of figures and space usage, varied significantly between depression and control subjects (Wright & McIntyre, 1982). Furthermore, certain symbols repeated themselves within the artwork of those with depression.

Those with depression used less space than those without (Dawson, 1984). This finding resulted from issuing numerous drawings to patients with depression ($n = 50$). These drawings paired with BDI scores for validity outcomes. This antithesis of Horror vacui tends to be a phenomena noted by many researchers.

Hundreds of drawings of those with depression stemming from sexual abuse underwent analysis (Wohl & Kaufman, 1985). For these researchers, forms could be indicators of an individual’s awareness of the outside world. Constriction therefore, was a reflection of low awareness. The impoverished use of form then suggested depression and low energy.
Gantt and Tabone (1998) concluded that constricted use of space related to depression and a general lack of energy. There was usually little to no drawn backgrounds within their PPAT drawings for those with depression. Figures tended to be smaller in the artwork of those with depression, leaving significant amounts of free space.

Cavézian et al. (2007) examined hemineglect and visual-spatial abilities between depressed patients and controls. Using the manual line bisection task (Halligan & Marshall, 1988), they discovered that those with depression significantly differed from the controls in this task with the line cued on the right side, at an alpha level of .05, $F(1, 27) = 7.96$, $p = .009$. This demonstrated some impairment in thinking, and this performance resembled that of schizophrenic individuals performing the same task.

Sketchy/hesitant line (items 1, 7, 8, 9, 11, 12, 13, 15, 16, 20, and 23). Individuals with diagnostic criteria for depression often drew disconnected or sketchy lines while conducting the Bender-Gestalt test (Hutt & Briskin, 1960). This occurred during the copying phase of the card task. The lines also resembled faint likenesses to the stimulus figures.

Individuals with depression likewise made their lines sketchy and hesitant in Bender-Gestalt responses in yet another study (Clawson, 1962). The response of drawn line resulted from the Bender-Gestalt spectrum. From this, individuals with depression made disconnected lines.

Unusually shaped lines occurred within the art productions of individuals with depression (Zimmerman & Garfinkle, 1963). These often resembled swirls or looped lines. This phenomenon was particularly prominent in bipolar patients in the manic stage of depression. Burns and Kaufman (1970) collected and analyzed numerous samples of KFD’s. Of these, individuals with depression displayed sketchy qualities of line within their family drawings. Figures took on a shaky appearance.

White and McGraw (1975) attempted to differentiate depression from organic brain disorders. With a sample of males ($n = 24$) and females ($n = 24$), they analyzed Bender Gestalt figures one and two. Pairing these results with MMPI scores, they did not find a significant difference between slanting figures and depression—which resulted from hesitant line—but pressed on to suggest that this occurred in a miniscule amount of cases. This study held potential to be valid, but required further research.
Cronin and Werblowsky (1979) discovered that individuals with diagnostic criteria often drew disconnected or sketchy lines. This tended to occur in spontaneous art production. Wadeson (1980) later confirmed this finding.

Some studies included comorbid conditions. For instance, those with depression and Dementia Alzheimer’s Type (DAT) tended to draw line that was hesitant and uncontrolled (Katzmann, 1987). This observation coincided with a study by Reisberg, Ferris, and Franssen (1985).

Researchers examined the drawing patterns of 15 students, and devised a series of rules for their kinematics actions (Thomassen et al., 1991). These include starting at the leftmost areas, threading, anchoring, vertical segment initiation, starting at topmost area, drawing from top to bottom, drawing parallel lines in equal length, and drawing horizontal lines from left to right. This was useful for the overall layout of the FSA-R template. With these rules in mind, the drawer generally starts at the top and leftmost part of templates, so, they are present in this fashion within the FSA-R.

Sabbe et al. (1996) analyzed standardized drawing and writing tasks of 22 individuals with depression and 22 control individuals through digital technology. The SDS confirmed the depression of the test individuals. They noted more errors in those with depression on the drawing tasks. Furthermore, they theorized that reaction time (RT) of visual-spatial drawing correlated with errors in these individuals. In other words, the more time and referencing it took to complete a drawing task, the more the errors. In Major Depression subjects, the simple movement of a very small line often produced errors.

Gantt and Tabone (1998) noted that those presenting with depression often drew sketchy or hesitant lines within the PPAT. In contrast to this, Urban (1963) considered drawings with even pressure and consistency as normal. This was a qualitative judgment.

Sabbe et al. (1999) used a computerized tablet and digital pen to analyze 10 drawing tasks of patients with depression and controls. Overall, those with depression performed the drawing tasks more slowly, particularly with the reproduction of 30-millimeter lines. Additionally, those with depression made notable errors in the reproduction of these lines. There appears to be a strong correlation between slow performance time and mistakes made in the drawing of lines—both of which improved with treatments of Prozac.
Tucha and Lange (2001) analyzed the handwriting and line movement of those with Major Depression and those without, and found distinctions. This occurred in both adult and child samples. The execution of handwriting movements revealed differentiations in kinematics, and consequently—line quality.

Hegerl et al. (2004) analyzed 16 inpatients presenting with depression and receiving Citalopram (30–60 milligrams/day) and 12 acutely depressed patients treated with Reboxetine (4–8 milligrams/day) drawing circles using a digitizing tablet for tracking movement dynamics. They discovered more controlled movement in drawing because of this pharmacological treatment, implying that depression affected the quality of line. A similar study entailed the issuance of Prozac to depressed patients with eventual improvement in line and figure copying tasks (Schrijvers, Maas, & Sabbe, 2009).

Pier et al. (2004a) researched psychomotor retardation (PR) in Major Depressive Disorder (MDD), and concluded that twenty years of literature revealed valid connections between PR and MDD. It is a significant correlation and symptom of MDD and it is evident in latency and RT in copy drawing tasks. This particular article found evidence for PR and MDD, but not dysthymia.

Pier et al. (2004b) produced a similar article asserting that PR was an important symptom of MDD. Earlier research also considered the measurement of PR as an aid to diagnosis (Greden & Carroll, 1981). It is measurable, holds prognostic and diagnostic value, and can even differentiate subtypes of MDD. These researchers understood that antidepressant medications affected PR, and assigned subjects that were titrated off these medications for more accurate results of studies. Using an array of figure copying tasks, they determined that unmedicated MDD patients \( n = 32 \) performed poorer than matched controls in drawing lines.

Depression and PR link together based on large amounts of research (Mergl et al., 2007). From this assumption, researchers analyzed the handwriting of 37 patients with depression and 37 controls to arrive at significantly less velocity in the handwriting marks of the former \( p = .001 \). From irregular and uncontrolled peaks in the lines of writing, they were able to determine which areas of the brain depression affected. This could translate to drawing, as the mechanism appears to be the same.

Lack of details (items 17, 18, 19, 20, 21, 22, 24, 27, 28, 29, 30, and 33). An analysis commenced with the paintings of patients with depression (Dax, 1953). Hundreds of patients
participated in the survey. The researcher concluded that their paintings simply lacked detail in contrast with normal subjects.

The drawings of individuals with depression featured lack of detail (Hammer, 1958). These individuals generally did not complete drawing tasks, and often left artwork unfinished. This was true even when the drawing did not require an appreciable amount of energy to complete. Wadeson (1971) sampled pictures of individuals at the peak of their depression ($n = 10$), with non-depressed patients serving as controls. Within their artwork, she observed less expenditure of energy on the final product, leading to incomplete drawings and an overall lack of detail. The paucity within the image adequately reflected the depressed state of the individual.

Wadeson (1971, 1975) explained the findings of numerous free drawings of those with depression. She discovered that an overwhelming amount of these drawings lacked detail. Individuals with depression expended little effort in their drawings, and thematic components often included hopelessness and isolation.

Dawson (1984) investigated differences between the drawings of depressed and non-depressed adults. The study hypothesized that the drawings of those scoring on various spectrums of the BDI would omit more details than those of the control participants. The depressed group excluded more details than the non-depressed group. The difference between the group means differed, but not significantly.

A lack of details resulted from those presenting with depression. This observation resulted from hundreds of drawings of those with depression stemming from sexual abuse (Wohl & Kaufman, 1985). Furthermore, a theory that items are indicators of an individual’s awareness of the outside world underpinned this research.

Through longitudinal research, Simon (1992) noted that paintings by those with depression tended to blur together, thus obscuring details and boundaries that the artist with depression wished to convey. The resulting work resembled an impressionistic style of painting; the author also referred to this as an *archaic transition* within a model of art therapy.

Significant improvements in the score of the Clock Drawing Test (CDT) occurred in those with dementia treated for depression (Lee & Lawlor, 1995). A pre- and post-treatment of depression transpired. This resulted in improved details, implying that this population with depression would draw the clock with less detail.
Okada, Takahashi, and Tokumitsu (1996) examined 36 patients with Major Depression diagnosed by DSM-III-R to find changes in brain oxygenation while conducting the mirror-drawing task (MDT). Results showed that the non-dominant hemisphere of the individual before depression might become dominant while in the midst of depression. This lead to the leftward bias of FSA-R test items; the individual with depression may have difficulty with this.

Within drawings of those with depression, an overt lack of detail is often evident (Malchiodi, 1998). Additionally, the figures within the drawings tend to be small, and the line quality featured light pressure. This gave the drawings a faint appearance.

Souza, Cendon, Cavalhero, Jardim, and Bogossian (2003) found that those with depression and Cardiopulmonary Disease (COPD) drew incomplete figures in the Human Figure Drawing Test. Common omissions included face details, facial hair, clothing, ground line, legs, trunk, hands, feet, and transparency (items 12, 14, 17, 19, 20, 23, 24, 27, 29, 30, and 33). Overall, the drawn human figures appeared incomplete.

People with depression often produced drawings with fewer details than those without depression (Kaplan, 2003). The artwork of those who exhibited depression usually appeared constrained, especially when compared to the work of those with mania. This constricted space is similar to free space.

Clinicians that used human figure drawing tests generally employed the sign method of determining pathology within the individual (Lilienfeld, Wood, & Garb, 2004). A fitting example included the omission of face details in the drawing as a sign of depression. This approach requires more research to solidify connections; the incremental validity that it contained outweighed the usefulness of such projective techniques.

Zomet, Amiaz, Grunhaus, and Polat (2008) analyzed results of a filling-in task of those with Major Depression \((n = 27)\). This is a concept similar to gestalt completion, of which many of the FSA-R test items are based upon. These researchers found that nearly all of the subjects were unable to fill in the test figures, while the controls \((n = 32)\) performed significantly better at this task.

### Justification for Main Instruments Used

**Face Stimulus Assessment (FSA) Platform**

Perception of the face is most significant to humans when presented in vertical format (Alley, 1998; Golomb, 1992). Infants use preexisting knowledge of face recognition that enables
them to identify familiar features of the face upon birth. This innate structural information steers face preference (Morton & Johnson, 1991). Difficulty in representing face image may be a hallmark of those with depression (Gantt & Tabone, 1998).

The human face is a proven source of information and emotional expression (Cacioppo & Petty, 1983; Kemp, Hopkinson, Gordon, Williams, & Clark, 2007). This suggests some cognitive interplay between depression and features within the face, reinforcing the importance of face stimulus. Individuals with depression performed significantly worse than controls in face emotion perception (Langenecker, Bieliauskas, Rapport, Zubieta, Wilde, & Berent, 2005). Recent findings indicated that individuals diagnosed with Major Depression exhibited bias towards the presentation of sad faces (Jukka, 2006). They chose sad faces more often than happy faces. Additionally, individuals with depression mentally processed sad images of faces quicker than happy faces (Moretti, Charlton, & Taylor, 2002). Furthermore, this particular finding suggested the development of face assessment tools for depression.

Recent work in neurological imaging has also shown that children with bipolar disorder encode information about the face differently from normal subjects (Dickstein et al., 2006). There was a significant increase in activity in the striatum of the brain when happy and angry faces presented to the bipolar disorder group compared to the normal group. One implication for this study is that child bipolar subjects may depict the face in a manner different from a normal group, which could become an important consideration for diagnostic studies using the face. There is evidence to say that face recognition is innate to newborns (Morton & Johnson, 1991).

The person is among the first recognizable objects drawn by children (Lowenfeld & Brittain, 1970). The face is usually part of the drawn person. Lowenfeld contended that no artistic expression occurred without some component of self-identification within an art piece. A drawing of a face may therefore reflect components of the artist. At the age of three, children can distinguish between race, species, and age, and develop narrowing profiles of familiar faces carrying well into adulthood (Kuefner, Picozzi, Macchi, & Bricolo, 2006). Children begin imitating facial responses of adults at an early age to form a repertoire of their own facial expressions (Meltzoff & Moore, 2000).

Self-portraiture may ultimately be a means of analyzing the self. Throughout history, many artists have painted self-portraits as reflections of themselves. Van Gogh and Da Vinci are examples, though other well-known artists did the same. Self-portraits are especially significant
for those with depression (Alter, 2007). With self-portraiture, the artist can step back from painful emotions and present them in vicarious form. Therefore, the face may be a significant form of self-portraiture.

The outline of the FSA face incorporated multicultural consideration (Betts, 2003). This is significant because the increasing diversity within society challenges the assessment field to construct ways to treat individuals of varied backgrounds with fairness (Prediger, 1994). Multicultural considerations are crucial to current assessment research (T. Fabian, personal communication, January 18, 2006). Additionally, the use of the face worked well with ethnic children and adolescents presenting with multiple disabilities, and design of the FSA outline resulted in response to these populations (Betts, 2001). The face is universal. Face recognition is innate to newborns (Morton & Johnson, 1991). Every individual has likely seen, or recognized, a face on some level. If individuals did not, they would likely experience failure to thrive (Zenel, 1997). The ability to perceive faces is universal.

Because of this universality, the face may be an effective referent. In addition, the outlines of the items in the template afford a means of consistency and standardization; this rationale follows that of Kellog (1978), in reference to the MARI test. The examinee drawing within the confines of a face stimulus template may require little direction to proceed (e.g., “I know what eye colors exist, so I will use the color green, blue, brown; I know what color skin is, so I will use pink, tan, brown…”). To convert the FSA into the FSA-R test required definite stimuli to foster constructed responses from examinees (Osterlind, 1990). In the case of the FSA-R, this is the outline of the face with subsequent stimulus templates. It appears that visual test items developed on the FSA-R equate to standard test items, based on criteria (Osterlind, 1998). The section within this dissertation on test item considerations further explains this link.

Beck Depression Inventory (BDI-II)

Groth-Marnat (2003) stated that over 1,000 research studies outlined the use and development of this inventory. Most researchers included the Hamilton Psychiatric Rating Scale for Depression (HRSD) in conjunction with their depression drawing tasks, though this is not a self-report instrument, like the BDI-II. As the aim of the FSA-R research is to discover answers through objective means, it seemed best not to include any more raters than necessary, hence the importance of the BDI-II, where the participant of the study would report symptoms. Aside from this, the BDI-II scores correlated highly with the HRSD ($r = .71$), and it is a valid and reliable
instrument for screening depression. From meta-analyses, internal consistency ranges from .89 to .94 (Beck et al., 1996).

Other advantages of this instrument include accessibility, ease-of-administration, short time limit, and reputation. The BDI-II is a level “C” test, meaning that persons with marginal qualifications may use this instrument. This was important for norming the FSA-R, as many supervisors of the FSA-R pilot test administrations hailed from varied professions, and it was possible that some would not be qualified to administer higher-level tests. Additionally, it is cost-effective; it is available for purchase with the manual, and additional forms within a packet. This instrument is face valid, and presentable within a single double-sided page of questions. Concerning time, it generally only takes 5 to 10 minutes to complete. The BDI-II is one of the most common and effective screening devices for depression in existence; it has been around for over 40 years. It is also backed by thousands of publications confirming appreciable levels of reliability and validity (Groth-Marnat, 2003).

