INFORMATION TO USERS

The most advanced technology has been used to photograph and reproduce this manuscript from the microfilm master. UMI films the original text directly from the copy submitted. Thus, some dissertation copies are in typewriter face, while others may be from a computer printer.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyrighted material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each oversize page is available as one exposure on a standard 35 mm slide or as a 17" × 23" black and white photographic print for an additional charge.

Photographs included in the original manuscript have been reproduced xerographically in this copy. 35 mm slides or 6" × 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.
Meta-analyses of test anxiety among college students

Harris, Mark Manning, Ph.D.

The Ohio State University, 1987
META-ANALYSES OF TEST ANXIETY
AMONG COLLEGE STUDENTS

DISSERTATION

Presented in Partial Fulfillment of the Requirements
for The Degree Doctor of Philosophy in the Graduate
School of the Ohio State University

By

Mark Manning Harris, B.S., M.A.

* * * * *

The Ohio State University

1987

Dissertation Committee:
Harold B. Pepinsky
W. Bruce Walsh
Theodore J. Kaul

Approved by

Advisor

Department of Psychology
To my Family and to my Family of Friends
ACKNOWLEDGMENTS

I would like to thank my family for their support throughout my academic career. I would like especially to thank my advisor, Dr. Harold B. Pepinsky, for his encouragement and unending support for my development as a professional. I would also like to thank Dr. Ted Kaul and Dr. Bruce Walsh for their support at all stages of my graduate career. Additional thanks go to Susan Kashubeck for her advice on this project and her never ending cheerfulness and willingness to help.

Finally, I would like to thank the numerous friends who have helped to make my years in graduate school the best yet. Among the many I owe a particular debt of gratitude to are; Robin August, Audrey Bahrick, Fran Davis, John Fennig, Steve Gravenkemper, Chuck Klink, Dan Morrison, Peg Mortensen, Jim Nolan, Chris Paterson, Karen Peterson, Jim Reusch, Deb Serling, and Todd Sevig.
VITA

November 13, 1959.................. Born - Newton, Iowa

1982............................... B.S., Iowa State
University
Ames, Iowa

1985............................... M.A., The Ohio State
University
Columbus, Ohio

1986-1987.......................... Pre-doctoral Intern
University of
Florida
Gainesville, Florida

1987-............................... Counselor and
Coordinator of
Programming
Miami University
Oxford, Ohio

Field of Study

Major Field: Counseling Psychology
# TABLE OF CONTENTS

DEDICATION ......................................................... ii
ACKNOWLEDGEMENTS ................................................ iii
VITA ................................................................. iv
LIST OF TABLES .................................................... vii

CHAPTER

I. THE PROBLEM .................................................. 1
   - Theories of Test Anxiety ....................... 1
   - The Relationship Between Test Anxiety and Performance  5
   - Interventions to Alleviate Test Anxiety .......... 6
   - Implications of the Published Research .......... 14
   - The Present Investigation ..................... 15

II. THE METHOD ............................................... 17
   - Locating Studies ............................... 17
   - Criteria for Including A Study ............. 19
   - Classifying Studies ......................... 20
   - Coding Characteristics of the Studies ........ 20
   - Calculation of Effect Sizes .................. 26
   - Meta-analysis of the Data ................... 28

III. RESULTS AND DISCUSSION ......................... 31
   - RESULTS ........................................... 31
# LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Identified Categories Employed in Studies</td>
<td>23</td>
</tr>
<tr>
<td>2. Meta-analyses Within Treatments</td>
<td>34</td>
</tr>
<tr>
<td>3. Cohen's Magnitude of Effect Sizes</td>
<td>36</td>
</tr>
</tbody>
</table>
CHAPTER I
THE PROBLEM

Test anxiety has become the most widely studied of the specific anxieties (Sarason, 1980). As a special case of general anxiety, it refers to the phenomenological, physiological, and behavioral responses that accompany concern about possible failure. These responses are to a class of stimuli that have been associated in the individual's experience of evaluation or testing (Sieber, 1980).

The term test anxiety was introduced by Mandler and Sarason (1952) to describe anxiety in evaluation situations. Following its introduction, there has been an increase in the publication of anxiety scales, questionnaires, and measures (Sarason, 1980). The past 25 years have shown considerable growth in the attention paid to test anxiety (Sieber, 1980).

Theories of Test Anxiety

Mandler and Sarason (1952) assumed test situations to be associated with various types of strong learned
drives. A first group, of learned task drives, depends upon the task, the test materials, and instructions for performing the task. These facilitate the completion of the task. A second set, of learned anxiety drives, derives from responses to previously learned anxiety-arousing stimuli, present in the testing situation. The maladaptive aspect of test anxiety has been described by Mitchell and Ingham (1970) as a "debilitating force which disrupts the capacity of the student to concentrate, think, and remember and is generally accompanied by states of extreme tension, restlessness, and in some cases muscular contraction, headache, and nausea." (page 69).

Liebert and Morris (1967) suggested that test anxiety has two major components, worry and emotionality. Worry (W) can be conceptualized as cognitive concern over performance. Emotionality (E) is the autonomic arousal aspect of anxiety. Several studies (Doctor & Altman, 1969; Liebert & Morris, 1967; Morris & Liebert, 1969, 1970; Spiegler, Morris, & Liebert, 1968) suggest that these components yield scores which vary predictably with time relationships to classroom examinations and with performance expectancies.
Wine (1971), in a review of the literature, concluded with an interpretation of the adverse effects which test anxiety has on task performance that was based on attention. The low test-anxious person was found to focus on task-relevant variables in the testing situation. Conversely, the highly test-anxious person was found to focus internally on self-evaluative, self-deprecatory thinking, coupled with negative perception of autonomic responses. Poor results are assumed to occur because test-anxious individuals divide their attention between internal cues and task cues and do not give the full attention to task cues required for a good performance. Support for this conceptualization has come from Sarason, who completed several studies which indicate that the self-deprecatory, self-ruminative thoughts are characteristic of highly test-anxious students (Sarason & Ganzer, 1962, 1963; Sarason & Koenig, 1965).

Further support for this conceptualization came from Holroyd, Westbrook, Wolf, and Badhorn (1978), who monitored autonomic activity during a testing situation and found that virtually identical changes in electrodermal activity and heart rate were obtained for both high-anxious and low-anxious students. The
authors concluded that test anxiety can most accurately be conceptualized as a cognitive and attentional phenomenon rather than as a state of heightened autonomic arousal.

Lazarus and Averhill (1972) proposed an alternate theory of test anxiety. They suggested that the arousal and cognitive components of test anxiety can be identified and defined as individual components, but that they interact as a single process in test anxiety. According to this model, two pathways are possible. Increased levels of arousal can lead to cognitive appraisals and produce subjective test anxiety. Also, cognitive appraisals of threat can result in increased arousal and produce subjective test anxiety.

Culler and Holahan (1980) proposed a broader model than those previously formulated in their belief that the relationship between test anxiety and academic performance is seen as partially dependent upon differences in behaviors relating to study between students who are highly test anxious and students who are not highly test anxious. This conceptualization is consistent with the work of Wittmaier (1972), who found students who were highly test anxious to have significantly lower levels of competence in study
Meichenbaum and Butler (1980) proposed a model of test anxiety that accounts for many features of earlier models. In their formulation, elements of test anxiety are seen as interacting components of behavioral, cognitive, and environmental influences that operate on one another in a self-perpetuating cycle, as in Bandura's (1978) reciprocal determinism. This cycle can operate at an automatic stereotyped level. Specifically, individuals' systems of meaning result in their interpreting physical symptoms of discomfort as anxiety, which leads to self-referent ideation, which in turn influences arousal, leading to avoidant behavior, which increases anxiety, and so forth. Within this framework, test anxiety is a construct that summarizes this complete chain of events (Meichenbaum & Butler, 1980).