Another reason for choosing this instrument is that Harcourt’s clinical measuring consultant stated that statistically comparing one projective test to another is useless (J. C. Hanson, personal communication, April 4, 2007). The FSA-R required a valid parallel test for depression. The Beck Depression Inventory became a clear choice because it measures stable traits and demonstrates both high reliability and validity (Beck et al., 1996).

**Formal Elements Art Therapy Scale (FEATS)**

Gantt and Tabone (1998) developed the FEATS for use with the PPAT. Much of the research fueling the development of the FSA-R stemmed from the graphic equivalents of symptoms present within the manual of the PPAT (Gantt & Tabone, 1998). The developers of this instrument linked Major Depression with DSM-IV-TR symptoms and corresponding observations from art therapy that manifested the formal elements. Most of these formal elements are testable through the FSA-R computerized rating system.

The goals of the FSA-R are similar to those of the PPAT and FEATS. Namely, the establishment of increased reliability and validity measures within art-based assessment that will possibly contribute to higher credibility within the assessment area of art therapy. The methods are also similar. Both instruments seek to distinguish DSM-IV-TR diagnostic categories through quantitative data and statistical methods.
Additionally, the structure of the FSA-R is similar to that of the PPAT. Both feature scoring manuals, descriptions of norms, and detailed outlines of theoretical underpinnings. Standardization measures for the FSA-R follow those of the PPAT; these include the populations, testing environment, directive, and materials.

Public Domain Image Analysis Software (PDIAS). The future of psychological testing depends on the trends of computer technology (Groth-Marnat, 2003). Furthermore, computerized assessment will offer faster scoring, fewer errors, complex decision rules, and a higher level of overall accuracy; all of these factors lead to increased test validity. Two PDIAS programs rated formal elements in this study. They include Image Java (ImageJ) and Measures of Vegetative Health (MVHimage).

The primary PDIAS program used for this study was ImageJ v.1.43, which is a highly accurate image analysis tool developed by the National Institutes of Health (NIH) for biotechnological imaging (Collins, 2007). Existing validity of the rating tools of ImageJ stems from testing the software prior to its release (Sonka & Hanson, 2001). In addition, by measuring the scaling features in the digital image, calibration results. Calibration is the method used to generate reliability and validity within ImageJ when tested with other measures.

The second program used is MVHimage v.8. For MVHimage, the software developer issued alpha testing of the software prior to public release (J. Pickle, personal communication, October 27, 2008). This served the function of detecting early errors and establishing some measure of reliability and validity (Kaner, Falk, & Nguyen, 1999).

**Philosophical and Theoretical Underpinning**

**Bell’s Philosophy**

Bell (1958) stated that personal experience of an emotion was the starting point for aesthetic response, and this response is particularly compelling for the recipient. Bell classified all emotions felt from an aesthetic response as of the same kind. They are reactions to all things visual. The emotion Bell spoke of was what he called the *aesthetic emotion*, and this was a necessary condition for experiencing aesthetic. Bell stated that if we could find a common quality across all of the visual objects that elicited aesthetic emotion, then we would essentially solve the central problem of aesthetics. This was the key to solving the mystery of aesthetic response. It could effectively distinguish works of art from all other objects.
Only a single formal quality is necessary for Bell’s aesthetic experience (Eldridge, 2003). The sign method that the FSA-R employs parallels this notion. The quality shared by all works that provokes aesthetic emotion is the *significant form*. Combinations of line, color, shape, and relations of these constitute significant form. In addition, boundaries of contrasting qualities within space are essential to separate form. Significant form is a universal quality; all aesthetic objects in question share it. Even the most likely opponent of Bell’s philosophy, George Dickie, stated that he did not doubt the existence of universals in his own theory, known as institutional theory (G. Dickie, personal communication, December 24, 2008).

**Test Theory**

Downing and Haladyna (2006) purported the existence of two basic theoretical approaches to test construction: *Classical Measurement theory* and *Item Response theory*. Clinical measurement theory is the clear choice for the development of the FSA-R because of the nature of the examined phenomena. Clinical measurement theory assumes that each person has a true score (T) that, with the addition of error (E), results in an observed score (Allen & Yen, 2002). This is the most influential theory used in the social sciences. Whereas item response theory is well suited for criterion-based testing, clinical measurement theory applies to norm-referenced instruments, such as the FSA-R. Incorporating theory may address test construction bias complicating the clinical measurement equation. Bias in test construction is reducible through several guidelines. These include, (a) incorporating an evaluative panel, (b) using the standard error of measure (SEM), (c) use of the Item Response theory or Classical Measurement theory, and (d) considering that if one race scores significantly higher than another, it may indicate a bias within the demographics (J. C. Hanson, personal communication, April 15, 2007).

The hierarchical analysis of examinee testing is crucial to efficient test development (Shrock & Coscarelli, 1989). The most popular hierarchical framework to use is Bloom’s Taxonomy (Bloom, 1956; Schunk, 2007). This describes a learner’s cognitive process, and each stage meets a prerequisite stage to move forward within the levels. The levels include evaluation, synthesis, analysis, application, comprehension, and knowledge. Each format of test involves some of these steps (e.g. multiple-choice, true/false/matching), but only one incorporates all—the essay.

The visual test also includes all the levels of learning and response, since this format of testing most closely resembles the essay test in that it contains a *stem* and an open response.
allotment. One might guess that including all of Bloom’s taxonomy results in a better test format, but drawbacks will likely exist for a visual test, just as they do for the essay. Disadvantages may include difficulty in reliability of scoring, scorer fatigue, image vagueness, long test-taking time, and long scoring time. However, many of these disadvantages of the essay test described by Shrock and Coscarelli (1989) are fixable through an objective scoring system, such as that used in conjunction with the FSA-R (Mattson, 2009).

**Questioned Documents Theory**

All handwriting begins as artistic renditions of letters and words (Norwitch, 2009). Art making and writing overlap each other (Vygotsky, 1978). As a result, some of the tenets of questioned documents theory (QDT) apply to making art. There are many such tenets to QDT, which differ from graphology (Osborne, 1978). Unlike graphology, QDT is both rooted in hard science and accepted within evidential forensics, making this a sound theoretical scaffold.

The main theories of QDT hold that, (a) given a large enough of a sample, no two individuals write the same, and (b) no single individual writes anything the same twice (F. Norwitch, personal communication, July 25, 2009). However, while much variation exists within writers, some components stay the same. These components are class characteristics (Smith, 1984).

Given a large societal pool of individuals schooled by a similar social system, these class characteristics will be evident in the style of the writing. For instance, the school system in the United States primarily teaches the Zaner-Bloser system of handwriting (Norwitch, 2009). With this, similarities in writing arise, though with the passage of time, each style of handwriting becomes increasingly distinct. Regardless, this shared collective style of teaching and learning makes indelible base styles of both writing and art making. The development of art also stems from known developmental levels of such researchers as Viktor Lowenfeld, fostered by standard school curriculum. Though currently, the field of art education is beginning to institute the Art for Life model (D. Gussak, personal communication, September 9, 2009). The development and institution of standard art models is important because it leads to the next tenet—that of internal variation.

Internal variation refers to significant variables in a writer that contribute to a style vastly different from the writer’s original style. Classic examples include age, nervousness, medical conditions, and mental health conditions. For instance, an individual schooled in a collective
system may write similarly to his or her classmates up to a point; if this individual has a stroke, the difference in this style will likely be apparent.

Art making and writing are similar. Children begin making artistic renditions of letters so that drawing and writing are essentially the same process when we first begin writing (Bissex, 1980; Clay, 1975; Gardner, 1980; Kellog, 1970). Barrs (1988) noted that children told verbal stories in the form of drawings. Gesture translated to drawing scribbles in early writing (Vygotsky, 1978). A split emerges when the use of the line differs in meaning (Kellog, 1970). Because of this similarity, both share similar tenets. The shared tenets of interest include the concepts that, (a) everyone draws similarly with class characteristics, and (b) internal variation of mental illness affects this class characteristic.

The notion that everyone draws similarly could arise from our developmental levels of drawing. Each age group up to 12 draws within relatively predictable stages (Kellog, 1970; Lowenfeld & Brittain, 1987). The adult will likely stay within the realism stage, as few advance beyond that. From this, most normal adults will draw accurate depictions of color, detailed faces with differentiated gender, ground lines with sky, and basic perspective (Malchiodi, 2003). Gardner (1980) demonstrated that drawing in childhood stemmed from reflections of children’s native culture. Culture could be an example of a class characteristic of drawing, hence making the drawing style similar across a given body.

Studies in depressive psychomotor retardation evidence internal variation (Thomassen et al., 1991; Hegerl et al., 2004; Mergl et al, 2007; Sabbe et al., 1996; Pier et al., 2004a; Pier et al., 2004b; Sabbe et al., 1999). Psychomotor retardation drawings are clear examples of how mental illness affects standard drawing tasks. Non-patient groups often contrasted with these studies.

In conclusion, art making and writing are similar. Both share similar tenets. The notion that everyone draws similarly could arise from our researched developmental levels of drawing. Studies in depressive psychomotor retardation evidence internal variation. Therefore, those with Major Depression will likely draw differently from their peers.

Conclusion

In this chapter, several sections outlined the literature composing this dissertation. The first section explained the use of current computer technology. This section incorporated material most similar to this dissertation, especially components of both computer color and free space rating, and the basic terminology surrounding these formal elements.
The subsequent section gave the reader historical accounts of using artwork for assessment. The review revealed that early forerunners of art-based assessment merely described artwork instead of developing diagnostic measures from it. In addition, much of the work in art-based assessment sprung from the 20th century, and instruments often evolved from each other. In the 20th century, clinicians found means of detecting depression through artwork. In this period, assessors delineated mental illness categories from “insane art”, which led to accurate assessment of depression. Nonetheless, areas of norming, reliability, and validity require additional studies. The history included the most recent works on detecting depression through artwork. Here, advances in the literature primarily originated from early sources and instruments. This section highlighted the emergence of more advanced statistical and procedural techniques in assessing artwork.

The following sections explained the literature search underlining the individual test items with their respectable domains. The individual test domains were determined there. They consisted of lack of color/dark colors, constricted space, lack of detail, and sketchy/hesitant line. Within each domain is a listing of sources pertinent to each test item.

The last section outlined the theories and philosophy of this dissertation. This dissertation philosophy follows Bell’s universal forms, where a single significant form determined the aesthetic of an art object. This is important because it is analogous to the sign method of art assessment. Additionally, it segued into questioned documents theory, where two premises emerged, (a) everyone draws similarly with class characteristics, and (b) internal variation of mental illness affects these class characteristics. Classical Measurement theory from Test theory aptly suits this study.

This overall literature search contributes to the construction of the FSA-R. It is possible that those with Major Depression would draw differently from their normal peers. In order to test this claim, the actual administration of the instrument follows in the subsequent methodology chapter.
CHAPTER 3
METHODOLOGY

Construction within a Framework

A theoretical framework guided the process of both constructing and administering the FSA-R. In order to develop an instrument that is both valid and reliable, it is essential to follow such a framework in all stages of test development (McDonald, 1999). This chapter illustrates the steps to instrument design proposed by Fishman and Galguera (2003); the steps include the completed literature review, conceptualization of components, drafting of item pool, formatting, and pilot testing/initial administration. The schematic of instrument design is in Appendix C. Throughout these steps are the restated research questions, null hypothesis, research methodology used, research design, population and sample selection, instrumentation procedures, and approaches to data analysis.

The first stage was the literature review. The validity of this test hinged on a thorough review of literature. Sources included information from the substantive field of depression, computer rating, and graphic traits posed by those with Major Depression. Research revealed the componentiality of this test. The four main dimensions are dark colors/lack of colors, constricted space, sketchy/hesitant line, and lack of details. A literature review commenced through the Florida State University library system, Florida Agricultural and Mechanical University library, Leroy Collins Library, Interlibrary Loan, electronic journal searches, archival searches, and personal communications. The resulting sources determined and supported the items of the test.

This research supports varying aspects of the Major Depression traits measured. Mislevy et al. (2003) outlined considerations of this step through a model, which included the overall plan of test construction. The desired outcome, test format, sources of validity evidence, purpose, desired inferences, psychometric model, timeliness, security, and quality control are all components of this step. The desired outcome of the FSA-R is the completion of a pilot study. The test format is visual in nature. Sources of validity evidence stem from the literature review. The purpose of the FSA-R is to form a screening instrument for Major Depression. The psychometric model used is Classical Measurement theory. Timeliness included a 3-month deadline from inception. Security stemmed from issuing select administrators the testing manual. Quality control resulted from controlling a host of variables (Christensen, 2001). The examinees did not receive information on the rationale of the test prior to its administration. Control of
recording errors resulted from training the test administrators on issuing the FSA-R; they also received identical test manuals prior to testing.

The second stage was the conceptualization of components. Results from past issuances of the FSA-R informed its current version (Mattson, 2008). From raw data collected from groups of those with elements of depression, certain styles of the drawings by individuals with depression became apparent. This required sound judgment of the literature and methods of test construction. Furthermore, authors Kane et al. (1999) explained that this stage involved phasing out alternative explanations of test answers so that the core justification of the test remained. To address this, literature came from the target populations of Major Depression.

The third stage involved drafting the item pool. A larger item pool than needed is required to account for deletions. Kline (1987) recommended twice as many items as thought necessary for the actual test. Simple language in the instructions helps respondents understand the items. Grouping of similar items tests ipsativity. It is best to include as many items as possible for each dimension of the test. Raymond (2002) mentioned the third consideration of test specification, that of listing behaviors and topics that a test purports to measure. The FSA-R purports to measure the drawing response to graphic item stimuli from those with Major Depressive Disorder. From this, the subfields are DSM-IV-TR symptoms connected to these responses. It measures the percentage of color, space usage, amount of details in the form of shape descriptors, and line deviations from stimulus tasks for those with Major Depressive Disorder.

The fourth stage was formatting. The FSA-R is an alternative format test. The nature of the questions is visual and continuous. The FSA instrument was originally three pages, but the FSA-R is condensed to a single page—based on the formats of the Person Picking an Apple from a Tree (PPAT) and Kinetic House-Tree-Person (K-H-T-P). Effort and graphic design layout revealed an appealing format with clear instructions, which are essential to a valid test (Fowler, 1993). A reasonable time limit was set at 30 minutes. Welch and Hoover (1993) offered advice on the fourth consideration of item development, that of technical consideration. The items underwent professional review from colleagues at FSU, and psychometricians at Professional Testing, Incorporated (PTI).

The fifth and final stage involved the initial pilot test administration. The subjects and conditions of the pilot test should approximate subsequent testing. Participants did not receive
the impression that this test was a pilot, but that it was the actual test. Elements in the testing room received standardization for experiments (Christensen, 2001; Cronbach, 1984). Administrators, who were clinical staff, made efforts to standardize and record lighting, seating arrangement, materials, ventilation, temperature, time, and other environmental factors. All administrators received a checklist for these variables. This is standard procedure for the senior administrator at PTI (D. M. Clark, personal communication, July 17, 2009). A post-interview about the structure of the test followed each administration. Roid and Haladyna (1982) proposed guidelines for the actual test design in the fifth consideration. It entailed planning, research, psychometric considerations, and pragmatics. Computerized data from the pilot will inform the planning of the second administration.

Prior to any of these stages, theory should drive the construction and function of a test. Most good assessment instruments, like the Miller Clinical Multiaxial Inventory (MCMI, Millon, 1997), work through some form of theory (J. Hanson, personal communication, February 27, 2006). The theories that underlie the FSA-R include Classical Measurement theory, and Questioned Documents Theory. The next section outlines the restated research questions for this study, followed by the actual instrument administration phase.

**Null Hypothesis**

The null hypothesis asserts that PDIAS cannot rate the formal elements of color and/or free space to distinguish artwork of those with Major Depression from the artwork of those without Major Depression.

**Research Questions**

The research questions and assumptions are as follows:

1. Can PDIAS rate the formal elements of color and/or free space to distinguish artwork of those with Major Depression from the artwork of those without Major Depression.
2. Can the FSA-R instrument become a standardized instrument for Major Depression?
3. Do common graphic indicators exist amongst the artwork of those with Major Depression?
4. Once completed, can components of this computer-rated instrument generalize to other assessments, or populations?
Research Methodology

A mixed method post-positivist approach constituted the experimental approach for this study (Trochim, 2000). The post-positivist critical realist recognizes that all observation is fallible and contains error, and that all theory is revisable. When trying to measure difficult concepts, such as personality types or indicators of mental illness, a single theory may be too narrow to address the amount of variables involved. All tests have some margin of error; the goal is to be as accurate as possible in measurement.

Because all measurement is fallible, the post-positivist emphasizes the importance of multiple measures and observations, each of which may possess different types of error, and the need to use triangulation across these multiple erroneous sources to achieve objectivity. Test construction requires the administrator to consider the behavior of the examinee, extraneous variables, medical history, demographical history, experience, and other variables. All these are possible sources for error, and require consideration. In addition, good tests feature some measure of concurrent validity with other tests to reduce some of this error (Kline, 1987).

In post-positivism, theories change. When one approach does not work, the researcher tailors it until it does. This trial-and-error approach is necessary for constructing a difficult test, especially a visual one.

Another tenet is that the way researchers think and work and the way we think in our everyday life are not distinctly different. Scientific reasoning and common sense reasoning are essentially the same process. A post-positivist can use the scientific method selectively. Constructing tests requires intuition and creativity.

This is a mixed-methods approach with an emphasis in quantitative study. This study includes the use of experimental design. The scientific method applied as stringently as possible, with some room for flexibility.

This study used convenience sampling. Convenience sampling entails selecting subjects that are nearby. Most subjects came from close facilities and organized support groups. A randomization of sample collection will replace convenience sampling in time. A random selection in sampling ensures that every member of a particular population has an equal chance of representation (Glass & Hopkins, 1996).
Research Design

The primary design of this research is comparative after-only design. After-only design involves recording the dependent variable one time after administration of the treatment condition to the experimental group (Christensen, 2001). There are two dependent variables measured (color percentage, free space area) compared to each independent group (Major Depression, control). Variables of analysis included Major Depressive Disorder and control as the independent variables. Dependent variables included color coverage and free space area. Controlled variables included the severity index of Major Depressive Disorder, age, sex, and socio-economic status. Extraneous variables included training, manuals, directives, and comfort environmental factors.

Statistical Analysis

A t-test for independent groups’ comparison of variables was used in this study (Glass & Hopkins, 1996). An alpha level of .05 evaluated the main hypothesis. Nine variables: blue, orange, red, green, purple, yellow, brown, black, and free space registered as measurements through PDIAS analysis. Raw scores from these nine categories were collected from 40 total samples (Major Depression n = 20, controls n = 20).

The next task was to convert the raw scores to t-scores. From this, the means of each of the nine categories compared through the t-test procedure. Analysis revealed the means with the most significant differences, and these differences determined the outcome of the original hypothesis.

Population and Sample Considerations

Participants included individuals presenting with Major Depressive Disorder (n = 20), and controls without this designation (n = 20). A sample size of 20 is sufficient in order to run most statistical procedures, and to assume a normally distributed test statistic (Dattalo, 2008; Steinberg 2007). It was necessary to include adults with varied diagnoses, as this may actually be more representative of the population. No research of the FSA-R exists to examine the effects on children or adolescents (Mattson, 2009). Therefore, children and adolescents were excluded from this study.

A nationally normed test is preferable, but a regionally normed test is sufficient in most circumstances (J. C. Hanson, personal communication, April 4, 2007). Norm groups hailed from
three locations across two states. The FSA-R used convenience sampling to pool test subjects in which little randomization occurred.

Sample size selection resulted through employing several measures. Using ten times more subjects than items is a sound rule for test construction (Adkins, 1960). If the test features 10 items, 100 subjects should be included in the population for the pilot stage of testing. However, this pilot test did not receive such high numbers. Nonetheless, it is important to make sure that the sample size is large enough to validate the hypothesis (Mertens, 2005).

The population selection resulted from the assumption that there is a connection between mental illness and artistic creation. That, at its very basic premise, the presence of mental illness will exhibit some change from a control in artwork products. Researchers may look to the neurological level to find patterns of brain activity that affect these regions during the process of artmaking (Andreasan, 2001). Another source backing the population choice was biographical data that showed positive links between mental illness and art (Jamison, 1993).

**Control Group**

Demographics of the control participants are in Table 1. The control conditions closely matched the experimental conditions. Controls were individuals without Major Depressive Disorder. To meet criteria for control, they needed to be relatively free of psychiatric conditions, as screened by the intake information in the participant packet. The control group consisted of students from a community college set in northern Michigan \((n = 24)\). As principal investigator for this region, I met with the Art Appreciation class instructor, and we reviewed the manual, materials, and procedures prior to the administration. I modeled the FSA-R administration to the instructor. All manners of instruction for the control groups closely matched the experimental groups.

Using random selection of 24 participants, I derived 20 samples from the pool (Christensen, 2001). The average age of the participants was 29 \((SD = 13.09)\). All were white, which resulted in a potentially biased sample. The discussion section explains the implications for this bias. Ten of them were male, while 14 were female. Most were single and reported at least some college experience. The average income was $12,100 annually. The participants were relatively free of psychiatric conditions.
Table 1

Demographics of Control Group

<table>
<thead>
<tr>
<th>Participant</th>
<th>Age</th>
<th>Sex</th>
<th>Race</th>
<th>Marital</th>
<th>Education</th>
<th>Income</th>
<th>Med/Psych</th>
<th>Rx</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>23</td>
<td>M</td>
<td>White</td>
<td>S</td>
<td>17</td>
<td>0-5,000</td>
<td>Substance Abuse</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>37</td>
<td>F</td>
<td>White</td>
<td>M</td>
<td>14</td>
<td>30,000+</td>
<td>--</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>18</td>
<td>F</td>
<td>White</td>
<td>S</td>
<td>13</td>
<td>0-5,000</td>
<td>--</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>21</td>
<td>M</td>
<td>White</td>
<td>S</td>
<td>14</td>
<td>0-5,000</td>
<td>Stroke</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>49</td>
<td>F</td>
<td>White</td>
<td>M</td>
<td>12</td>
<td>25,000-30,000</td>
<td>Asthma</td>
<td>No</td>
</tr>
<tr>
<td>6</td>
<td>20</td>
<td>F</td>
<td>White</td>
<td>S</td>
<td>13</td>
<td>0-5,000</td>
<td>Thyroid</td>
<td>No</td>
</tr>
<tr>
<td>7</td>
<td>22</td>
<td>M</td>
<td>White</td>
<td>S</td>
<td>16</td>
<td>10,000-15,000</td>
<td>Anxiety Disorder, Alcoholism, OCD</td>
<td>Yes</td>
</tr>
<tr>
<td>8</td>
<td>52</td>
<td>F</td>
<td>White/Other</td>
<td>M</td>
<td>18</td>
<td>15,000-20,000</td>
<td>Cancer</td>
<td>No</td>
</tr>
<tr>
<td>9</td>
<td>22</td>
<td>M</td>
<td>White</td>
<td>S</td>
<td>13</td>
<td>0-5,000</td>
<td>OCD</td>
<td>No</td>
</tr>
<tr>
<td>10</td>
<td>18</td>
<td>F</td>
<td>White</td>
<td>S</td>
<td>13</td>
<td>0-5,000</td>
<td>--</td>
<td>No</td>
</tr>
<tr>
<td>11</td>
<td>21</td>
<td>F</td>
<td>White</td>
<td>S</td>
<td>15</td>
<td>25,000-30,000</td>
<td>Asthma</td>
<td>No</td>
</tr>
<tr>
<td>12</td>
<td>19</td>
<td>F</td>
<td>White</td>
<td>S</td>
<td>15</td>
<td>10,000-15,000</td>
<td>--</td>
<td>No</td>
</tr>
<tr>
<td>13</td>
<td>30</td>
<td>M</td>
<td>White</td>
<td>M</td>
<td>13</td>
<td>30,000+</td>
<td>Alcoholism, Substance Abuse</td>
<td>No</td>
</tr>
<tr>
<td>14</td>
<td>20</td>
<td>M</td>
<td>White</td>
<td>S</td>
<td>14</td>
<td>0-5,000</td>
<td>--</td>
<td>No</td>
</tr>
<tr>
<td>15</td>
<td>54</td>
<td>M</td>
<td>White</td>
<td>M</td>
<td>12</td>
<td>5,000-10,000</td>
<td>High Blood Pressure</td>
<td>No</td>
</tr>
<tr>
<td>16</td>
<td>19</td>
<td>M</td>
<td>White</td>
<td>S</td>
<td>13</td>
<td>0-5,000</td>
<td>--</td>
<td>No</td>
</tr>
<tr>
<td>17</td>
<td>22</td>
<td>M</td>
<td>White</td>
<td>S</td>
<td>13</td>
<td>10,000-15,000</td>
<td>--</td>
<td>No</td>
</tr>
<tr>
<td>18</td>
<td>19</td>
<td>F</td>
<td>White</td>
<td>S</td>
<td>16</td>
<td>0-5,000</td>
<td>--</td>
<td>No</td>
</tr>
<tr>
<td>19</td>
<td>44</td>
<td>F</td>
<td>White</td>
<td>S</td>
<td>14</td>
<td>0-5,000</td>
<td>--</td>
<td>No</td>
</tr>
<tr>
<td>20</td>
<td>49</td>
<td>M</td>
<td>White</td>
<td>M</td>
<td>14</td>
<td>30,000</td>
<td>Asthma</td>
<td>No</td>
</tr>
</tbody>
</table>

Note. Med/Psych = Medical or psychiatric conditions; Rx = Pharmaceutical treatments.
Major Depression Groups

Demographics of the experimental participants are in Table 2. Those with an existing diagnosis of 296.3x Major Depressive Disorder (MDD) participated in this study. If they did not have this diagnosis, participants needed to indicate that they held criteria for MDD through the intake form, and a score of at least 14 (moderate depression) on the BDI-II. As indicated in the literature review, individuals completing this test received instructions to answer the items in accord with how they felt within a 2-week period. They answered these items by means of an intensity scale from 0 to 3. The higher the total score, the more depression. The cutoff scores range from 0 to13 as having minimal depression, 14 to 19 mild, 20 to 28 moderate, and 29 to 36 severe (Beck et al., 1996). Appendix D lists the DSM-IV-TR (2000) criteria for MDD. Depressed mood or loss of interest or pleasure is hallmark of MDD.

The depressed mood generally lasts most of the day, nearly every day, for 2 weeks, or more as indicated by subjective report, or observation made by others. Markedly diminished interest or pleasure in most activities occurs, and weight or appetite change. Other symptoms include insomnia or oversleeping, psychomotor agitation or retardation, fatigue, feelings of worthlessness or guilt, diminished ability to think or concentrate, inability to make decisions, or recurrent thoughts of death or suicide.

The MDD group came from two locations. The first group was from a psychosocial facility modeled after a clubhouse setting where residents with mental illness from the community gathered for meetings and supportive social venues. Major demographic considerations of this site were the fact that most all were White, with a low median income and level of education; a significant portion of the participants reported schizophrenia as a main diagnosis ($n = 6$). Recruitment stemmed from collaborations with the case manager of the facility. The case manager addressed the attendees three weeks prior to the administration to rally a survey of those who would attend. The actual administration occurred with the case manager serving as proctor following brief training on the instrument, which occurred on two dates within one month prior to the administration. As principal investigator, I administered the instrument and collected the data, following all procedures outlined in the Administrator’s Manual. I arrived 30 minutes early to the main activity area to control for variables. A small group of participants ($n = 11$) showed for the administration. Of these, five met the criteria for MDD. Four out of five of these presented with comorbidity.
The second group came from a depression support group in northern Florida. Recruitment stemmed from the group leader, and this leader received correspondence packets and instructions on how to administer the FSA-R instrument. Major demographic features of this group included an almost entirely white designation, except for two individuals; one reported *other* in the race category, and the other reported *Hispanic*. The demographics form of the clubhouse and support group also revealed a sample slightly older than the control group; the average age for the combined groups was 46 ($SD = 14.61$), with great variation in other variables, such as income level and marital status. Twenty-six individuals participated in the administration. Of these, one individual declined, some did not meet the criteria for MDD, and random sampling phased out the remaining samples, which resulting in the remaining viable samples ($n = 15$).
Table 2