The Relationship Between Inferred Test Anxiety and Measured Performance

The relationship between test anxiety and performance has been well documented. For example, Spielberger (1962) reported that students who were identified as highly anxious received lower grades and had a higher academic failure rate than did students of
equal intelligence who were not anxious. Walsh, Engbretson, and O'Brien (1968) showed that measured test anxiety has a significant negative correlation with performance on academic tests. Similar findings have been discussed by other investigators (Allen, Lerner, Wayne, & Hinrichsen, 1972; Carrier & Jewell, 1966; Desiderato & Koskinen, 1969). This relationship is reported to hold for non-academic measures such as the Wonderlic Personnel Test (Wonderlic, 1970) and the Digits Symbols Test of the WAIS-R (Matarazzo, 1972). The seriousness of test anxiety is underscored by the realization that the measurement of academic performance is based substantially on the ability of the student to perform successfully on examinations.

Interventions to Alleviate Test Anxiety

Most of the research on attempts to reduce test anxiety has centered on behavioral interventions, for example, the use of systematic desensitization or training for relaxation. Use of these devices rests on the belief that physiological arousal is the most important indicator of test anxiety. Foremost among these have been interventions utilizing systematic desensitization. Systematic Desensitization has

Training students in techniques to evoke relaxation has also been a method of reducing the physiological arousal believed to be associated with test anxiety. In some studies (Aponte & Aponte, 1971; Finger & Galassi, 1977; Holahan, Richardson, Puckett, & Bell, 1979), relaxation training has been shown to be effective in reducing anxiety while failing to do so in others (Bedell, 1976; Johnson & Sechrest, 1968). Training in relaxation has produced mixed results on
measures of performance. Some studies have demonstrated significant gains on measures of performance (Chang-Liang & Denney, 1976; Denney, 1974; Holahan, Richards, Puckett, and Bell, 1979), while in other studies improvements in performance were not evident (Aponte & Aponte, 1971; Finger & Galassi, 1977; Johnson & Sechrest, 1968; Kirkland & Hollandsworth, 1980; Romano & Cabianca, 1978).

Cue-controlled relaxation (Russell, Miller, & June, 1974) consists of two basic steps: Progressive muscle relaxation and repeated pairing of the relaxed state with the cue word "calm." The use of cue-controlled relaxation has been successful in reducing test anxiety (Russell, Miller, & June, 1975) and in reducing anxiety while at the same time increasing performance on a test of mental abilities (Counts, Hollandsworth, & Alcorn, 1978). In other studies, no significant improvements of experimental over control groups were found on either anxiety measures (Marchetti, McGlynn, & Patterson, 1977; McGlynn, Kinjo, & Doherty, 1978; Russell & Lent, 1982) or on performance measures (Russell et al., 1975).
Another technique used to reduce negative physiological arousal associated with test anxiety is implosion (Stampfl & Levis, 1967). Implosive therapy is a behavioral technique which reduces maladaptive anxiety by emphasizing the presentation of the stimulus which has high anxiety-evoking characteristics until the stimulus is no longer able to evoke anxiety. This type of intervention has met with mixed or partial success in the reduction of anxiety (Cornish & Dilley, 1973; Dawley & Wenrich, 1973; Graff, MacLean, & Loving, 1971; Prochaska, 1971; Smith & Nye, 1973). Similarly, it has also produced mixed results on measured performance, with some studies showing improvement (O'Brien, 1976; Prochaska, 1971) and other studies failing to do so (Cornish & Dilley, 1973; Levine & O'Brien, 1980; Smith & Nye, 1973).

A widely used method of reducing students' test anxiety by teaching them study skills (e.g., as used by Culler & Holahan, 1980) rests on the belief that such anxiety is related to a skill-deficit. Research indicates that highly test-anxious students have poorer study skills than do their less anxious counterparts (Desiderato & Koskinen, 1969). Therefore, teaching study skills to these students might also result in
lowering their anxiety. Studies which have used training in study skills to alleviate test anxiety include those of Cornish and Dilley (1973), Horne and Matson (1977), and Osterhouse (1972). Increases in performance were found by Allen (1973), Horne & Matson (1977), and Lurie and Steffen (1980). Several other investigators found no increases in performance for subjects given training in study skills (Altmaier & Woodward, 1981; Cornish & Dilley, 1973, Dendato & Diener, 1986; Osterhouse, 1972; Ricketts & Galloway, 1984).

Covert positive reinforcement (Cautela, 1970) is predicated on the assumption that any behavior that can be influenced overtly by operant conditioning may also be influenced covertly. In this procedure, the student imagines making non-anxious responses in a test situation. These responses are followed immediately by the imagining of reinforcing stimuli. This procedure has been shown to be effective in the reduction of test anxiety in several studies (Guidry & Randolf, 1974; Kostka & Galassi, 1974; Wisocki, 1973) but not in another (Bajtelsmit & Gershman, 1976). The effects on measures of performance have also been mixed, with investigators recording both positive findings (Kostka
& Galassi, 1974) and indeterminate results (Bajtelsmit & Gershman, 1976).

Not all researchers believe that the emphasis on behavioral interventions has been an appropriate one. Wine (1971) was one of the first reviewers to note the role played by intrusive, task-irrelevant thoughts on arousal and performance. Within the cognitive interventionist's school of thought, it is believed that concentrating on emotional responses, defined as those of physiological and affective arousal, focuses on the transient situational cues present only in the stressful teaching situation itself. Cognitive responses indicative of worry, such as preoccupation with outcomes and fear of failure, are seen as more stable characteristics of an individual than are the transient emotional responses. Therefore it is assumed that treatment of such cognitive manifestations of test anxiety are more likely to effect performance changes than is treatment of the emotionality component (Finger & Galassi, 1977).

Several researchers have successfully reduced anxiety by teaching individuals to modify unrealistic beliefs (Goldfried, Linehan, & Smith, 1978; Holroyd, 1976; Meichenbaum, 1972). There have also been studies
in which gains in performance were demonstrated (Crowley, Crowley, & Clodfelter, 1986; Holahan et al., 1979; Holroyd, 1976; Horne & Matson, 1977; Meichenbaum, 1972). In other studies there were no significant improvements (D'Alelio & Murray, 1981; Dendato & Diener, 1986; Finger & Galassi, 1977; Lurie & Steffen, 1980; Ricketts & Galloway, 1984).

Counseling has not been shown to be effective in reducing test anxiety. Investigators report no significant changes between subjects in counseled and non-counseled groups (Anton, 1976; Lomont & Sherman, 1971). Group counseling has been shown to produce gains in performance in some instances (Doctor, Aponte, Burry, & Welch, 1970) but not in others (Anton, 1976; Lomont & Sherman, 1971; Mitchell & Ng, 1972).