Demographics of Experimental Group

<table>
<thead>
<tr>
<th>Participant</th>
<th>Age</th>
<th>Sex</th>
<th>Race</th>
<th>Marital</th>
<th>Education</th>
<th>Income</th>
<th>Med/Psych</th>
<th>Rx</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>45</td>
<td>F</td>
<td>White</td>
<td>S</td>
<td>12</td>
<td>5,000-10,000</td>
<td>Major Depression, Anxiety, Schizophrenia</td>
<td>Yes</td>
</tr>
<tr>
<td>2</td>
<td>34</td>
<td>M</td>
<td>White</td>
<td>S</td>
<td>13</td>
<td>5,000-10,000</td>
<td>Major Depression, Substance Abuse, Schizophrenia, Thyroid, Epilepsy</td>
<td>Yes</td>
</tr>
<tr>
<td>3</td>
<td>26</td>
<td>M</td>
<td>White</td>
<td>--</td>
<td>12</td>
<td>0-5,000</td>
<td>Major Depression</td>
<td>Yes</td>
</tr>
<tr>
<td>4</td>
<td>31</td>
<td>M</td>
<td>Hispanic</td>
<td>S</td>
<td>14</td>
<td>10,000-15,000</td>
<td>Major Depression</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>19</td>
<td>M</td>
<td>White</td>
<td>S</td>
<td>13</td>
<td>0-5,000</td>
<td>Anxiety, Major Depression, OCD</td>
<td>No</td>
</tr>
<tr>
<td>6</td>
<td>48</td>
<td>M</td>
<td>White</td>
<td>D</td>
<td>20</td>
<td>25,000-30,000</td>
<td>Major Depression, Substance Abuse, OCD</td>
<td>Yes</td>
</tr>
<tr>
<td>7</td>
<td>39</td>
<td>F</td>
<td>White</td>
<td>M</td>
<td>16</td>
<td>30,000+</td>
<td>Major Depression</td>
<td>No</td>
</tr>
<tr>
<td>8</td>
<td>44</td>
<td>F</td>
<td>White</td>
<td>M</td>
<td>16</td>
<td>30,000</td>
<td>Major Depression</td>
<td>No</td>
</tr>
<tr>
<td>9</td>
<td>41</td>
<td>M</td>
<td>White</td>
<td>D</td>
<td>13</td>
<td>10,000-15,000</td>
<td>Major Depression, High Blood Pressure</td>
<td>No</td>
</tr>
<tr>
<td>10</td>
<td>60</td>
<td>F</td>
<td>White</td>
<td>--</td>
<td>16</td>
<td>30,000+</td>
<td>Major Depression, Anxiety Disorder, ADHD, PTSD, OCD, Asthma, Thyroid</td>
<td>No</td>
</tr>
<tr>
<td>11</td>
<td>67</td>
<td>F</td>
<td>Other</td>
<td>D</td>
<td>17</td>
<td>5,000-10,000</td>
<td>Major Depression, ADHD, PTSD</td>
<td>No</td>
</tr>
<tr>
<td>12</td>
<td>31</td>
<td>F</td>
<td>White</td>
<td>D</td>
<td>16</td>
<td>25,000-30,000</td>
<td>Major Depression, ADHD</td>
<td>Yes</td>
</tr>
<tr>
<td>13</td>
<td>39</td>
<td>F</td>
<td>White</td>
<td>M</td>
<td>14</td>
<td>0-5,000</td>
<td>Major Depression, Paranoia</td>
<td>Yes</td>
</tr>
<tr>
<td>14</td>
<td>68</td>
<td>F</td>
<td>White</td>
<td>D</td>
<td>20</td>
<td>5,000-10,000</td>
<td>Major Depression, PTSD, Heart Conditions, High Blood Pressure, Thyroid, Myasthenia</td>
<td>Yes</td>
</tr>
<tr>
<td>15</td>
<td>53</td>
<td>F</td>
<td>White</td>
<td>D</td>
<td>13</td>
<td>0-5,000</td>
<td>Major Depression, Cancer</td>
<td>Yes</td>
</tr>
<tr>
<td>16</td>
<td>49</td>
<td>M</td>
<td>White</td>
<td>D</td>
<td>10</td>
<td>5,000-10,000</td>
<td>Major Depression</td>
<td>No</td>
</tr>
<tr>
<td>17</td>
<td>68</td>
<td>M</td>
<td>White</td>
<td>M</td>
<td>17</td>
<td>30,000+</td>
<td>Major Depression, PTSD, Heart Conditions, High Blood Pressure</td>
<td>No</td>
</tr>
<tr>
<td>18</td>
<td>45</td>
<td>F</td>
<td>White</td>
<td>M</td>
<td>16</td>
<td>0-5,000</td>
<td>Major Depression, ADHD, PTSD, High Blood Pressure, Thyroid</td>
<td>Yes</td>
</tr>
<tr>
<td>19</td>
<td>69</td>
<td>F</td>
<td>White</td>
<td>M</td>
<td>10</td>
<td>--</td>
<td>Major Depression, Asthma</td>
<td>No</td>
</tr>
<tr>
<td>20</td>
<td>43</td>
<td>M</td>
<td>White</td>
<td>M</td>
<td>12</td>
<td>0-5,000</td>
<td>Major Depression</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Note. Med/Psych = Medical and/or psychiatric conditions; Rx = Pharmaceutical treatment.
Instrumentation

The FSA-R

The Face Stimulus Assessment-Revised (FSA-R) is a visual instrument constructed from components of the Face Stimulus Assessment (FSA, Betts, 2003). It consists of a single sheet of paper measuring 21.59 centimeters x 27.94 centimeters. The paper contains an outline of a pre-drawn face. Surrounding the face is a series of visual items, with copy tasks lining the bottom.

The face contains features that are half-drawn for allowing participants to project more information onto the image (Mattson, 2009). The face also underwent downscaling to allow for additional room on the paper for color, or additional imagery. The early version of the FSA-R contained numerous items within a grid (see Appendix F).

After conceptualizing the items that were detailed in the literature review, computerized tools assisted in drafting them. After drafting, they underwent sizing to fit within the paper relative to the pre-drawn face. Initial placement of the copy tasks was random.

A series of informal administrations of the early version of the FSA-R shaped its current design (see Appendix G). These participants were colleagues, students, and test constructors. Feedback from the informal groups included reducing the size of the face, arranging the copy tasks in a logical position, and eliminating the graph lines encompassing the items. These groups also offered input on the choice of markers, ultimately leading to a change from alcohol-based to water-based markers.

The FSA-R instrument is an adaptation of the FSA templates (see Appendix G). The heavy emphasis of the first drawing comes from the kinetic models of Burns (1987), where all three elements of the Kinetic House-Tree-Person (KH-T-P) were included on the same page because of the rich interactions of client insight that resulted. It also resulted from the single-page drawing of the Person Picking an Apple from a Tree (PPAT, Gantt & Tabone, 1998). Additionally, the drawing features testing of psychomotor retardation (PR) through copying tasks (Thomassen et al., 1991; Hegerl et al., 2004; Mergl et al, 2007; Sabbe et al., 1996; Pier et al., 2004a; Pier et al., 2004b; Sabbe et al., 1999). The FSA-R instrument varies from standard PR tasks in that the stimulus designs were adapted and transformed into artistically depicted objects, making this an art-based test. For example, the manual line bisection task became the knife picture from Silver’s screening for depression (Silver, 1996). The stimulus item of the Fitt’s task became the shape of a tree, rather than successive circles seen in conventional PR tasks.
The boundaries within the drawing allow for separation of test items. They may also test for blurring and obscuring details, as this manner of drawing is characteristic of the depressed individual (Simon, 1992). Colors and lines often bleed through boundaries on the paper, often giving an impressionistic style to the result.

It is theorized that the individual with depression will likely not be able to complete the tedious task of drawing within each individual boundary, hence stopping at some point, skipping over items, or drawing within items in a haphazard manner. Those with depression often lacked energy when faced with drawing tasks, which often resulted in incomplete or haphazardly drawn imagery (Dawson, 1984; Dax, 1953; Gantt & Tabone, 1998; Hammer, 1958; Lee & Lawlor, 1995; Lilienfeld et al., 2004; Malchiodi, 1998; Souza et al., 2003; Wadeson, 1971; Wohl & Kaufman, 1985). These are all testable points comparable to a normal control population.

There exists a tendency for individuals to draw based on similarity and proximity (Arnheim, 1974). This gestalt tenet also guides the rationale for stimulus drawings. An individual naturally connects the gaps present within line drawings. This holds true even when the figure is incomplete. I needed to determine what people with depression draw, and include it as a positively or negatively scored item. The templates represent common objects that are perceptually incomplete.

Perceptual learning occurs within the human brain when it experiences certain objects repeatedly (Doniger, Foxe, Schroeder, Murray, Higgins, & Javitt, 2001). The objects of a sun, cloud, ground line, and tree are included because of their perceptual universality. The developer of the H-T-P explained that the objects of the house, tree, and person are items that all age groups are familiar with (Buck, 1987). One may also assume that most all humans have seen and are thus familiar with these particular objects. This is an attribute in considering the FSA-R as a sound multicultural instrument. There are several dimensions to this test. These include dark colors, lack of colors, sketchy/hesitant line, constricted environment, lack of details, and line error. Line dimensionality will be tested in a future study.

A time limit accompanies the instrument. Many art-based assessments do not make use of a time limit, however, a limit is important because it preserves test standardization. Time limits may or may not affect performance. Performance did not increase under any category except reading when behaviorally disordered adolescents were given unlimited time for proscribed testing (Forness & Dvorak, 1982). Similarly, results of Rudman and Raudenbush (1986)
indicated that extra testing time presented a positive linear effect on reading comprehension and total reading scores. The benefit of excess time was most evident at 5 minutes posttest.

Personal experience in using the original FSA administration instructions resulted in frustrated and confused participants (Mattson, 2008). For verbal tests instructions, one needs to more effectively define the task, direct attention, and motivate the participant (Sidowski & Lockhard, 1966). As there is an attempt to parallel the procedures used by verbal tests, the FSA underwent modification for the FSA-R. The only line extracted from the original FSA instructions includes the command for using markers. The remaining structure of the instructions stems from the instructions of the BDI-II.

The BDI-II

The Beck Depression Inventory in its abbreviated form (BDI-II) tests a wide range of depression symptoms (Beck & Alford, 2008; Groth-Marnat, 2003). The BDI-II is a 21-item self-report of depressive symptoms derived from psychiatric research. It was widely used for screening depression amongst psychiatric patients for many years (Steer et al., 1999). This screening tool is able to detect depression in psychiatric populations and normal individuals (Beck et al., 1996).

Individuals completing this test receive instructions to answer the items in accord with how they feel within a 2-week period. They answer these items by means of an intensity scale from 0 to 3. The higher the total score, the more depression one is experiencing. The cutoff scores range from 0 to 13 as having minimal depression, 14 to 19 mild, 20 to 28 moderate, and 29 to 36 severe. Typical questions include dimensions of low mood, irritability, guilt, failure, and disturbances in sleep and appetite (Groth-Marnat, 2003). After the administration of the inventory, which usually takes from 5 to 10 minutes, the scores are added and assessed against an interpretive scale. Clinically depressed individuals generally score between 14 and 28 (Beck et al., 1996).

PDIAS

For this study, PDIAS programs were adapted for the analysis of the FSA-R. The software includes downloadable programs such as Image Java v.1.43 (ImageJ), and Measuring Vegetative Health v.8 (MVHimage). National Institutes of Health (NIH) developed ImageJ to analyze biomedical images with high degrees of accuracy, which stemmed from comparing known to measured areas within an image (W.S. Rasband, personal communication, November
ImageJ accuracy also resulted from extensive beta testing of its tools, and a measurements scale set by its users (Sonka & Hanson, 2001). Educational programs and color analysis of satellite imagery are primary applications of MVHimage. The developer of MVHimage established its accuracy by testing select items prior to release, though independent testing is pending (J. Pickle, personal communication, October 27, 2008).

**Consent Form**

According to Human Subjects research, a consent form is required if a study involves human participants. The form used for this study is in Appendix H. It lists basic requirements of Human Subjects testing, such as duty to inform, reasonable expectations, possible deleterious consequences, protocol for managing emergencies, benefits, procedures, participant rights and responsibilities. The Florida State University Office of Human Subjects approved this study. The approval form is in Appendix I.

**Demographics Form**

A one-page form issued to participants garnered background information. This form is in Appendix J. Such categories of information include age, sex, living location, socio-economic status, medical history, and mental health history. These informed the norming of the FSA-R test.

**Post-Interview Form**

A one-page questionnaire was included in the test administration packet at the end of the FSA-R instrument. This questionnaire is in Appendix K. This is a technique highly recommended by test constructors (Fishman & Galguera, 2003). It asks questions directed towards improving the second run of the test. Such questions include those related to overall format, troublesome test items, and time limit. Additionally, it offers the participants debriefing in the instance stress or other unanticipated events occurred (Christensen, 2001).

**Materials**

The FSA-R participants used eight Crayola Multicultural® and eight Sharpie Flip Chart® markers. This choice resulted from field-testing over 34 different varieties of marker, and seven different types of paper to obtain optimum surface coverage. Additionally, I considered the average statistics of consumer color preference, color preference research, and color preference amongst those with depression (Burkitt, Barrett, & Davis, 2004; Dittmar, 2001; Frank, 1993; Richards & Ross, 1967; Scott, 1969; Valdez & Mehrabian, 1986). The markers presented in an
order of most to least preferred and included colors normally preferred by those with depression, in accord with the research.

**Procedures**

**Scoring Manual**

Since many testers ultimately issued the FSA-R, standardized procedures in a specially written manual guided the testing procedure (Christensen, 2001). The manual is 21 pages long, and for this reason, it is not included as an appendix (Mattson, 2011). The guidelines for the PPAT and the H-T-P inspired the construction of the FSA-R manual. The completed FSA-R manual is modeled after several existing manuals of instruments, including the FEATS (Gantt, & Tabone, 1998), the H-T-P (Buck, 1970), the BDI (Beck, 1996), the Minnesota Multiphasic Personality Inventory (MMPI, Hathaway & McKinley, 1940), and numerous standard manuals issued from Professional Testing, Incorporated (PTI). Based on preliminary trials, scoring of the FSA-R proved simple because of the automated processes of computerized rating features.

**Scoring Procedures**

Nine categories: blue, orange, red, green, purple, yellow, brown, black, free space, were measured. PDIAS analyzed the scanned results and converted them to raw scores. Raw scores from these 12 categories were collected from 40 total samples (Major Depression \( n = 20 \), control \( n = 20 \)). Statistical operations transformed the raw scores into \( t \) scores. Color and free space was measured by percentages of highlighted pixels over the surface area of the images.

**Instructions**

Instructions serve the purpose of defining the task, directing participant attention, developing sets, and motivating participants (Sidowski & Lockard, 1966). To be effective, they must not be vague. Instructions may be lengthy, demanding several operations at once from a participant, but they must be clear and specific (Sutcliffe, 1972). The following instructions resulted from empirical research and pre-trials of issuing the FSA-R by Mattson (2011):

Using the markers provided to you, color, draw, or mark within the sheet. Make sure whatever you do reflects how you have been feeling during the past two weeks, including today. If you arrive at an item that is already drawn and divided in the square by a line, please copy this item in the space provided to the right of it. You will be given 30 minutes to complete this activity. When you are finished, please turn the page and fill out
the post-interview. When completely finished, please remain quietly seated while all other individuals finish. (p. 16)

If the participant asks for clarification, the administrator states, “I can only repeat the instructions. If you have further questions, please reserve them until the end of this activity.” A more detailed script for administrators exists within the testing manual, which is typical of thorough testing manuals (D. M. Clark, personal communication, July 16, 2009).

Time Limit

Time limit was set at 30 minutes. The past informal administrations of the FSA-R and general testing guidelines determined this time. The general rule-of-thumb is to allow 1.5 minutes completion time per test item (Osterlind, 1998). However, this will vary based on the format and difficulty of the test (D. M. Clark, personal communication, July 16, 2009).

Data Analysis

Equipment

Analysis of the data commenced with a flatbed Hewlett-Packard Printer–Scanner–Copier (HP PSC) 1410 with 4800×1200 dots per inch (dpi) resolution. Image processing involved a computer with 2.4 gigahertz (GHz) of read-only memory (ROM) and 512 megabytes (MB) of random access memory (RAM). These specifications handled the main programs.

The main programs used for statistical analyses included Microsoft Excel 2010 ©, and the Statistical Package for the Social Sciences (SPSS) v. 17.0.1. Excel accepted raw scores from the images and organized initial data. SPSS permitted statistical operations.

Processing

The basic steps to image processing include acquisition, discrimination, segmentation, measurement, and stereological interpretation (Russ, 1990). The final step of stereological interpretation normally pertains to three-dimensional objects, so this study excluded it. Acquisition, however, was necessary for image processing. Conversion to Tagged Image File Format (TIFF) from an analog state further readied the image for processing. The Joint Photographic Experts Group (JPEG) format, although high fidelity, does not allow for true image analysis (Francus, 2004). Therefore, this study retained the TIFF format. Scale of the image format was set for accuracy at 84 pixels per centimeter, and filtering techniques lowered noise.

Filtering is a component of processing that addresses noise. Noise can be introduced at various points of image acquisition and processing, and may alter the results of analysis. This
analysis used a *Gaussian blur* filter. The Gaussian blur filter is a common tool used in image analysis (Russ, 1997). This filter rejects certain types of noise. It is possible to use additional methods to filter images prior to analysis to improve appearance (e.g., auto-contrast, sharpen, smoothen, brighten, and subtract). However, it is best to avoid these techniques because most will introduce bias into the analysis (Hedges, 2006). This study minimized image analysis filtering.