There has been a trend toward the increasing use of interventions that combine several approaches (Dendato & Diener, 1986). Katahn, Strenger, and Cherry (1966) used a combination of counseling and desensitization and found a reduction in test anxiety. Other investigators have also employed some form of counseling in addition to, or in combination with systematic desensitization (Allen, 1971; McManus, 1971, Mitchell & Ng, 1972). When Decker and Russell (1981)
combined study skills counseling with a decision-making model, they found significant improvements in grade point averages. Similarly, Dendato & Diener (1986) and Kirkland and Hollandsworth (1980) used a combination of cognitive therapy and study-skills training to improve students' academic performance as measured by grade point averages. A combination of study-skills training and systematic desensitization has been reported to be effective and superior to either component alone, both in reducing measured anxiety and in improving academic performance (Allen, 1971; Lent & Russell, 1978; Mitchell, Hall, & Piatkowska, 1975).

Other studies have not demonstrated gains in performance from using combined approaches. Altmaier and Woodward (1981), using a combination of desensitization and study-skills training, found no significant increases on either course grades or on grade point averages. Finger and Galassi (1977), who employed a combination of relaxation and cognitive approaches, also failed to find significant improvements in performance as measured by the Wonderlic Personnel Test (Wonderlic, 1970) and the Digit Symbols Test (Matarazzo, 1972).
Implications of the Published Research

A review of the literature on test anxiety raises questions. It is unclear which interventions are comparatively more successful than are other interventions in reducing subjective anxiety among students. It is also unclear which interventions result in the largest increases on performance measures. With performance measures, there have been conflicting results. Overall, previous reviews have indicated that performance gains occurred in less than half the studies (Allen, Elias, & Zlotlow, 1980; Wine, 1971). How much faith can we have in our ability to reduce test anxiety and to improve performance?

Perhaps the greatest need in research on test anxiety is not for more studies, but rather for an analysis of what we know now and what we ought to know and how this is to be determined (Allen et al., 1980). In keeping with the recommendation of the research group at the Conference for the Future of Counseling Psychology (held April 3 through April 6, 1987, in Atlanta, Georgia), it would seem that there should be more research synthesis. An integrative review would serve the purpose of clarifying and reconciling discrepancies in and drawing overall conclusions from
The many separate studies that have addressed the need for alleviating test anxiety.

The recently developed statistical procedures of meta-analysis (Glass, 1976) claims to provide the tools necessary for just such a systematic review of the literature. Meta-analysis is heralded as a quantitative way for organizing and describing a body of research literature. The major advantage of meta-analysis is said to be that it provides an objective and quantitative means for evaluating the findings (Kazdin, 1985). Another purported advantage is that meta-analysis is more likely to be able to detect reliable effects of the variables which moderate findings than are narrative reviews (Kazdin, 1985). For a more complete discussion of meta-analysis, see Appendix A.

The Present Investigation

There are several goals for the present investigation.

First, to determine the overall effectiveness of interventions in reducing test anxiety as measured by self-report. Second, to determine whether there are differences
in the degree of effectiveness among interventions, and, if so, to account for differences based on the use of moderator variables.

Third, to examine the concomitant effect of these interventions on measures of performance, again so as to arrive at an overall level of effectiveness.

Fourth, if there are differences among these interventions on measures of performance, to account for differences based on the use of moderator variables.
Locating Studies

Relevant studies were located by means of procedures outlined by Cooper (1984). As a guideline, Cooper advises employing multiple channels of locating studies so the chances of strong unidentified bias distinguishing included from unincluded studies is small. Bias might occur if the review were limited in the methods used for obtaining relevant studies, such as in searching only for studies that appeared in prestigious journals. Examples of the channels that Cooper recommends include on-line computer searches, the tracking of references cited in other publications, and informal sources such as those studies identified by colleagues.

Studies for this meta-analysis were originally located through the indexing services of The Educational Resources Information Center (ERIC) and
Psychological Abstracts. Searches run on four separate occasions yielded studies considered for inclusion.

On November 8, 1985 an ERIC search was performed using the descriptors of "study habits," "study skills," and "self-control" crossed with "college students." This search produced 463 references. On April 8, 1986 an ERIC search was again run, this time using the descriptors "academic" crossed with "effectiveness," "academic" crossed with "interventions," and "college students" crossed with "study skills" and with "interventions." This search produced 37 references.

On June 20, 1987 a Psychological Abstracts search was run covering the literature from 1967 until June 1987. The descriptors used for this search were "test anxiety" crossed with "college students." This search yielded 320 references. On June 28, 1987 an ERIC search was run covering the time period from January 1981 through March 1987. This search also used the descriptors "test anxiety" crossed with "college students." This search yielded 52 references.

From these articles, appropriate studies were identified by reference to criteria discussed below. Next, the ancestry approach (Cooper, 1984) was used to
locate additional studies. The ancestry approach involves "tracking down" the research cited in already obtained relevant research. The reference list of each article obtained was checked for relevant citations and these were then checked against the master list of studies for inclusion. This was done until all relevant studies were identified.

Criteria for Including a Study

The population of studies for this meta-analysis was limited to empirical investigations centered on the decrease of test anxiety among college students. Several criteria had to be satisfied before a study could be included for analysis. The first criterion was that the study be published in a scientific journal. Next, studies were limited to those using college students as the population from which samples were drawn. Studies were further limited to those which had used a design containing an adequate control group. There were 90 studies that met the above conditions. The number of studies eliminated because they did not contain enough data to meta-analyze was 20. This left a total of 70 studies for inclusion in the final analysis.
Classifying Studies

The first step in coding was to determine to which category of interventions a specific study belonged. Here there were ten possible categories to which a study could belong. These categories were systematic desensitization, relaxation, cue-controlled relaxation, implosion, study skills, reinforcement, cognitive, counseling, "other," and combined treatments. The "other" category was added for treatments that were not appropriate for inclusion in any of the initial categories. Decisions regarding which category was appropriate were made based on a combination of what the intervention was called by the investigator(s), and on my subjective judgments and those of another investigator familiar with the literature. These judgments were based on a careful reading of the "Methods" section in each study. Any disagreements in classification were resolved by discussions between the judges, to arrive at consensus.

Coding Characteristics of the Studies

After the research studies were identified and selected, various characteristics of the studies were
coded. This coding process was undertaken to aid in addressing the questions underlying the meta-analysis: To determine the specific effectiveness of these interventions, and to account for how the findings vary from one type of study to another. "Moderator variables" are characteristics of the studies which might influence outcome. Some of the moderator variables were selected because of their apparent impact in previous studies. Others were selected intuitively, or included because they were readily available. According to Cooper (1984), the first rule for constructing a coding sheet is that any information that might have the slightest chance of being relevant to the analysis should be retrieved from each study. It is impossible to know beforehand which moderator variables will have the strongest influence on the results. It is also much more difficult to return to previously coded studies to add new moderator variables than it is to discard mistakenly coded information.

The variables coded for analysis as potential moderator variables included type of control group, duration of the intervention, method of treatment administration, age of subjects, method of recruitment, test anxious status of the subjects, random assignment
of subjects, subjects' gender, level of therapists' training, and setting of the research. All of these classifications decisions were made independently by two separate investigators. In the event the investigators arrived at different conclusions, an attempt was again made to reach consensus. This occurred in all cases. In some cases not enough information was available to reach a conclusion. This resulted in some missing data in the matrix of categories. Table 1 contains a summary of the categories that were coded.

The first category of classification to be coded was the type of control group. The sub-categories were either "no significant contact" or "placebo." In the first category subjects were merely tested and no other contact occurred. In the "placebo" category some attempt was made to control for the effects of attention, usually by providing some form of inert treatment.