After filtering, *grayscale* was applied to the scanned image by simple conversion. Grayscale pertains to neutral gray shades ranging from black to white on an intensity spectrum. This prepared the image for further processing. Grayscale conversion is included in the discrimination step (Russ, 1990). Binary conversion followed grayscale representation. Binary conversion reduces images to distinct black and white points, leading to easier measurement. This also allows for greater figure-ground discernment. Thresholding the image then separated additional elements from these relationships by means of pixel brightness discrimination. The segmentation step further processed binary images. Segmentation refers to partitioning pixel regions (Shapiro & Stockman, 2001). After segmentation, the measurement step generated data. This was the final stage used in this study. The following conclusion summarizes the main points of this chapter.

**Conclusion**

This chapter presented the methods and materials used for the administration of the FSA-R instrument. The construction of this instrument follows basic test development guidelines through comparative after-only research design. The participant samples stemmed from various facilities, and standardization techniques interlaced the testing methodology. This chapter featured the forms, materials, hardware specifications, and image analysis techniques used in the study.
CHAPTER 4
RESULTS

The primary interest of this study was to determine if those with Major Depressive Disorder (MDD) drew different levels of formal elements than those without MDD, and to construct a computer-rated visual screening known as the Face Stimulus Assessment-Revised (FSA-R) to screen these differences. The previous chapter outlined the procedures that informed the construction of the FSA-R. This chapter presents the results as they relate to the hypothesis.

In response to the initial hypothesis that the formal elements of color and/or free space from the FSA-R, rated by public domain image analysis software (PDIAS), can distinguish Major Depression artwork from the artwork of those without Major Depression, results indicate that certain colors and free space distinguish the groups. Those with Major Depression drew with less purple, less orange, and more free space than controls.

Although this study is primarily quantitative in nature, it is equally important to explain the qualitative aspects of art-based assessment instruments. This ensures that clinicians receive comprehensive pictures of their clients (Betts, 2005; Gantt & Tabone, 1998). John Buck (as cited in Bowyer, 1970), developer of the House-Tree-Person assessment, stated that researchers should aim for a hybrid of quantitative and qualitative assessment in what he called the “quality of the quantity” (p. 24). A section on results that are qualitative in nature follows the quantitative results.

Quantitative Results

Results of the statistical analyses are in Table 3. There were two dependent variables measured (color percentage, free space area) for each independent group (Major Depression, control). An independent-samples t-test was conducted to compare color percentages in experimental and control conditions. This study employed an alpha level of .05, meaning that in each case, there existed a 5% chance of Type I error. The confidence intervals of group means revealed that with 95% confidence, the intervals captured the true population means. There was a significant difference in color scores between the experimental group \( (M = .16, SD = .53, 95\% \ CI [-.07, .39]) \) and the control group \( (M = 3.5, SD=4.89, 95\% \ CI [1.36, 5.64]) \) for the color purple, \( t(38) = -2.95, p = .05, d = -.96 \), and there was a significant difference in color scores between the experimental group \( (M = .08, SD = .18, 95\% \ CI [0, .16]) \) and the control group \( (M = 1.64, SD = 3.13, 95\% \ CI [.27, 3.01]) \) for the color orange, \( t(38) = -2.28, p = .05, d = -.70 \).
Cohen’s $d$ reflected the effect sizes. The standardized means exhibit negative values because the effect (Major Depression) decreased the color usage. The effect sizes in these two instances are *large*, and *medium*, respectively (Cohen, 1988).

An independent-samples $t$-test was conducted to compare free space percentages in experimental and control conditions. There was a significant difference between the experimental group ($M = 78.98$, $SD = 20.14$, 95% CI [70.15, 87.81]) and the control group ($M = 64.56$, $SD = 19.32$, 95% CI [56.09, 73.03]) for free space usage, $t(38) = 2.26$, $p = .05$, $d = .73$. The effect size in this instance is *medium*.

For hypothesis testing these three variables, the absolute value of $t$ exceeded the critical value. Additionally, the intervals did not capture what the null hypothesis specified. Therefore, the decision is to reject the null hypothesis for these variables. Because $n_1 = n_2$, this test is robust to the equal variance assumption (Agresti & Finlay, 2008), meaning no post-hoc adjustments are required.

**Computerized Results**

The images in this study underwent computerized analysis from public domain image analysis software (PDIAS) programs. The average time that it took to scan the images was 51 seconds, while the average time for actual analysis was approximately 4 minutes. Pixel size of each image was 292,000.

Manual conversion to Tagged Image File Format (TIFF) prepared the images for processing through PDIAS programs. ImageJ converted all images to black and white grayscale, a necessary condition for processing. After conversion, a Gaussian blur operator filtered the image with a *sigma* value of 1.0. This essentially reduced the brightness of pixels in a neighboring matrix, giving the image a blurred appearance. This blurring reduced image detail so that color prominence became more apparent. A standard brightness histogram typically used in image analysis is in Figure 2. It depicts an image completed by a member of the control group. The Y-axis of the graph displays the number of pixels, while the X-axis displays the *brightness value (BV)* from 0 to 250, 0 representing black and 250 denoting white. The narrow peaks on the far right of the graph are typical manifestations of *clipping*, commonly seen when images contain prominent whites or overexposure. The Gaussian blur filter increased the contrast of the image, evidenced by widening the peaks in the graph. It also smoothed the image, noted by less peaks in
the second graph. These were typical effects observed in pre- and post-filtering of the other images.

The actual analysis involved a region of interest (ROI) tool. This was a square drawn around the area that required measurement. After selecting the area, the PDIAS program automatically calculated the percentages of pixels represented in the square and displayed the results in the form of an image mask. The average pixel size of the ROI tools was 318.56.

Figure 1. Brightness histograms before (left), and after (right), the application of a Gaussian blur filter to an image (below) from the control group.
Table 3

Differences in Formal Element Percentages between Experimental and Control Group

<table>
<thead>
<tr>
<th>Var</th>
<th>M</th>
<th>SD</th>
<th>M</th>
<th>SD</th>
<th>t(38)</th>
<th>p</th>
<th>LL</th>
<th>UL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blue</td>
<td>3.18</td>
<td>7.13</td>
<td>4.02</td>
<td>5.74</td>
<td>-.4</td>
<td>.05</td>
<td>-2.42</td>
<td>1.62</td>
</tr>
<tr>
<td>Green</td>
<td>.44</td>
<td>1.11</td>
<td>5.11</td>
<td>6.43</td>
<td>-1.19</td>
<td>.05</td>
<td>-3.21</td>
<td>.83</td>
</tr>
<tr>
<td>Purple</td>
<td>.16</td>
<td>.53</td>
<td>3.5</td>
<td>4.89</td>
<td>-2.95*</td>
<td>.05</td>
<td>-4.97</td>
<td>-.93</td>
</tr>
<tr>
<td>Red</td>
<td>1.11</td>
<td>2.22</td>
<td>1.21</td>
<td>1.74</td>
<td>-.15</td>
<td>.05</td>
<td>-2.17</td>
<td>1.87</td>
</tr>
<tr>
<td>Black</td>
<td>1.22</td>
<td>1.85</td>
<td>2.67</td>
<td>3.88</td>
<td>-1.49</td>
<td>.05</td>
<td>-3.51</td>
<td>.53</td>
</tr>
<tr>
<td>Brown</td>
<td>10.89</td>
<td>14.09</td>
<td>5.78</td>
<td>11.26</td>
<td>1.27</td>
<td>.05</td>
<td>-.75</td>
<td>3.29</td>
</tr>
<tr>
<td>Orange</td>
<td>.08</td>
<td>.18</td>
<td>1.64</td>
<td>3.13</td>
<td>-2.28*</td>
<td>.05</td>
<td>-4.29</td>
<td>-.27</td>
</tr>
<tr>
<td>Yellow</td>
<td>.26</td>
<td>1.13</td>
<td>2.53</td>
<td>6.38</td>
<td>-1.53</td>
<td>.05</td>
<td>-3.55</td>
<td>.49</td>
</tr>
<tr>
<td>Free</td>
<td>78.98</td>
<td>20.14</td>
<td>64.56</td>
<td>19.32</td>
<td>2.26*</td>
<td>.05</td>
<td>.25</td>
<td>4.27</td>
</tr>
</tbody>
</table>

Note. CI = confidence interval; LL = lower limit; UL = upper limit; *significant difference (p = .05).

Additional Results

To understand more fully the differences between groups not detected by the computer, this section offers additional information about the FSA-R drawings. Although the computer detected differences in quantity, it was unable to detect the magnitude of these differences. For instance, cases existed where differences in means existed in tenths of a percent. While the computer effectively distinguished the groups in such cases, it would require human observation to see how little this difference actually was. The differences in additional formal elements (Gantt & Tabone, 1998) are apparent when comparing samples from the MDD experimental group with samples from the control group (see Figure 1). The differences included color, line, form, details, emotional tone, and symbols.

Color. Many of the results from the experimental group appeared either monochromatic or dichromatic. The limited use of color in artistic expression often indicates a general disdain for color (Gage, 2000). This means that the artist reports dislike for arrays of color, and generally avoids using colors from a variegated palette. Heavy use of brown permeated the experimental group results. Brown is diametric to creativity and sensation (Lüscher, 1971). It is important to note this because those with Major Depression often experience a loss of perceptual sensation.
The shading included varieties of brown seen in the Crayola Multicultural© markers, but more typically—dark brown appeared prevalent. Many participants used dark brown for the hair color. However, some used it exclusively in the face and environment area of the FSA-R instrument.

Line. The quality of the line for most experimental group drawings appeared sketchy. The sketchy quality of line contrasts with straight, controlled line in that it often evokes a feeling of tension or anxiety (Leibowitz, 1999). The graphic nature of this line is forceful, pressured, and uncontrolled. In such instances, it is not unusual to see lines that are zigzagged or showing throw the back of the paper. This is in contrast to healthy individuals, who often draw smooth lines in a controlled manner. Most participants from the experimental group colored outside the line boundaries of the template and mixed colors more so than the control group, demonstrating lack of control. The psychomotor retardation (PR) tasks revealed similar results.

Form. The forms for the PR tasks in the experimental group results did not appear neatly represented and integrated, whereas the control group’s PR tasks did. For the experimental group, houses looked warped, images did not fit within the grid cells, and participants drew lines in each direction of the cued bisecting line task. The cue on the line templates was a number on the right side that guided the direction of the participant’s line (see Appendix G). Some participants in the experimental group left the tasks incomplete or blank, in contrast to the control group.

Details. There appeared to be far fewer details in the results of the experimental group drawings versus the control groups; these included eyes, face, and background details. However, the number of drawings featuring jewelry or accoutrements appeared the same for both groups. The experimental group did not draw as many ears on the face profile as compared to the control group, and not a single drawing from the experimental group showed a smiling face, whereas the control group featured several. The control group added more detail to the sky, including clouds, birds, and colored suns, whereas the experimental group featured less of these details. The control group added more ground line detail, such as grass, hills, trees, mountains, and flowers. No participants from the experimental group filled in the rainbow stimulus. While the experimental group as a whole did not elaborate on the flower stimulus template, the majority of control group participants colored it. The experimental group produced several images of a crying face, while no such detail appeared in the control group.
Emotional tone. The drawings of the control group appeared more positive than the experimental group, evidenced through exhibiting upturned smiles, lack of crying, lack of dark colors, and the addition of details such as birds, rainbows, and flowers. Several drawings from the experimental group showed themes of aggression, which included excessive use of reds (Silver, 2005). Other symbols of aggression (Rubin, 1997; Silver, 1996/2005) included knives stabbed into the face, lightning bolts (Gussak, 2007), and stitches. A couple drawings gave the sense of someone wounded, as bandages and scars plastered tear-strewn faces. The impressions that come to mind when viewing the lot of these drawings is that they were sad, futile, hopeless, wounded, angry, uncertain, and empty. Malchiodi (1998) also used these words in her analysis of free drawings completed by those with depression. The sadness resulted from down turned smiles. Futility and hopelessness emanated from the general incompleteness of the drawings, as though someone simply “gave up.” The general lack of direction in the lines of the drawings evoked a sense of uncertainty, and the fact that so much free space dominated the images could potentially leave the observer feeling empty.

Symbols. The most prominent symbols in the drawings of the experimental group were stitches and question marks. Question marks usually occurred in a dialogue bubble, also absent in the control group drawings. Several drawings from the experimental group featured lightning bolts. One individual from the experimental group drew a picture of what appeared to be a family holding hands, but then crossed it out. Symbols of death permeated the drawings of those with MDD, as noted in past observations by other authors. These symbols included the following: a skull (Malchiodi, 2007), knives (Silver, 2005), blood (Hass-Cohen, Kaplan, & Carr, 2008), bullet holes (Malchiodi, 1999), stitches (Wadeson, 2000), black birds, and a black sun (Hammer, 1981).

Throughout history, civilizations used the skull as an artistic symbol of death (Tresidder, 2005). The skull often represented the end of mortality and the death of the flesh. Though it occurred only a few times in the drawings of those with MDD, it could be significant to the drawings of those presenting with depression.

The knife is often a symbol of an aggressive means of obtaining power and control (Turner, 2008). This, however, is just one possible interpretation. Throughout many civilizations, it became a symbol of utility, or was depicted in records of early hunting.
The more modern symbol of aggression or hurt may come in the form of bullets, or bullet holes. Bullets often represent warfare or hurt (Kadish, 1983). The effects of bullets could bring depression, as in the aftermath of a war.

The high occurrence of black symbols seen in the experimental group may represent the depression felt (Silver, 1988/1993). Black is part of a duality, and could represent the negative aspect of nature (Jung, 1968). However, it is important to note that not all cultures view black as a negative symbol of life. In some countries, it is used in positive ceremonies (Gage, 2000).
Conclusion

The basic purpose of this study was to determine if those with Major Depressive Disorder (MDD) drew different levels of formal elements than those without MDD. Computerized technology, combined with the FSA-R instrument, tested the hypothesis of differentiating experimental groups with MDD and control groups without MDD. Quantitative results yielded differences in the purple, orange, and free space variables. Computerized results demonstrated timesaving qualities and the efficacy of the Gaussian blur filtering technique.

Because computers are incapable of analyzing all variables present in artwork, a section of this chapter outlined additional results that were qualitative in nature. For color, dark brown appeared more prevalent in the experimental group. Line quality in the experimental group appeared sketchy in contrast with smooth lines apparent in the control group. There appeared to be less integration in form for those in the experimental group as opposed to the control group. The experimental group featured less detail overall, compared to the control group. Emotional tone appeared more negative in the experimental group than the control group. Lastly, symbols of death and depression permeated the experimental group drawings, while featuring less
positive symbols than the control group. The next chapter describes all the measured variables as they relate to similar studies, and discusses the overall results of this study.
CHAPTER 5
DISCUSSION

This chapter presents a discussion of the results. This discussion is important because it evaluates the implications of the results in respect to the original hypothesis in the context of weaknesses and strengths. An examination of validity and reliability is a central component to any study regarding test construction (Fishman & Galguera, 2003). The quantitative results are discussed first, followed by computerized results, and additional results. There are existing studies that relate to the quantitative, computerized, and additional results of this work; they are the most pertinent to the actual results. This chapter then includes an overview of the possible threats to the validity and reliability encountered during the initial FSA-R administration, followed by their potential remedies. Finally, potential meanings extrapolated from the overall results conclude this chapter.

Quantitative Results

Many statistical procedures yield inaccurate results when the sample sizes are below 20 (Grissom & Kim, 2005). This is also true in the case of calculating confidence intervals. The intervals in this study appeared wide relative to capturing the population mean, the samples did not include a significant amount of individuals, especially those with varied ethnicity. It is difficult to ascertain the accuracy of the confidence intervals at this time.

The benchmark for interpreting Cohen’s $d$ effect sizes (small, medium, and large) was meant merely as a guide. Researchers need to take into account other aspects of the experiment before making a true judgment of effect size. If results are replicable, then even a small effect size, such as .2, may be significant (Cohen, 1988). Conversely, large effect sizes, as seen in the results for the variable purple ($d=-.91$), may be inflated from instrument response bias (Thorndike, 1997), outliers (Grissom & Kim, 2005), and methodological error (Trusty, Thompson, & Petrocelli, 2004). Many possible methodological errors could have skewed the quantitative results of this study. They take the form of threats to overall instrument reliability and validity subsequently discussed, along with possible remedies.

For the large effect size of purple ($d = -.91$), and the medium effect sizes of orange ($d = -.70$) and free space ($d = .73$), it is possible that they are inflated because statistical outliers were left in the calculations. Other possible reasons for these figures could stem from response bias of the healthy control group in relation to those with Major Depression. Those that were healthy
simply exhibited more energy in completion of the drawing tasks, which registered as larger percentages of color and less free space. These variables were considered when determining the overall results. All statistical inference, including this one, contains subjective judgment (Huberty & Morris, 1988).

**Studies Related to Quantitative Results**

The ANOVA results of the Hacking et al. (1996) study showed differences between those with depression and those without for the color variable orange \( n = 9, p < .02 \). Cronbach’s alphas for orange rated by six independent raters and an experimenter was .99 and .98, respectively. The ANOVA results of the follow-up study revealed those with depression significantly differing from controls in the use of free space, \( F(28,9) = 2.86, p = .05 \) (Hacking & Foreman, 2000).

Another main study that offered similar results was a study conducted by Dawson (1984). The drawings of those with depression \( n = 50 \) scoring on spectrums of the BDI similar to those seen in the FSA-R exhibited less color and more free space. Like Dawson’s study, the group with depression in the FSA-R results left more free space in their drawings and excluded more details than the group without depression.

There are several studies not supporting the results of the FSA-R. The study using the Shepherd Pratt Art Rating Scale (SPARS) to quantify artwork revealed almost no differences in color usage between depression and control groups (Bergland & Gonzalez, 1993). The intraclass reliability coefficient for two raters for color was .83 \( p = .001 \). The Draw-a-Person (D-A-P), paired with the Beck Depression Inventory (BDI), revealed slight differences in drawn figure size in those with depression, but not in color or free space variables (Holmes & Wiederholt, 1982). Nolan, Dai, and Stanley (1995) studied a group of undergraduate students \( n = 72 \). A portion of this sample presented with depression. Using the BDI, they discovered that those scoring above a 10 (minimal depression) preferred the color black. This was not the case for the results of the FSA-R pilot study, where those scoring above a 10 used little black.

**Computerized Results**

The time taken to analyze the images averaged 4 minutes. Though expedient, an inter-rater study using human raters and computers will need to be done to give context to this time. As described in subsequent sections, researchers conducting similar studies (Kim, Bae, & Lee, 2007; Kim, 2008) reported analysis times in seconds, largely because of the automatic functions.
of their computer programs. In general, studies in rating artwork through computerized methods appear to save time over conventional hand scoring (Mattson, 2010).

This study used Gaussian blur filtering, though many other types of filters exist that could yield different results (Russ, 1997). The sigma value used for the Gaussian blur filter was 1.0, though this is variable. Different results may be obtained through different values, and this study did not explore them. The brightness histograms for the Gaussian blur filter often revealed clipping, which is a form of noise that should be addressed in additional studies. It is possible that the clipping in the scanned images resulted from the prominent whites in the images, though further study will be required to find the true source of this noise. The region of interest (ROI) tool is also variable; this study used a tool pixel size of 318.56 only because this value fit within most drawing regions.

Studies Related to Computerized Results

Previous studies support the computerized results of this study. Image processing in this study used established computerized techniques. Russ (1997) produced brightness value histograms of Gaussian blur filters similar to those observed in this study. Kim, Bae, and Lee, (2007) used median filters in their expert systems. The Gaussian blur filter is similar to the median filter in that it is just as adaptable and works in a similar manner (Russ, 1997). Recent experimentation with filters revealed that the Gaussian blur produced more efficient results than the median filter for processing art-based images, given the PDIAS programs used (Mattson, 2010).

Hedges (2006) also used ImageJ to analyze artistic images, but for dating, not assessment. Regardless, Hedges used techniques similar to those used in this study. Hedges also observed that minimal filtering of artistic images is crucial to preserving the original artwork. The time it took to process the images in the FSA-R study compared closely to the results produced by other researchers.

For placement and color-related formal elements in the D-A-P, an expert system required 20 seconds to scan each image and 45 seconds to analyze it (Kim, Bae, & Lee, 2007). For determining basic and main colors within crayon drawings from two case drawings, the Expert System for Diagnosis in Art Psychotherapy (ESDAP) required an average of 35 seconds for scanning and 42 seconds for analysis (Kim, 2008). These results compare to the time averages
recorded in the FSA-R study for scanning, although analysis through the image analysis techniques used in this study required more time than expert systems.

**Additional Results**

The results that were qualitative in nature supplemented the quantitative results. The decision to retain the qualitative results arrived from judging the notable differences in color, line, form, detail, symbol, and emotional tone. It was clear even from unaided viewing that the images discriminated the groups.

The experimental group used less color, less details, sketchy line, non-integrated form, symbols associated with death and depression, and negative emotional tone. The control group used more color, more details, integrated the forms, used symbols representing positive themes, and tended to exude more positive emotional tone. The abundance of these qualities in the control group appeared evident in contrast with the experimental group when placing the drawings next to each other (see Figure 1).

**Studies Related to Additional Results**

Previous studies supported additional results of the FSA-R. Silver (1988, 1993) noted recurrent themes of aggression within the drawings of those presenting with depression. Within these themes, the object of the knife or wound site was prominent.

The D-A-P directive with depression drawings sometimes revealed a general lack of detail in the face, including an absence of mouth, eye, or ear (Naglieri et al., 1991). This was also seen in the MDD results of the FSA-R. Wadeson (1971, 1975) and Malchiodi (1998) explained a lack of detail in numerous free drawings completed by those with depression. Wadeson (1980) also viewed the images as appearing hopeless and isolated in terms of emotional tone.

Many of the MDD participants of the FSA-R did not complete the templates. A recent study also noted this same effect (Zomet et al., 2008). When the individuals did fill in the templates, they tended to blur the colors across lines, an action seen in past research (Simon, 1992).

Studies supported the observation that the quality of line in those with MDD appeared hesitant and sketchy (Gantt & Tabone, 1998; Hutt & Briskin, 1960; Pier et al., 2004a; Pier et al., 2004b; Wadeson, 1980). This often took the form of psychomotor retardation (PR), a prominent symptom in those with depression. The presence of PR is a diagnostic aid for those with MDD. There is support for the cuing of the line bisection task seen in the FSA-R results. Using the
manual line bisection task (Halligan & Marshall, 1988), researchers discovered that those with depression significantly differed from the controls in this task with the line cue on the right side, at an alpha level of .05, $F(1, 27) = 7.96, p = .009$ (Cavézian et al., 2007).

Some studies did not support the additional results of the FSA-R. Lewinsohn and May (1963) attempted to discriminate figure drawings of those with depression and those without, based on frowning or negative face expressions. The results did not differentiate the groups.

Dawson (1984) investigated differences between the drawings of those with depression and those without. The group with depression excluded more details than the group without depression. These details included environmental features, like variations in clouds and birds. The differences, however, were not significant.

Despite support, the main outcome of the FSA-R pilot study is provisional. Not all variables of color distinguished the two groups. Internal and external factors are likely responsible for this. Internally, a main reason lies in the instrument itself. Some of the items may not have elicited sufficient color usage, and require revising. One possible revision may include the addition and deletion of items, explained in the subsequent chapter. Another theory, proposed by Gantt, is that paint actually results in more color coverage than markers, and paint could become a possible medium to use for the FSA-R in the future (Williams et al., 1996). Externally, since this is one study with a small sample size, results are not definitive. A larger sample captured during the peak of MDD onset would likely reveal improved statistical results (Williams et al.). These and other threats to reliability and validity permeated this study.

**Threats to Reliability and Validity**

**Reliability**

The most pertinent threats to the FSA-R study stemmed from the experimenter effect, subject and observer reliability, instrument reliability, and data processing reliability (Robertson & Williams, 2009). Though many factors affect reliability, these appear to be most applicable to the development of the FSA-R. Constructing a reliable instrument depends on factors within the instrument itself, along with all the processes surrounding its development.

As the principal investigator, I did not have a chance to train all the administrators. As such, variations likely occurred in the directions given to the participants at different sites, regardless of issuing standardized manuals and protocol. This is an example of the experimenter effect. This is a clear threat to the reliability of the pilot administration (Pintrich & Maehr, 2002).
In addition to this, materials could be mishandled, and testing centers presented with slightly differing variables, such as room temperature, and seating arrangement.

The reliability of both the participant and the administrator also comes into question. Those with Major Depression sometimes experience low energy levels, fatigue, irritability, and many other symptoms that could present with differing graphic outputs (American Psychiatric Association, 2000; Gantt & Tabone, 1998). Those in the control groups could present with varying levels of mood. The administrator is also prone to these symptoms as part of normal human variance in mood. Researcher bias possibly entered into the administration that the principal investigator conducted.

This study did not include tests of instrument reliability. It is important to note that existing computer analysis of art has not been evaluated through inter-rater reliability (Whiston, 2005). Errors could be present in the form of faulty items, directions, or procedures—all factors in affecting instrument reliability. Even after the collection of the data, it is possible to miscode the data, or process it in an inconsistent manner. There were several instances where the statistical procedures used to calculate the color percentages were not followed in exact order.

Validity

Internal validity. There were several threats to the internal validity of instrument experimentation. Those that pertained to this study include selection bias, history during testing, instrumentation, mortality, and subject effect (Ary et al., 2009). As principal investigator, I targeted subjects in existing treatment groups or controlled settings. A better practice would be to randomize the samples. This would likely resolve many of these threats. Additionally, most selected in each group were white. This study did not capture a multicultural sample; the results did not represent ethnic diversity. The main implication for this is that the FSA-R will not currently be suitable for those from ethnic backgrounds.

During the actual testing, some changes that occurred in one group may not have occurred in another. Some variables were not controlled during administration. For instance, in the clubhouse setting, the temperature was 68 degrees Fahrenheit when it was supposed to be 72 \( \pm 2 \) degrees Fahrenheit. The administrator at the community college did not read the scripts in their entirety. Many participants talked and asked questions during the administrations, some walked in and out of the room during testing.
Administrators noted distractions in the community college setting and the clubhouse setting that were not present in other groups. While testing, the actual instrument could have affected validity. The FSA-R was built in part from items featured in haphazardly constructed art-based instruments (Betts, 2005a). A most likely cause of inconsistency in the instrument could also be the order in which some group members turned the pages.

Some members opted out of testing, and chose to sit idly in place of testing. This reflected a small level of mortality, nonetheless important to the validity of a study. Fortunately, attrition was not high enough to threaten the integrity of the study. The subject effect, however, was notable in many of the qualitative observations during testing. Some of the attitudes of participants reflected disinterest or uncertainty, particularly in the experimental group. Many participants of the experimental group reported feeling neutral about the testing experience in the post-interview forms.

External validity. Threats to external validity of this study included convenience sampling, interaction of personal variables, multiple treatment interference, Hawthorne effect, Novelty effect, and interaction of history (Houser, 2009). The low power of the sample size likely affected the validity in that it was not adequately representative of the population; this increases the chance of a Type II error (Henry, 1990). Sampling came from individuals chosen at convenience. Personal variables such as artistic ability possibly interfered with drawing results of either group. Most individuals participating received pharmaceutical treatment and psychological treatment, which likely affected the amount of depression experienced at the time of drawing. Individuals who thought they were being tested may have tried to improve their results, spurring the Hawthorne effect. The Novelty effect of the drawing design and computer rating may have also increased performance, and the history of each individual would have likely affected outcome.

The results of this research should eventually join forces with the work of Kim, Bae, & Lee (2007), Hacking and Foreman (2001), Gantt (1998) and Betts (2003) to form a measure of convergent validity. It is usually a good practice to combine pilot instruments with established instruments. This practice could help to establish further credibility of art-based instruments that make use of computerized rating scales.
Remedies to Threats

Reliability

The administrators of the FSA-R offered input on changing the forms in the future; this could assist in increasing instrument reliability. During the first administration to an experimental group, the administrator noted that the cue to start the BDI-II was not present in the script. The introductory script seemed too formal, and requires revision. Filling out the appendices in the administrator manual proved difficult. The checklists and observational forms require cropping, as many parts of the completed forms remained blank after testing.

To address the experimenter effect and observer effect, future research should involve standardized training for all administrators. Additionally, the forms for the FSA-R require revision to make them more consistent. The demographics form is improvable in the future. It ought to contain question stems that inquire about specific diagnoses and personality disorders, as these were a diagnostic axis not mentioned, which could have affected the results.

Many measures exist to improve instrument reliability. Future experimenters should employ reliability measures on the intra- and inter-rater levels of computerized rating methods (Mattson, 2010; Whiston, 2005). A sensible technique to accomplish this is to use human raters in conjunction with the output of the computerized programs (Mattson & Veldorale-Brogan, 2010).

Validity

Internal validity. To remedy selection bias in the future, the experimenter should have a system of recruitment that is independent of personal judgment. This would entail random selection of individuals for the groups. Both control and experimental groups should be randomized, and run by experimenters other than the creators of the instrument. The creators of the instruments rated by computerized assessment programs are often a source of bias because they are the sole experimenters of the programs (Garb, 2000; Mattson, 2009).

One of the most effective ways to control the influence of history during actual testing is to attenuate extraneous variables (Evans & Rooney, 2008). Effective ways to do this would be elimination of certain variables from the testing environment. Additionally, the researcher could hold more variables constant, as there is room for additional constancy in the FSA-R protocol.

The FSA-R requires further statistical work for it to become a refined instrument. Analysis should include a focus on internal reliability. During analysis of the FSA-R, several
issues arose. For one, the more filters used, the more distorted the analysis. The current version of MVHimage (version 8.0) allowed for analysis of TIFF images, while older versions did not. Also, during experimentation, a single iteration of the Gaussian blur filter resulted in more accurate analysis than the 3 x 3 median filter so commonly used in art analysis (Kim, Bae, & Lee, 2007; Mattson, submitted). The individual items should undergo analysis to determine difficulty, correlation, and discriminability (Fishman & Galguera, 2003). The process of addition and deletion of items should be ongoing.

To lessen the impact of mortality, the experimenter should include larger sample sizes in the future (Ary et al., 2009). In accordance with Institutional Review Board (IRB) protocol, participants will always be allowed to refuse taking the FSA-R at any point, and have the option to participate in free drawings in lieu of the FSA-R. Some mortality in experimentation is expected.

An effective means of reducing the subject effect is to conduct double-blind studies (Straub, 2006). The FSA-R researcher could combine the MDD experimental groups with the control groups, and use passive deception (Christensen, 2001) in the explanation of the experiment. This ensures that both groups will be blind to each other. This could improve internal validity.

External validity. Many of the samples in both the control and experimental groups presented with multiple diagnoses and comorbid conditions, which could affect external validity (Weiner, Freedheim, Schinka, & Velicer, 2003). Additionally, samples did not undergo screening for personality disorders because such disorders could have also tainted the sample (L. Annis, personal communication, April 4, 2010). Better samples could result from improved methods of recruitment, such as randomization, and advertisement targeting those specifically with MDD. Randomization is preferable to convenience sampling, and should constitute the main recruitment method in future administrations (Christensen, 2001). Despite randomizing samples, it is difficult to account for all the possible interactions of personal variables that could affect the ecological processes of testing. There may be multiple variables interacting through countless combinations.