The duration of the intervention was coded according to the total number of minutes that comprised the intervention. This was thought to be a more accurate estimate than to use number of sessions, because the actual amount of time involved per session
## Table 1

Identified Categories Employed in Studies

<table>
<thead>
<tr>
<th>Category</th>
<th>Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment Type</td>
<td>Relaxation, Systematic Desensitization, Study Skills, Cognitive, Counseling, Implosion, Covert Positive Reinforcement, Cue-Controlled Relaxation, Combination, Other</td>
</tr>
<tr>
<td>Control Type</td>
<td>Placebo, No Significant Contact</td>
</tr>
<tr>
<td>Treatment Administration</td>
<td>Individual, Group, Live, Automated</td>
</tr>
<tr>
<td>Recruitment</td>
<td>Invitation, Money or Point Reward, Required, Unknown</td>
</tr>
<tr>
<td>Previous Test-Anxious Status</td>
<td>Not Measured, Chosen Anxious</td>
</tr>
<tr>
<td>Random Assignment</td>
<td>Yes, No</td>
</tr>
<tr>
<td>Gender</td>
<td>Percentage Male</td>
</tr>
<tr>
<td>Therapist Experience</td>
<td>Undergraduate or Graduate Professional, Unknown</td>
</tr>
<tr>
<td>Category</td>
<td>Levels</td>
</tr>
<tr>
<td>------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>Setting</td>
<td>Treatment Agency</td>
</tr>
<tr>
<td></td>
<td>All Other</td>
</tr>
</tbody>
</table>
varied from as little as eight minutes (Bruch, 1978) to as much as 90 minutes (Allen, 1971).

Administration of treatment covered two basic considerations. First, interventions were classified in terms of whether they were administered individually or in groups. Second, interventions were classified as either being live or automated, the latter utilizing audiotape or videotape.

The age of subjects was coded using the following scheme: Subjects identified as "freshmen," "first year," or aged 18 to 18.99 were called "freshmen," those described as "sophomores," "second year," or aged 19 to 19.99 were called "sophomores." A similar system was used for Juniors and for Seniors. If not enough information was provided, the age of the subjects was coded as "unknown."

There were three possible classifications for method of recruitment. The first included subjects who responded to an invitation for participating in the intervention. The second class was for those subjects who received either a monetary reward or earned points toward a course grade for participation. There was a third category for not enough information to classify.
Also classified was whether there was a previously measured high level of test-anxiety among subjects in the study. In order to meet this condition, subjects had to have been given some type of measurement of test-anxiety and to have met a criterion score of anxiety for inclusion. On this basis, studies were divided according to whether or not subjects were required to fulfill such a criterion.

Random assignment was classified by whether this was part of the design used. Studies which utilized a stratified random assignment design were also considered to have met this criterion. The gender category was constructed using the percentage of subjects that were male. Either gender could have been selected for this purpose.

Finally, the level of the therapist's experience was classified as that of either graduate student, professional, or unknown. The setting category was based on whether the intervention took place in a special treatment agency setting or in a classroom.

**Calculation of Effect Sizes**

Effect sizes for the studies were coded according to Hedges' $g$ as the indicator of effect size (Hedges,
An effect size is an attempt to quantify the post-treatment advantage of having been in a group which received the treatment. Hedges' $g$ is comparable to Cohen's $d$ and Glass's $g$ (Rosenthal, 1984). Hedges' $g$ is calculated by subtracting the mean of the control group from the mean of the experimental group and dividing this quantity by their pooled standard deviation. When the results were in the expected direction (i.e., where the intervention led to a grade increase) Hedges' $g$ was a positive value. When the results were not in an expected direction, Hedges' $g$ acquired a negative value.

Several times, researchers reported only that an effect size was "not significant." When this occurred, these effects were assigned the value of zero and included in the analysis. This method allowed for the inclusion of valuable data by recourse to a relatively conservative estimate of effect size (Cooper, 1984).

On several occasions a study reported results for more than one dependent variable related to anxiety, to performance, or to both. If each effect size for each dependent variable were entered into the analysis, rules of statistical independence would be violated, because each result would be counted as though it were
a separate study (Wilson, 1985). To avoid this violation of statistical independence, Rosenthal (1984) recommended that each sample of subjects contribute only a single effect size to the total in any overall analysis. One commonly used method of obtaining a single estimate of effect size is to calculate the average of the effect sizes for all the dependent variables of each type included in the study. This procedure, called the mean effect size procedure (Rosenthal, 1984), was used in this study.

For those studies with incomplete reporting of means and standard deviations, the methods described by Hedges and Olkin (1983) were used to derive \( \eta \) from \( t \) and \( F \) scores. All such calculations were cross-checked by another investigator familiar with the procedure.

Meta-analysis of the Data

Two separate meta-analyses were conducted. The first meta-analysis dealt with measures of anxiety used as dependent variables. Examples of these include the Test Anxiety Questionnaire (TAQ; Mandler & Sarason, 1952), the Test Anxiety Scale (TAS; Sarason 1958), the Achievement Anxiety Test (AAT; Alpert & Haber, 1960), the Otis-Lennon Mental Abilities Test (Otis & Lennon,
1967) and the Trait Anxiety Inventory (TAI, Spielberger, Gorsuch, & Lushene, 1970).

The second meta-analysis dealt with performance measures. These included both academic measures, such as grades, grade point averages, or test scores, and other performance measures such as the Wonderlic Personnel Test (Wonderlic, 1970) and the Digit Symbols Test (Matarazzo, 1972).

A test of heterogeneity of effect sizes (diffuse comparison chi-squares; Rosenthal, 1984, p. 78) was performed for each meta-analysis to test the hypothesis that the effect sizes were homogeneous. Variations in results across studies too great to be explained by sampling error were isolated in terms of the coded moderator variables. Where moderator variables were identified as having a significant influence on the results, each level of that moderator variable was meta-analyzed separately. The same procedure was followed as for the initial meta-analysis. Where additional moderator variables were found, separate meta-analyses were performed on each level. This process is referred to as "disaggregation." Patterns in the variation of population effect sizes were explored until an adequate model of the effect sizes was
obtained, such that studies could be classified into relatively homogeneous categories (Hunter, Schmidt, & Jackson, 1982) or until the pool of studies could no longer be divided into smaller groups. This was accomplished with an SAS computer programming package.
CHAPTER III
RESULTS AND DISCUSSION

RESULTS

This study contained two meta-analyses. The first was concerned with anxiety measures, and the second was concerned with performance measures. This provided information separately for the four questions this study was designed to address: Did these interventions reduce subjective anxiety, and what accounted for variations, if any, across studies? Also, did these interventions increase performance and, again, what accounted for any variation across studies?

The first analyses performed were tests to determine whether there was heterogeneity among effect sizes. The test statistic Q indicated that significant heterogeneity existed for both anxiety measures [$Q = 250.02, \chi^2(111) = 144.02, p < .005$] and for performance measures [$Q = 182.93, \chi^2(86) = 113.45, p < .005$].
For anxiety measures, there was insignificant variance among seven of the 10 treatment types. The three heterogenous categories were systematic desensitization, combined treatments, and cue-controlled relaxation. For performance measures, there were also seven categories that were homogeneous. Here, the three categories that were not heterogeneous were systematic desensitization, study skills, and combined treatments.