Consider, for instance, multiple treatment interferences. The participants in all FSA-R groups often reported being on some form of medication; this possibly affected the graphic results of those with MDD, as studies have shown this to be true in some instances (Pierre et al.,
Additionally, participants could have been in different stages of treatment for their illnesses, since many of them came from voluntary groups.

Some participants were subject to the Hawthorne effect. Control groups sometimes desire to outperform the experimental groups, or vice versa. Fortunately, this pilot study addressed the Hawthorne effect in two main ways that should continue in subsequent administrations. For one, the consent form did not disclose the true purpose of the study. Only in the debriefing afterwards did the participants understand the goal of the study. In addition, participants of each group were not aware of the fact that this was a multi-site study. Therefore, participants in the control group would likely not compete with the experimental group for better scores.

The novelty effect could be a threat to the measurement of the dependent variable. This effect is particularly potent if the administration is short-lived. The way to remedy it is to extend administration time in the future to allow for the true influences of MDD on graphic representations (Houser, 2009).

Because administration locations of the FSA-R were diverse, the interaction of history on measurement is a real possibility. Events that occurred around the time of administration could have affected the ecological validity of this pilot test. Controlling more variables in subsequent administrations could remedy this.

Possible Meanings

This study used objective computer programs to successfully rate artwork. Technological and quantitative components detected differences in group artwork. This study showed that quantitative differences existed between those with mental illness and those without.

This could mean that computer programs are a viable means of rating art-based instruments, and there are, in fact, notable differences in formal elements between those with a mental illness designation and those without, as detected by these programs. Because of the objectivity and standardization employed, this study marks an initial step in scientifically affirming the long-held notion that there is a link between mental illness and art (Batey & Furnham, 2008; Jamison, 1993; Lombroso, 1891/2010; Nettle, 2006; Preti & Vellante, 2007).

This study also demonstrated expedient rating techniques. Computerized analysis often saves time over conventional hand scoring (Kim, Bae, & Lee, 2007; Kim, S. I., Kang, & Kim, 2008). Implications for this include saving time on scoring assessments so that more time can be devoted to client care (Mattson, 2010).
Furthermore, the face of clinical assessment is changing. Out of the 10 most popular psychological tests used by clinicians, nearly half exist in computerized form (Lubin, Larsen, & Matarazzo, 1984). The future of psychological assessment depends on computerized technology (Groth-Marnat, 2000). The FSA-R pilot is an initial study in computerizing art-based instruments. The results of this study are congruent with the digital revolution our society is currently experiencing (Parsons & Oja, 2010). The technology used in the computerized assessment of art is part of the growing constellation of advancements in computers, communications, and digital media.

**Conclusion**

This chapter discussed the results of the FSA-R pilot study. The results are quantitative, computerized, and additional to the study. There are studies that relate to the results. The effect sizes of the quantitative results, though high, may be inflated due to multiple errors. An examination of validity and reliability revealed areas of standardization and potential weaknesses. Possible remedies to the threats of validity and reliability followed, as these are central components to test construction.

The results of this study imply several possible meanings. The computerized technology used in this study mark an initial scientific route to discovering mental illness traits in artwork. The technology used for this purpose is congruent with the new technological revolution currently faced in our society. The next chapter provides an overview of the background, technology, and methods used in this study, and answer the original research questions posed at the beginning of this work.
CHAPTER 6
CONCLUSIONS

“[When conducting a pilot study] whatever can go wrong, will go wrong.” (Fishman & Galguera, 2003, p. 100).

Previous chapters outlined the construction of the FSA-R instrument. The literature review supported the individual items comprising the FSA-R through an examination of graphic indicators commonly seen in those presenting with depression. The construction of the FSA-R incorporated established theoretical tenets and standardized methods that extended into the experimental phase of the study. Subsequent chapters featured the results and discussion of the results as they related to the hypothesis.

This chapter offers conclusions on the study. It begins with answers to the research questions originally proposed at the beginning part of this work. Limitations of the study follow the questions, and future considerations for the FSA-R instrument are discussed.

Research Questions Addressed

Question One

Can PDIAS rate the formal elements of color and/or free space to distinguish artwork of those with Major Depression from the artwork of those without Major Depression?

The results of this particular study supported the hypothesis that color and/or free space distinguished the Major Depression group from the control group, and can be accomplished through the assistance of PDIAS programs. It is possible that because of the accuracy of computerized methods (Sonka & Hanson, 2001; Kim, Bae, & Lee, 2007; Kim, Kang et al., 2009) differences in the groups became more apparent than if mere human perception rated the images. With such miniscule differences in statistical operations calculated by the computer to differentiate groups, even minor errors in human judgment could skew these results.

Computerized assessment reduces human error, saves on time, is easy to use, allows for early detection of conditions, improves statistical measures, and is resilient amid healthcare reform (Mattson, 2010). For these reasons, it offers advantages over the exclusive use of subjective assessment. Furthermore, the variables that did not distinguish the groups could distinguish them, if this study employed differing levels of probability through the medium of this computerized technology.
The results supported the hypothesis, yet are far from conclusive, as they stem from a small sample size within the framework of a single pilot test. Additionally, this is only one study. In regards to this question, more work using similar methods yielding results in support of this study is the viable answer at this time.

**Question 2**

Can the FSA-R instrument become a standardized instrument for Major Depression?

Viewing that significant differences emerged in this study, it is quite possible that the FSA-R can screen those with Major Depressive Disorder in the future. This study used standardized methods of testing. Standardization occurred throughout the entire process of construction. All administrators received the same materials and the same directives.

Despite all efforts to standardize completely the instrument, some components require more work. For instance, the training the administrators received was inconsistent. The manufacturers of the markers discontinued the colors yellow and red from their packets halfway through the administration process. Not all rooms where the administrations occurred were alike. It is probable that the FSA-R instrument can become a standardized tool for screening Major Depression, but future studies should involve more stringent standardization practices.

**Question 3**

Do common graphic indicators exist amongst the artwork of those with Major Depression?

To address this question, one needs to consider also the qualitative interpretation of the drawings that the Major Depression group completed. Although the most effective studies in art therapy assessment contain an objective component, it is also necessary to include components of subjective input (Betts, 2005; Mattson, 2009). A survey of the completed instruments that was qualitative in nature revealed differences particular to the Major Depression group.

These included an extensive use of monochromatic and dichromatic color applications, particularly with shades of brown. This prevalence did not show in computer analysis, but through subjective judgment. In terms of line or marking, a sketchy/hesitant quality did exist throughout the drawings of those with Major Depression more so than the control group. This is consistent with previous findings (Hutt & Briskin, 1960; Burns & Kaufman, 1970; Hegerl et al., 2004; Pieret et al., 2004a; Wadeson, 1980).
Concerning other graphic indicators, the wound stitch and the question mark appeared in several drawings, but not enough to be significant. Hammer (1981) once noted phenomena of recurring graphic signs within the artwork of those with particular mental illnesses. These signs, however, were rare, but Hammer continued to assert that they could be particular to mental illnesses and required attention, and so he dubbed them rare signs. It is possible that such signs were particular to the drawings of those with Major Depression, but were too rare to register in a small sample of such small magnitude. The area of rare signs requires additional work in the future. The results of this study are inadequate to answer the general nature of this question.

**Question 4**

Once completed, can components of this computer-rated instrument generalize to other assessments, or populations?

There are two approaches to explaining the possible generalizability of the FSA-R—qualitative, and quantitative. Many existing art-based instruments are offshoots of predecessors (Burns, 1987, Knoff, & Prout, 1985; Koppitz, 1968), and Gantt’s (1998) rating scale now generalizes to the FSA (Hamilton, 2008). It appears that only minor alterations to original instruments is necessary for most generalizations, such as using the second drawing in the FSA series for rating by the FEATS, or adding kinetic components to the series drawings from the H-T-P. In the case of generalizing PDIAS rating to other existing art-based instruments, it is possible because these other instruments can be digitally resized for analysis, receive the same directive under similar demographics, and contain or incorporate items from artwork samples of those with Major Depression. The PDIAS programs are capable of generating data on most all forms of artwork.

Quantitatively, after the pilot study, regression analysis of this data will yield possible future scores of the FSA-R within a certain margin of error. Regression analysis, in such an instance, will forecast test scores (Freedman, 2005). This will aid in determining cutoff scores for other populations taking the instrument. In response to this question, it is possible that components of the FSA-R can generalize to other art-based instruments and populations. However, further work in varied applications of the instrument is needed to solidify this claim.
Limitations of the Study

Several types of limitations exist in this study. Previous chapters discussed aspects of these limitations. These include limits in the following areas: computerized methods used in assessment, standardization, inter-rater reliability, and validity.

Limits of Computerized Assessment

Because the computerized assessment of art-based instruments is a relatively new component of art therapy, it presents with a unique set of limitations. The limitations are: (a) the prototypical nature of these programs, (b) validity concerns, (c) an inability to account for the vast array of intake variables, (d) bias, (e) Barnum effect, and (f) confidentiality (Mattson, 2010).

All programs used to analyze artwork are currently prototypes (Kim, Bae, & Lee, 2007; Kim, Kang et al., 2008). As such, they do not exhibit appreciable levels of validity (Lichtenberger, 2006). Even with current advances in technology, computers lack the ability to account for all variables that a client presents with (Grove & Meehl, 2006). Bias stems from the software developers being the only individuals testing the software (Garb, 2000). The software users sometimes receive the impression that computers are infallible, a prime example of the Barnum effect (Matarazzo, 1983). With any computer-based form of testing, issues of security arise, especially as this pertains to client confidentiality, since individuals not related to a treatment team may have access to electronic results (Groth-Marnat, 2000).

Limits in Standardization

Despite rigorous attempts to preserve testing standards, this study did not succeed in achieving complete standardization. Not all administrators received consistent training (Pintrich & Maehr, 2002). Errors occurred in the order of instructions, variables remained uncontrolled, data may have been miscoded consistently, and the materials were sometimes inconsistent due to the manufacturer discontinuing certain colors within the marker sets at the halfway point of administering the FSA-R instruments.

Inter-Rater Reliability Studies

This study did not evaluate inter-rater reliability. Most studies involving computerized rating have not undergone studies of inter-rater reliability (Whiston, 2005), and this includes rating art-based assessments. A technique currently used to test the reliability of computerized programs involves a comparison between computer rating and human rating (Mattson &
Inter-rater reliability, along with validity measures, should follow the pilot stage of test construction (Fishman & Galguera, 2004).

**Validity Studies**

It is difficult to determine if the FSA-R test items measured what they should have, which is the quintessential trait of a valid instrument (Groth-Marnat, 2003). There are other explanations for the test results besides the distortional effect Major Depression has on graphic representation. For instance, the experimental group may have been more tired than the control group, since fatigue often affects the results of testing (Christensen, 2001). Other possible reasons include environmental factors, materials, or the effect of an observer in the testing room. It is difficult to evaluate truly the validity of the FSA-R instrument until further studies commence.

**Future Considerations**

The following are future considerations of this study. Additionally, each section addresses possible avenues for subsequent study of some of the areas of this study that require improvement. The future considerations involve test construction stages, population, item analysis, and technological improvements.

**Test Construction Stages**

The next stage in the construction of the FSA-R will include further revision of its items (Fishman & Galguera, 2004). After deleting items based on statistical analysis, the second administration occurs, and the test construction cycle will repeat (see Appendix C). Based on the results of the initial administration, the FSA-R instrument will likely benefit from a large number of future improvements. These include changes in population, amendments to forms, changes in methods, and adaptations of materials.

**Population**

To improve the norms in the future, the instrument will require a wider range of ethnicities taking the instrument on a national level. A nationally normed instrument is preferable to a regionally normed instrument (J. C. Hanson, personal communication, April 4, 2007). This study took small samples from only two regions of the United States. The larger the sample, the more the data curve approaches the normal distribution (Glass, & Hopkins, 1996), and thousands of participants support valid and reliable instruments (Millon, 1969). There should
be more ethnic dispersion in future samples. Ethnic considerations are paramount for the future direction of assessment (Betts, 2006; Prediger, 1994).

The population in this study also included multiple diagnoses. Although researchers can strive to obtain pure diagnoses of Major Depressive Disorder (MDD) in the future, there were several reasons for including multiple diagnoses in this pilot study. For one, based on surveying several sample pools from different areas in the country, the presence of dual diagnoses appeared to be the norm. This being the case, the samples chosen for this pilot study are actually more representative of the population. Secondly, variants of MDD, or those with secondary diagnoses, scored similarly on the BDI-II in terms of score means and score standard deviations, which is the main reason why Beck et al. (1996) combined all subtypes of depression in norming the BDI-II. The FSA-R similarly combined subtypes of MDD.

Item Analysis

The FSA-R post-interview resulted in a wide variety of feedback from the participants. The post-interview forms asked detailed questions about the nature of the instrument (see Appendix K). Administrators also commented on the process of giving the test to participants. Additionally, subject matter experts at the Florida State University (FSU) offered advice on the formation of new visual items. This panel offered visual question stems and possible answer options. From the results, variations on face gestures may be included in future versions of the FSA-R. Based on the results of the pilot study, initial item deletions will include numbers 6, 11, 19, and 26 (see Appendix F). Because of the low response rate and/or difficulty level, these items will not be part of the second administration. All groups involved in the pilot test offered valuable input that will shape the continued development of the FSA-R, particularly in terms of validity and reliability.

Technological Improvements

The techniques used in image analysis are constantly evolving. Increased sophistication in programming and the broadening of image analysis applied to art could result in better processing techniques. Such developments will eventually permit scoring of less structured assessments, such as free drawings (Whiston, 2005).

Eventually, transparencies based on norms could overlay the FSA-R template, thereby increasing the efficiency of hand-scoring (Mattson, 2008). Additionally, the FSA-R is capable of becoming a computer-assisted test (CAT) (Mattson, 2010). Participants will be able to draw on
an electronic tablet using a stylus, and results will show on the screen. The administrator could then rate and save the results of the instrument directly on the same computer, or from a remote location. This fits in directly with healthcare reform requirements, and the ever-expanding federal initiative to digitize patient charts (Mattson, 2010).

As computer technology evolves, so too will the equipment and software used in FSA-R analysis. Future research should include the use of computers with larger RAM and ROM, along with processors featuring greater speed and efficiency. These three main components—RAM, ROM, and improved processing—will increase computing power (Bansal, 2004). Hewlitt-Packard continues to release printer-scanner-copiers with improved features over the one used for this study.

Ultimately, the FSA-R should have its own image analysis software instead of using adaptations of existing programs designed for other purposes (Mattson, 2009). There may come a time when the test is completely computer-based; it may even have its own cell phone app available to professionals. There is much potential for technological improvement with this instrument. The FSA-R is adaptable to the surge of technological advancements permeating allied medical fields.

**Conclusion**

The results of this study supported the initial hypothesis. The research questions centered on the hypothesis were answered. The first question was answered by presenting the variables that distinguished the experimental group from the control group. Question two offered a provisional answer because standardization occurred throughout the study, but not completely. Question three explored the possibility of common graphic indicators drawn by those with Major Depression, and concluded that additional work will be needed to detect truly their presence. Question four asked about the possibility of generalizing procedures and results of the FSA-R instrument to other art-based instruments. The answer to this question is positive in regards to the scoring techniques, but concludes that additional work will be required for a complete answer.

The limitations to this study are: (a) the prototypical nature of these programs, (b) validity concerns, (c) an inability to account for the vast array of intake variables, (d) bias, (e) Barnum effect, and (f) confidentiality. The final section addressed these limitations through the stages of test construction, population, item analysis, and future directions.
Future research of the FSA-R should include further revision of the items, the population, and the image analysis software used to assess artwork. The FSA-R is adaptable to the rapid advancements in the digital technology of our current society. There is much potential for the FSA-R to become a standardized instrument for screening Major Depression, all in the context of today’s digital revolution.
## APPENDIX A

### TEST ITEM PROPERTIES

<table>
<thead>
<tr>
<th>Test Item Property</th>
<th>Graphic Remedy</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Definition of test Item:</td>
<td>“A test item in an examination of mental attributes is a unit of measurement with a stimulus and a prescriptive form of answering; and, it is intended to yield a response from an examinee from which performance in some psychological construct (such as knowledge, ability, predisposition, or trait) may be inferred.” (Osterlind, 1990, p.3)</td>
<td>Osterlind, S.J. (1998). <em>Constructing test items: Multiple-choice, constructed-response, performance, &amp; other formats</em> (2nd ed.). London, UK: Kluwer Academic.</td>
</tr>
</tbody>
</table>
## APPENDIX B

**DSM-IV SYMPTOMS, FEATS SCALES, & FSA-R CORRESPONDENCE**

<table>
<thead>
<tr>
<th>DSM-IV-TR Symptoms</th>
<th>Graphic Equivalents</th>
<th>FEATS Scales</th>
<th>FSA-R Graphic Item Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depressed mood</td>
<td>Lack of colors, dark colors</td>
<td>#1 Prominence of Color #2 Color Fit</td>
<td>5, 6, 25, 26, 33</td>
</tr>
<tr>
<td>Loss of energy</td>
<td>Constricted use of space</td>
<td>#3 Energy #4 Space</td>
<td>2, 3, 4, 10, 14, 20, 31, 32, 34, 35</td>
</tr>
<tr>
<td>Diminished interest</td>
<td>No environment Lack of detail</td>
<td>#7 Realism #10 Details #12 Person</td>
<td>17, 18, 19, 20, 21, 22, 24, 27, 28, 29, 30, 33</td>
</tr>
<tr>
<td>Diminished ability to think or concentrate</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Psychomotor retardation or agitation</td>
<td>Sketchy/Hesitant line</td>
<td>#11 Line Quality</td>
<td>1, 7, 8, 9, 11, 12, 13, 15, 16, 20, 23</td>
</tr>
</tbody>
</table>
APPENDIX C
SCHEMATIC OF INSTRUMENT DESIGN

I. Instrument Design

   Literature Review → Conceptualization of Components → Drafting of Item Pool → Version One

II. Version One

   Formatting → Pilot-Testing → Administration (1) → Selecting items and/or Drafting New Items
   → Item Difficulty (1) → Inter-item Consistency (1) → Item Discriminability (1) → Instrument Reliability (1) → Version Two

III. Version Two

   Administration (2) → Comparison with Version One—Selecting and/or Revising Items
   → Item Difficulty (2) → Inter-item Consistency (2) → Item Discriminability (2) → Instrument Reliability (2) → Version Three (if desired)
   → External Criterion Validity

IV. Version Three

   Administration (3) → Comparison with Version Two—Selecting and/or Revising Items
   → Item Difficulty (3) → Inter-item Consistency (3) → Item Discriminability (3) → Instrument Reliability (3) → Version Four... Version N (if desired)
   → External Criterion Validity

APPENDIX D

DSM-IV-TR CRITERIA FOR MAJOR DEPRESSIVE EPISODE

A. Five (or more) of the following symptoms have been present during the same 2-week period and represent a change from previous functioning; at least one of the symptoms is either (1) depressed mood or (2) loss of interest or pleasure.

(1) depressed mood most of the day, nearly every day, as indicated by either subjective report (e.g., feels sad or empty) or observation made by others (e.g., appears tearful). **Note:** In children and adolescents, can be irritable mood.
(2) markedly diminished interest or pleasure in all, or almost all, activities most of the day, nearly every day (as indicated by either subjective account or observation made by others)
(3) significant weight loss when not dieting or weight gain (e.g., a change of more than 5% of body weight in a month), or decrease or increase in appetite nearly every day. **Note:** In children, consider failure to make expected weight gains.
(4) insomnia or hypersomnia nearly every day
(5) psychomotor agitation or retardation nearly every day (observable by others, not merely subjective feelings of restlessness or being slowed down)
(6) fatigue or loss of energy nearly every day
(7) feelings of worthlessness or excessive or inappropriate guilt (which may be delusional) nearly every day (not merely self-reproach or guilt about being sick)
(8) diminished ability to think or concentrate, or indecisiveness, nearly every day (either by subjective account or as observed by others)
(9) recurrent thoughts of death (not just fear of dying), recurrent suicidal ideation without a specific plan, or a suicide attempt or a specific plan for committing suicide

B. The symptoms do not meet criteria for a Mixed Episode.

C. The symptoms cause clinically significant distress or impairment in social, occupational, or other important areas of functioning.

D. The symptoms are not due to the direct physiological effects of a substance (e.g., a drug of abuse, a medication) or a general medical condition (e.g., hypothyroidism).

E. The symptoms are not better accounted for by Bereavement, i.e., after the loss of a loved one, the symptoms persist for longer than 2 months or are characterized by marked functional impairment, morbid preoccupation with worthlessness, suicidal ideation, psychotic symptoms, or psychomotor retardation.

Adapted with permission from http://www.mental-health-today.com/dep/dsm.htm
# APPENDIX E

## STANDARDS FOR EFFECTIVE TEST DEVELOPMENT

<table>
<thead>
<tr>
<th>Steps</th>
<th>Example Test Development Tasks</th>
<th>Example Related Standards</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Overall plan</td>
<td>Systematic guidance for all test development activities: construct; desired test interpretations; test forms(s); major sources of validity evidence; clear purpose; desired inferences; psychometric model; timelines; security; quality control.</td>
<td>Standard 1.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Standard 3.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Standard 3.9</td>
</tr>
<tr>
<td>2. Content definition</td>
<td>Sampling plan for domain/universe; various methods related to purpose of assessment; essential source of content-related validity evidence; delineation of construct.</td>
<td>Standard 1.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Standard 3.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Standard 3.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Standard 14.8</td>
</tr>
<tr>
<td>3. Test specifications</td>
<td>Operational definitions of content; framework for validity evidence related to systematic, defensible sampling of content domains; norm or criterion referenced; desirable item characteristics.</td>
<td>Standard 1.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Standard 3.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Standard 3.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Standard 14.8</td>
</tr>
<tr>
<td>4. Item development</td>
<td>Development of effective stimulus formats; validity evidence related to adherence to evidence-based principles; training of item writers, reviewers; effective item editing; CIT owing to flaws.</td>
<td>Standard 1.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Standard 3.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Standard 3.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Standard 14.8</td>
</tr>
<tr>
<td>5. Test design and assembly</td>
<td>Designing and creating test forms; selecting items for specified test forms; operational sampling by planned blueprint; pretesting considerations.</td>
<td>Standard 3.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Standard 3.8</td>
</tr>
<tr>
<td>6. Test production</td>
<td>Publishing activities; printing or CBT packaging; security issues; validity issues concerned with quality control.</td>
<td>N/A</td>
</tr>
<tr>
<td>7. Test administration</td>
<td>Validity issues concerned with standardization; ADA issues; precoring; security issues; timing issues.</td>
<td>Standard 3.18</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Standard 3.19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Standard 3.20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Standard 3.21</td>
</tr>
<tr>
<td>8. Scoring test responses</td>
<td>Validity issues; quality control; key validation; item analysis.</td>
<td>Standard 3.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Standard 3.22</td>
</tr>
<tr>
<td>9. Passing scores</td>
<td>Establishing defensible passing scores; relative vs. absolute; validity issues concerning cut scores; comparability of standards; maintaining constancy of score scale (equating, linking).</td>
<td>Standard 4.10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Standard 4.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Standard 4.19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Standard 4.20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Standard 4.21</td>
</tr>
<tr>
<td>10. Reporting test results</td>
<td>Validity issues; accuracy, quality control; timely; meaningful; volume issues; challenges; ranks.</td>
<td>Standard 8.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Standard 11.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Standard 11.12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Standard 11.15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Standard 13.19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Standard 13.10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Standard 13.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Standard 6.4</td>
</tr>
<tr>
<td>11. Item banking</td>
<td>Security issues; usefulness, flexibility; principles for effective item banking.</td>
<td>Standard 3.1</td>
</tr>
</tbody>
</table>

*Abbreviations: ADA, Americans with Disabilities Act; CBT, computer-based testing; CIT, consistent-test-retest variance.*

APPENDIX G

REFINED VERSION OF THE FSA-R USED IN THE STUDY
APPENDIX H

CONSENT FORM

Face Stimulus Assessment-Revised
Consent Form

Name_______________________Date______________

1. **Research:** This is a research project, and should not be considered a health care procedure. However, secondary benefits of this project may include a chance for you to create a drawing. For humanity, information gathered from this project may benefit the allied healthcare field of art therapy through improving assessment procedures.

2. **The purpose of this research:** Is to review, evaluate, and compare drawings to assess your creative potential.

3. **Description of procedures:** You will receive a series of forms to fill out, followed by a drawing activity. You will draw within a designated space of paper upon listening to a set of instructions. Overall, you will be given 30 minutes to complete the activity.

4. **Risks and discomforts that may be reasonably expected:** It is unlikely that you will feel discomfort during this experiential. However, if discomfort is felt at any time, you may stop and discuss concerns with issuer.

5. **Alternative treatments:** If you not wish to participate in this study, you may decline, or draw freely without directives.

6. **Privacy and confidentiality:** The results of this study, including the pictures produced, may be published for the information of others without any identifying information aside from the facility in which it was conducted. Specific identifying information (e.g., your name, social security number, address) is not required of this study. This will be a group activity, so other individuals you may or may not know may be present. No other group member shall view or discuss your work. You shall not view or discuss the work of others.

7. **Contact:** In the event of research-related injury, please notify Donald C. Mattson.

8. **Voluntariness:** You have the right to withdraw from this study at any time. Refusal to participate will not result in penalty or loss of benefits.

I understand that a committee of clinical and non-clinical people periodically review and approve this research for scientific and ethical merit. I will be told of any future information that could affect my willingness for ongoing research. I may leave the research at any time. Such a decision will not affect my future with this, or any other organization.

I acknowledge that I have fully reviewed and understood the contents of this consent form. A copy of this consent is available to me if I request it.

Signature_______________________Date____________________

For additional concerns or questions, please contact:

Human Subjects Office
2010 Levy Avenue
Suite 276-C
Tallahassee, FL 32306-2743
Ph: (850) 644-7900
Fax: (850) 644-4392

Department Chair
Dr. David Gussak, PhD, ATR-BC
Chair, Department of Art Education
Florida State University
301 Francis Eppes,
Tallahassee, FL 32306
Ph: (850)-645-5663
APPENDIX I

PERMISSION FOR STUDY

Office of the Vice President For Research
Human Subjects Committee
Tallahassee, Florida 32306-2742
(850) 644-8673 · FAX (850) 644-4392

APPROVAL MEMORANDUM
Date: 2/16/2010
To: Donald Mattson
Dept.: ART EDUCATION
From: Thomas L. Jacobson, Chair
Re: Use of Human Subjects in Research
Face Stimulus Assessment

The application that you submitted to this office in regard to the use of human subjects in the research proposal referenced above has been reviewed by the Human Subjects Committee at its meeting on 02/10/2010. Your project was approved by the Committee.

The Human Subjects Committee has not evaluated your proposal for scientific merit, except to weigh the risk to the human participants and the aspects of the proposal related to potential risk and benefit. This approval does not replace any departmental or other approvals, which may be required. If you submitted a proposed consent form with your application, the approved stamped consent form is attached to this approval notice. Only the stamped version of the consent form may be used in recruiting research subjects.

If the project has not been completed by 2/9/2011 you must request a renewal of approval for continuation of the project. As a courtesy, a renewal notice will be sent to you prior to your expiration date; however, it is your responsibility as the Principal Investigator to timely request renewal of your approval from the Committee. You are advised that any change in protocol for this project must be reviewed and approved by the Committee prior to implementation of the proposed change in the protocol. A protocol change/amendment form is required to be submitted for approval by the Committee. In addition, federal regulations require that the Principal Investigator promptly report, in writing any unanticipated problems or adverse events involving risks to research subjects or others. By copy of this memorandum, the Chair of your department and/or your major professor is reminded that he/she is responsible for being informed concerning research projects involving human subjects in the department, and should review protocols as often as needed to insure that the project is being conducted in compliance with our institution and with DHHS regulations.

This institution has an Assurance on file with the Office for Human Research Protection. The Assurance Number is IRB00000446.
Cc: Dave Gussak, Chair
HSC No. 2009.3554
APPENDIX J

DEMOGRAPHICS FORM

Helpful Information

Date:_______________
Age:______________________
Date of Birth:______________
Zip Code:_________________
Please circle:

Sex:           M / F
Ethnicity:   White   Black  Hispanic  Asian   Native American   Other
Marital History:    Single    Married    Never Married   Widowed   Divorced
Education in Total Years (include High School): 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Income level: 0-5,000   5,000-10,000   10,000-15,000   15,000-20,000   25,000-30,000   30,000+

Current Medical or Psychiatric Condition(s)  (please circle all that apply):

Depression/Major Depression       Asthma
Anxiety Disorder                  Heart Conditions
Impulse Disorder                  High Blood Pressure
Bipolar I                        Stroke
Bipolar II                       TBI
Alcoholism                       Thyroid
Substance Use                    Alzheimer’s
ADHD                             Epilepsy
PTSD                             Paranoia
Schizophrenia                    Dementia
OCD                              Cancer
Autism                           Other_______________      None

Are you currently receiving treatment for the psychiatric condition(s)? YES  NO
If yes, which ones________________________________________________

Are you currently receiving treatment for the medical condition(s)?    YES  NO
If yes, which ones, and what medications?____________________________
APPENDIX K
POST-INTERVIEW FORM

Please circle one of the following:

A. I felt _________ about this drawing activity.

POSITIVE       NEUTRAL       NEGATIVE

B. On a scale of 1-10, 10 being the most serious, I was this serious when conducting the drawing test:

____________

1) Were the instructions clear for this activity?
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________

2) What thoughts (if any) did you have before, during, or after this activity?
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________

3) What do you think about the markers?
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________

4) Did you have trouble with figuring out what some of the images were in the picture? If so, which ones?
________________________________________________________________________
________________________________________________________________________

5) How was the time limit?
________________________________________________________________________
________________________________________________________________________

6) Any other improvements you think this activity could use?
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
REFERENCES


Gravitz, M. (1969). Figure size as an index of depression and MMPI scores in normal adults. *Journal of Clinical Psychology, 7*, 143-144.


117


BIOGRAPHICAL SKETCH

Donald C. Mattson has attended five major universities across three states beginning with Michigan Technological University in 1997 and ending with The Florida State University where he continues to conduct research in art-based assessment. During his extensive school career, he obtained a bachelor’s degree in psychology, a master’s degree in art therapy, one-half of a master’s degree in community counseling, and a doctorate degree in art therapy with a counseling emphasis.

While working in various clinical settings for community counseling, Donald C. Mattson continues to explore the connection between mental illness and art. He does this through an objective methodology, and is most interested in finding ways to quantify art for increasing the credibility in the field of art therapy. His research has gained recognition in Germany where he has been considered for a fellowship.