The overall $g$ scores for both the anxiety and the performance variables indicated that these interventions were effective in reducing anxiety and, to a lesser extent, they were also effective in improving performance. The overall $g$ for the anxiety measures was $0.7517 (Z = 20.26, p < 0.001)$. The overall $g$ for performance measures was $0.3511 (Z = 8.165, p < 0.001)$.

Cohen (1977) commented on levels of effect sizes that can be thought of as large, medium, and small. According to Cohen, an effect size of 0.7 or greater can be thought of as large. An effect size of 0.5 can be considered medium. Finally, an effect size of 0.3 or less can be considered small. Using these definitions, the overall effect size for the reduction of anxiety
was large, whereas the overall effect size for increasing performance was small.

In addition to finding the overall $g$ scores for both anxiety and performance measures, this study also examined $g$ scores for both anxiety and performance for each of the ten categories of interventions in order to compare the differential effectiveness of treatments. This information is summarized in Table 2. The ranking of these categories of interventions according to Cohen's ranges (1977) is summarized in Table 3.

**Effectiveness of Various Treatments**

The first category of treatment was systematic desensitization. The overall $g$ for anxiety was .873, which could be considered large by Cohen's standards. This figure was significantly greater than zero ($Z = 13.88, p < .001$). Initially there was heterogeneity in this score. This heterogeneity was eliminated by grouping studies first according to the type of control group, then the level of counselor experience, then the mode of treatment administration, then the gender of the subjects. A $g$ score of .2 was obtained for performance measures. Although this score was small, it was still significantly greater than zero ($Z = 2.61,$
<table>
<thead>
<tr>
<th>Treatment</th>
<th>Anxiety</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Systematic</td>
<td>0.873</td>
<td>0.20</td>
</tr>
<tr>
<td>Desensitization</td>
<td>z = 13.88</td>
<td>2.61</td>
</tr>
<tr>
<td></td>
<td>p = &lt;.001</td>
<td>&lt;.005</td>
</tr>
<tr>
<td></td>
<td>Q = 121.50</td>
<td>47.99</td>
</tr>
<tr>
<td></td>
<td>x = 55.76</td>
<td>38.89</td>
</tr>
<tr>
<td></td>
<td>df = 37</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>p = &lt;.001</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>Relaxation</td>
<td>g = 0.667</td>
<td>0.286</td>
</tr>
<tr>
<td></td>
<td>z = 5.15</td>
<td>1.15</td>
</tr>
<tr>
<td></td>
<td>p = &lt;.001</td>
<td>&lt;.02</td>
</tr>
<tr>
<td></td>
<td>Q = 16.908</td>
<td>21.42</td>
</tr>
<tr>
<td></td>
<td>x = 16.92</td>
<td>15.51</td>
</tr>
<tr>
<td></td>
<td>df = 9</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>p = &gt;.05</td>
<td>&gt;.05</td>
</tr>
<tr>
<td>Cue-Controlled</td>
<td>g = 0.423</td>
<td>0.244</td>
</tr>
<tr>
<td>Relaxation</td>
<td>z = 2.43</td>
<td>1.13</td>
</tr>
<tr>
<td></td>
<td>p = &lt;.007</td>
<td>&gt;.10</td>
</tr>
<tr>
<td></td>
<td>Q = 16.74</td>
<td>.28</td>
</tr>
<tr>
<td></td>
<td>x = 12.59</td>
<td>3.62</td>
</tr>
<tr>
<td></td>
<td>df = 6</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>p = &lt;.025</td>
<td>&gt;.50</td>
</tr>
<tr>
<td>Implosion</td>
<td>g = 0.644</td>
<td>0.196</td>
</tr>
<tr>
<td></td>
<td>z = 5.05</td>
<td>1.20</td>
</tr>
<tr>
<td></td>
<td>p = &lt;.001</td>
<td>&gt;.10</td>
</tr>
<tr>
<td></td>
<td>Q = 6.60</td>
<td>3.99</td>
</tr>
<tr>
<td></td>
<td>x = 15.51</td>
<td>9.49</td>
</tr>
<tr>
<td></td>
<td>df = 8</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>p = &gt;.50</td>
<td>&gt;.25</td>
</tr>
<tr>
<td>Study Skills</td>
<td>g = 0.42</td>
<td>0.232</td>
</tr>
<tr>
<td></td>
<td>z = 3.93</td>
<td>1.89</td>
</tr>
<tr>
<td></td>
<td>p = &lt;.001</td>
<td>&lt;.03</td>
</tr>
<tr>
<td></td>
<td>Q = 20.04</td>
<td>23.35</td>
</tr>
<tr>
<td></td>
<td>x = 21.03</td>
<td>18.31</td>
</tr>
<tr>
<td></td>
<td>df = 12</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>p = &gt;.05</td>
<td>&lt;.01</td>
</tr>
</tbody>
</table>
### Table 2 Continued

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Anxiety</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Covert Positive</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reinforcement</td>
<td></td>
<td></td>
</tr>
<tr>
<td>g = 0.758</td>
<td>0.121</td>
<td>0.405</td>
</tr>
<tr>
<td>z = 3.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p = &lt;.002</td>
<td>&gt;.30</td>
<td></td>
</tr>
<tr>
<td>Q = 0.514</td>
<td></td>
<td></td>
</tr>
<tr>
<td>x = 5.99</td>
<td>3.84</td>
<td></td>
</tr>
<tr>
<td>df = 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p = &gt;.750</td>
<td>&gt;.50</td>
<td></td>
</tr>
<tr>
<td>Cognitive</td>
<td></td>
<td></td>
</tr>
<tr>
<td>g = 0.787</td>
<td>0.61</td>
<td></td>
</tr>
<tr>
<td>z = 8.51</td>
<td>5.36</td>
<td></td>
</tr>
<tr>
<td>p = &lt;.001</td>
<td>&lt;.001</td>
<td></td>
</tr>
<tr>
<td>Q = 18.19</td>
<td></td>
<td>16.41</td>
</tr>
<tr>
<td>x = 22.36</td>
<td>16.919</td>
<td></td>
</tr>
<tr>
<td>df = 13</td>
<td></td>
<td>9</td>
</tr>
<tr>
<td>p = &gt;.10</td>
<td></td>
<td>&gt;.05</td>
</tr>
<tr>
<td>Counseling</td>
<td></td>
<td></td>
</tr>
<tr>
<td>g = 0.472</td>
<td>0.17</td>
<td></td>
</tr>
<tr>
<td>z = 1.42</td>
<td>0.587</td>
<td></td>
</tr>
<tr>
<td>p = &gt;.07</td>
<td>&gt;.27</td>
<td></td>
</tr>
<tr>
<td>Q = 0.039</td>
<td></td>
<td>2.72</td>
</tr>
<tr>
<td>x = 3.84</td>
<td>5.99</td>
<td></td>
</tr>
<tr>
<td>df = 1</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>p = &gt;.75</td>
<td>&gt;.250</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td></td>
</tr>
<tr>
<td>g = 0.792</td>
<td>0.294</td>
<td></td>
</tr>
<tr>
<td>z = 2.48</td>
<td>0.62</td>
<td></td>
</tr>
<tr>
<td>p = &lt;.007</td>
<td>&gt;.25</td>
<td></td>
</tr>
<tr>
<td>Q = 0.19</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>x = 3.84</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>df = 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p = &gt;.50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Combination</td>
<td></td>
<td></td>
</tr>
<tr>
<td>g = 0.986</td>
<td>0.708</td>
<td></td>
</tr>
<tr>
<td>z = 8.88</td>
<td>6.55</td>
<td></td>
</tr>
<tr>
<td>p = &lt;.001</td>
<td>&lt;.001</td>
<td></td>
</tr>
<tr>
<td>Q = 25.91</td>
<td></td>
<td>51.27</td>
</tr>
<tr>
<td>x = 22.36</td>
<td>23.69</td>
<td></td>
</tr>
<tr>
<td>df = 13</td>
<td></td>
<td>14</td>
</tr>
<tr>
<td>p = &lt;.025</td>
<td>&lt;.001</td>
<td></td>
</tr>
</tbody>
</table>
### Table 3
Cohen's Magnitude of Effect Sizes

<table>
<thead>
<tr>
<th>Magnitude</th>
<th>Anxiety Treatment</th>
<th>G</th>
<th>Performance Treatment</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large (.7 +)</td>
<td>Combined</td>
<td>.986</td>
<td>Comb.</td>
<td>.708</td>
</tr>
<tr>
<td></td>
<td>Sys. Des.</td>
<td>.873</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>.792</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cognitive</td>
<td>.787</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Reinforc.</td>
<td>.758</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medium (.5 - .7)</td>
<td>Relaxation</td>
<td>.667</td>
<td>Cognitive</td>
<td>.610</td>
</tr>
<tr>
<td></td>
<td>Implosion</td>
<td>.644</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Small (&lt;.5)</td>
<td>Counseling</td>
<td>.472</td>
<td>Other</td>
<td>.294</td>
</tr>
<tr>
<td></td>
<td>C.C. Relax.</td>
<td>.423</td>
<td>Relaxation</td>
<td>.286</td>
</tr>
<tr>
<td></td>
<td>St. Skills</td>
<td>.420</td>
<td>C.C. Relax.</td>
<td>.244</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>St. Skills</td>
<td>.232</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Sys. Des.</td>
<td>.200</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Implosion</td>
<td>.196</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Counseling</td>
<td>.170</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Reinforc.</td>
<td>.121</td>
</tr>
</tbody>
</table>
This score also contained heterogeneity \( Q = 47.99, \chi^2 = 38.89, p < .01 \).

The heterogeneity of this score was not able to be eliminated through grouping studies by any of the coded moderator variables. The studies did not systematically vary with respect to effect size. There are several reasons why this could have happened. The first possible explanation was that there was information contained within the reports of the studies that was not coded for this meta-analysis. The second possible explanation is that factors affecting outcome were not included in the research reports from which data for this study were obtained. Either of these explanations could be responsible for the failure to reduce this heterogeneity to non-significance.

A relatively large \( g \) score for anxiety measures of .667 was obtained for the next intervention category, relaxation. This score was significantly greater than zero \( (Z = 5.15, p < .001) \). There was no significant heterogeneity in this score \( [Q = 16.908, \chi^2 = 16.92, p > .05] \). The \( g \) score for performance was .286, a relatively small score but still significantly greater than zero \( (Z = 2.15, p < .02) \). There was no heterogeneity in this score, either \( [Q = 13.42, \chi^2 = \ldots] \).
Relaxation techniques both reduced anxiety and increased performance, although the latter to a lesser extent.

In the cue-controlled relaxation category, an overall $g$ score of $0.423$ was obtained for anxiety measures. This score was significantly greater than zero ($Z = 2.43$, $p < 0.007$). Heterogeneity was eliminated by grouping studies according to the type of control group. The performance measures for the cue-controlled relaxation group produced an overall $g$ score of $0.244$, which was not significantly greater than zero ($Z = 1.13$, $p > 0.10$). There was no heterogeneity in this score [$Q = 0.28$, $X^2_j = 3.62$, $p > 0.50$].

In the implosion category, there was also a significant reduction in anxiety. The overall $g$ score here was $0.644$, a figure significantly greater than zero ($Z = 5.05$, $p < 0.001$). There was, however, no significant improvement on performance measures. Here the $g$ score of $0.196$ was not significantly greater than zero ($Z = 1.20$, $p > 0.10$). There was no significant heterogeneity for either the anxiety score [$Q = 6.60$, $X^2_g = 15.51$, $p > .50$] or the performance score [$Q = 3.99$, $X^2_u = 9.49$, $p > .25$]. Implosion is designed to reduce anxiety and appears to do so with relative economy. It
does not appear viable, however, as an agent for increasing performance.

In the study skills category, anxiety measures yielded an overall $g$ score of .42. This number was significantly greater than zero ($Z = 3.93, p < .001$). No significant heterogeneity was found in this score [$Q = 20.04, X_{c}^{2} = 21.03, p > .05$]. The $g$ score for performance (.232) was also significantly greater than zero ($Z = 1.89, p < .03$). Significant heterogeneity in this score was eliminated by grouping studies based on both the type of control group and the percentage of male subjects.

For the covert positive reinforcement category, the $g$ score of .758 for anxiety measures was large. This score was significantly greater than zero ($Z = 3.04, p < .002$). There was no significant heterogeneity in this score [$Q = .514, X_{c}^{2} = 5.99, p > .75$]. Once again, the performance score for this category was not large. The overall $g$ score of .121 was not significantly greater than zero ($Z = .405, p > .30$). There was no significant heterogeneity in this score, either [$Q = .299, X_{c}^{2} = 3.84, p > .50$]. Like implosion, reinforcement appears to be successful in reducing anxiety while it has not been shown to increase
performance.

For cognitive interventions, the overall $g$ score for anxiety was .787, a relatively large figure that was significantly greater than zero ($Z = 8.51$, $p < .001$). There was no significant heterogeneity in this score [$Q = 18.19$, $X^2_{13} = 22.36$, $p > .10$]. The overall $g$ score of .61 was obtained for performance. This score was also significantly greater than zero ($Z = 5.36$, $p < .001$). There was no significant heterogeneity in this score [$Q = 16.41$, $X^2 = 16.919$, $p > .05$]. Interventions in the cognitive category also proved successful in reducing anxiety. They were also more successful than many other techniques in increasing performance.

For the counseling category, neither the anxiety $g$ of .472 ($Z = 1.42$, $p > .07$) nor the performance $g$ of .17 ($Z = .587$, $p > .27$) were significantly greater than zero. Again, there was no significant heterogeneity for either the anxiety [$Q = .039$, $X^2 = 3.84$, $p > .75$] or the performance [$Q = 2.72$, $X^2 = 5.99$, $p > .25$] scores. This suggests that the value of traditional counseling approaches for reducing the detrimental effects of test anxiety has not been substantiated for either anxiety or performance measures.
In the "other" category, the overall $g$ score for anxiety measures was 0.792, a figure significantly greater than zero ($Z = 2.48, p < .007$). There was no significant heterogeneity in this score [$Q = 19$, $X^2_1 = 3.84$, $p > .50$]. Performance measures, however, were less impressive ($g = 0.294$). The performance score was not significantly greater than zero ($Z = 0.62, p > .25$). There was no significant heterogeneity in this score [$Q = 0$, $X^2_0 = 0$]. There was only one study in this category.

In the combined category, the overall $g$ for anxiety was a very high 0.986. This number was significantly greater than zero ($Z = 8.88, p < .001$). Significant heterogeneity within this score was eliminated by grouping studies according to the type of control group and the number of sessions. The performance score $g$ for the combined treatment group was 0.708. This score was also significantly greater than zero ($Z = 6.55, p < .001$). The heterogeneity in this score was accounted for by grouping studies according to the type of control group and by the total number of minutes used in the treatment.

One-way ANOVA's were run on the $g$ scores for the different treatments. For the anxiety measures, there
were no significant differences among the treatments in terms of effect size, $F(9, 102) = 1.63$, $p = .1154$ (n.s.). There were also no significant differences between the treatments for performance measures, $F(9, 77) = 1.36$, $p = .2209$ (n.s.).

**Planned Comparisons**

Planned comparisons were made to examine the moderator variables by their different levels. Those examined were the level of counselor experience, whether the assignment of subjects was random, type of recruitment, previous test anxious status, type of control group, and two types of variables related to administration: individual versus group, and live versus automated.

The first planned comparison dealt with the level of counselor experience. Here, those at the graduate or undergraduate level were compared to those at the Doctoral level. It was hypothesized that the main effects size for Doctoral level counselors would be significantly larger than the main effects size for those without Doctorates because of the difference in training level. This hypothesis was not supported.

For both anxiety measures and for performance measures
those without Ph.D.s were more successful ($Z = 2.431, p < .001$ for anxiety and $Z = 1.96, p < .025$ for performance).

The next planned comparison concerned random assignment. It was hypothesized that the g for non-random assignment studies would be higher than the g for studies that utilized random assignment. This hypothesis was chosen because the studies employing randomization utilized a more rigorous methodology. Here, there were no significant differences between studies that used random assignment and studies that did not for either anxiety ($Z = .877, p > .19$) or for performance ($Z = 1.393, p > .08$) measures.

Method of recruitment was also subjected to a planned comparison. Here it was hypothesized that the g for subjects who volunteered for treatment without further inducements would be higher than the g for subjects that were given class credit or other rewards, because of the presumed greater level of motivation for those who volunteered for intrinsic reasons. There were no significant differences between subjects who had volunteered and subjects who received course credit or renumeration on the anxiety measures ($Z = .449, p > .32$). There were significant differences, however,
between these groups on the performance measures \( (Z = 2.62, p < .005) \) with the subjects who volunteered having a higher \( g \) scores than those receiving some type of reward. This finding is in support of the hypothesis.

The next planned comparison dealt with whether there was a criterial cut-off level of test anxiety as measured by an anxiety instrument to gain inclusion in the study. It was hypothesized that the \( g \) score for subjects in studies that had such a measurable cut-off would be higher than the \( g \) score for those in studies that did not have such a criterion. This hypothesis was made because subjects with a previously measured high level of anxiety are more likely to demonstrate improvements, whether legitimately or through regression toward the mean. Here, the hypothesis was supported for anxiety measures, with tested subjects having a significantly higher \( g \) than for subjects that were not tested \( (Z = 1.78, p < .038) \). For performance measures there were no significant differences between the two groups \( (Z = .038, p > .48) \).

The type of control group used in a study was also the subject of a planned comparison. Here, it was hypothesized that designs using a no contact control would have significantly higher \( g \) scores than those for
designs using a placebo control due to the former studies containing a less stringent methodology. This hypothesis was supported for the anxiety measures, where placebo designs had significantly lower g scores than did no contact designs ($Z = 3.96, p < .001$). Somewhat surprisingly, for performance measures, placebo groups had significantly higher g scores than did their no contact counterparts ($z = 1.795, p < .037$). It is not clear how this result occurred.

The next planned comparison concerned whether the treatment was administered individually or in groups. Here, it was hypothesized that the g scores for individually administered treatments would be higher than the g scores for group administered treatments. This hypothesis was chosen because of several assumed advantages of individual treatment. There is both a greater likelihood of individually tailored treatment plans being formulated and of more time being spent working directly with a given subject in the context of individual treatment. For both anxiety measures ($Z = 3.33, p < .001$) and for performance measures ($Z = 1.99, p < .024$) this hypothesis was supported. Individuals appeared to show significantly greater levels of improvement when treated individually as opposed to
being treated in groups.

The final planned comparison dealt with whether the intervention was administered live or in an automated fashion. Here, it was hypothesized that the immediacy of live administration would produce significantly greater improvement than would automated presentation. This hypothesis was supported for anxiety measures ($Z = 2.7, p < .004$) but not supported for performance measures ($Z = .45, p > .32$).

DISCUSSION

Anxiety Measures

Several conclusions can be put forth based on the findings of the present study. Overall, it appears that interventions designed to reduce subjective anxiety have been successful. This is consistent with previous reviews (Allen et al., 1980; Brown & Nelson, 1983). Of the ten categories of interventions used in this study, nine had $g$ scores significantly greater than zero. One-way ANOVA's further indicated that there were no significant differences among the treatments in terms of effect size. One possible conclusion to be drawn from this is that for reducing
anxiety, one treatment is as good as another. A second possibility is that this meta-analysis did not contain a sufficient number of studies to show significant differences among treatments. The only category of intervention that did not result in the significant reduction of anxiety was the counseling category. Apparently counseling alone is not effective in reducing the anxiety associated with testing situations.

Performance Measures

The performance measures reflected less positive results. This also is consistent with previous reviews (Allen et al., 1980; Brown & Nelson, 1983; Finger & Galassi, 1977; Mitchell & Ng, 1972). Here, only five of the ten categories had \( g \) scores that were significantly greater than zero. One-way ANOVA's run on these scores also failed to disclose significant differences between treatments. Significant gains belonged to the cognitive, relaxation, study skills, and systematic desensitization categories. Those categories where interventions failed to produce statistically significant gains were the cue controlled relaxation, implosion, counseling, covert positive
reinforcement, and the "other" categories.

**General Considerations**

One important consideration is the goal of treatment. If the goal is reduction of subjective discomfort, then most interventions seem to do an adequate job of meeting this goal. If the goal is to increase performance, then half of the categories of interventions reviewed will likely lead to success. It would be important to determine specifically what the desired outcomes are when designing programs to address the problem of test anxiety.

A review of the effectiveness of interventions to decrease test anxiety must deal with the problem of success criterion. An alternative to the subjective decision as to which criterion is most useful is the approach taken in this quantitative review. That is, to perform multiple analyses based on the various available outcome criteria. The reader is then left to decide which, if any, is more applicable or important to the task at hand.

Results of planned comparisons also yielded some interesting results. For both anxiety and performance measures, graduate student-administrators fared better
than did their professor counterparts. Also, it appears that for both classes of dependent variables, individual administration is preferable to administration in groups. This finding needs to be balanced against the savings in therapist contact time that groups afford. In neither case did it matter whether there was random assignment. Perhaps this was because there were included very few studies in the present analysis that did not use a random design.

In two categories the moderator variables made a difference for anxiety measures but not for performance measures. As for pre-measured test anxious status, studies that used these samples had significantly higher $g$ scores than those found in studies that did not. Also, studies that used live administration had significantly higher $g$ scores than studies that utilized a taped administration. In neither of these instances were there significant difference among performance variables.

Conversely, for performance measures, volunteer subjects had significantly higher $g$ scores than did those subjects given some kind of inducement. There were no significant differences on this dimension for anxiety measures.
Methodological Considerations

First, there are methodological recommendations based on the current state of design and reporting of studies. It is important that researchers utilize carefully controlled studies, preferably also controlling for the effects of attention, as with a design that includes a placebo control (Fiske, 1983; Kirschenbaum & Perri, 1982).

Adequate reporting of the methodology used is also an important issue (Fiske, 1983). It is also important that researchers carefully report all the operations that took place in their study. This is very important in order to evaluate the importance of the results or to replicate the study. Very often the reviewer is forced to make inferences based on incomplete information. While often such inferences can be made from context, this is an unnecessary step which can lead to errors and confusion.

Among those areas in which poor reporting is most evident is the failure of these studies to specify precisely and adequately characteristics of the subjects. Examples of characteristics not usually reported are the subject's year in school and the
percentage of subjects classified by gender. Also a problem has been the failure to specify adequately what contingencies surrounded the participation of subjects in the study, i.e., was their participation truly voluntary.

A useful piece of information rarely supplied was the specification of attrition rates. Even less common was the indication of whether these attrition rates were equally distributed across cells.

The first overall recommendation, then, is that researcher's use appropriate methodologies for their studies. From choice of design to implementation the focus should be on quality. Once this has occurred, it is equally important that the study be written in sufficient detail to allow the reader their own interpretation as well as to assist in any efforts to replicate or extend upon findings.

Recommendations for Future Research

One purpose of a meta-analysis is to identify deficiencies among pertinent research studies and to suggest directions that researchers might take (Fiske, 1983; Kazdin, 1985). There are several valid areas of inquiry for future research in this area. First, it is
evident that the preferred method for addressing the negative effects of test anxiety is to use a multimodal treatment. Future research should focus on which specific combinations of treatments are most effective for this purpose. Perhaps the most appropriate question is the now standard "which approaches produce the best results for which students under which condition" (Paul, 1967). For this question, primary research is crucial. Meta-analysis provides a basis for drawing conclusions from the literature, however, the specific conclusions are too far removed from the performance of individual clients to serve as a basis for making decisions about a particular individual. Here the data from single-case studies and analysis of individual performance patterns in group studies on specific sorts of measures may provide better guidelines (Kazdin, 1985).

In addition, the lack of correspondence between anxiety and performance indicators suggests not only that the link between test anxiety and performance may be weaker than previously assumed but also that the apparent common practice in test anxiety outcome research of treating all test anxious students as alike and assuming (at least implicitly) that they all
perform poorly in school may be unfounded (Brown & Nelson, 1983). The belief that test anxiety causes impaired performance has guided attempts to improve academic performance by the direct alleviation of test anxiety (Dendato & Diener, 1986). Although studies (Mitchell & Ingham, 1970; Spielberger, 1962; Spielberger & Katzenmeyer, 1959; Walsh et al., 1968) have reported that highly anxious students receive lower grades and have a higher academic failure rate than nonanxious students of equal intelligence (Paul & Shannon, 1966), some students who are anxious about tests perform quite adequately, while the reverse is true for other students. One area of inquiry would be to define as clearly as possible these distinctions and to ascertain their cause.

The planned comparisons also suggest some interesting questions that could be addressed empirically. One of the many uses of evidence generated by reviews, however equivocal it might be, is to test relationships which may never have been examined by primary researchers (Cooper, 1984). The planned comparisons perform the function that has been described by Strube and Hartmann (1983) as the "predictive function" that meta-analysis can serve.
Potential areas for future primary research that have been generated by the planned comparisons in this study include exploring what specifically accounts for the finding that graduate student or undergraduate counselors are more effective than their Doctoral level counterparts, what aspect of individual treatment makes it more effective than group treatment, and what are the elements that provide for a successful automated treatment program?

Although the experimental literature is rich in its implications for the improvement in performance among test-anxious persons and for reduction in test anxiety, the results of pertinent research are shown to be disappointingly meager in this regard. Perhaps the major difficulty involved in achieving beneficial effects with specific manipulations in the school environment is the pervasive atmosphere of evaluation in the schools. Since test anxiety is related to both individual and situational factors, it may be possible to reduce the debilitating effects of this syndrome by reducing the evaluative component of examinations (Allen, 1972). Until such wholesale changes can be implemented within the current system of college education in this country, researchers should continue
to focus on methods of identifying and reducing the negative effects of test anxiety for the individual student.
APPENDIX
METADATA

Meta-analysis, as a set of statistical procedures for integrating findings from different studies, has had a controversial past. The term "meta-analysis" was coined by Glass (1976) to refer to "the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings" (p. 3). Glass contrasts this statistical approach with the "casual, narrative discussions of research studies which typify our attempts to make sense of the rapidly expanding research literature" (p. 3).

Meta-analysis arose when the pace of new empirical studies outgrew the individual reviewer's ability to organize the increasing amounts of information and to draw conclusions from the myriad of methodologies, variables, and results. There was obviously a need to become more formal and systematic in attempts to organize the literature. Previous methods such as concentrating on several well-designed studies, listing variables that have been shown to have an effect, and "vote taking" had been criticized for their
shortcomings (Light & Smith, 1971). A growing frustration in the social sciences had been a lack of orderly progression and organization of the large amounts of existing information (Rosenthal, 1984). It appeared necessary that for social science knowledge transmitted through research reviews to be objective and believeable "research reviewers must be required to use the same rigorous methodology that is required of primary researchers" (Cooper, 1984, p. 10).

Meta-analysis uses many of the same statistical methods that are employed in primary data analysis, although in meta-analysis, the study becomes the individual unit of analysis. For experimental studies, it may be asked whether there is a consistent experimental effect across studies. If there is shown to be an effect, the attempt is made to estimate the magnitude, or "effect size," of that relationship. In cases where there are inconsistencies in the relationship, an attempt is made to isolate moderator variables that would explain such inconsistencies.

From the initial trickle of articles and studies concerning the meta-analytic techniques has come a flood of publications. Many of these articles were themselves reports of meta-analyses done on a variety
of topics. Many other articles were critiques and commentaries on what had become a major source of disagreement in the field, that is, the question of the utility of meta-analysis as a technique for distilling additional information from already completed studies. Rosenthal (1984) notes the increase from 6 papers on the topic published in 1976, to 120 papers on a similar topic published in 1982.

The flurry of activity surrounding meta-analysis may be evidence for the continuing important role it plays in social science research. Those supporting its usefulness have published both theoretical papers on the techniques (Cooper, 1984; Glass, 1976; Rosenthal, 1984) and have done a variety of meta-analyses themselves. Those most vocal in their opposition to meta-analysis have raised some well-taken cautions as to the appropriateness and limitations of these techniques. Perhaps the only agreement to be reached about meta-analysis is that it has forced those who are responsible for integrating information in our discipline to examine both our techniques for gathering information and our methods of extracting generalizations from the literature. For this reason alone meta-analysis can be viewed as a welcome
There have been a variety of criticisms directed toward meta-analysis. Rosenthal (1984) provides an excellent framework for organizing the various criticisms of meta-analysis. The categories he developed include: sampling bias and the file drawer problem, problems related to loss of information, problems of heterogeneity, problems of independence, and the exaggeration of significance levels. The following is an examination of each of these issues, the adequacy of meta-analytic techniques in addressing the problems, and recommendations for dealing more effectively with these concerns.

The first of these criticisms deals with the possibility that the studies retrieved do not reflect the population of studies conducted. Especially troublesome is a tendency for studies with statistically significant results to be published with more likelihood than studies without statistical significance (Greenwald, 1975). Two things must be taken into account here. The first is that this criticism applies equally well to traditional narrative reviews. The second is that there are procedures available to minimize the impact of this effect. Meta-
analysts are well advised to follow comprehensive guidelines available for reducing retrievability bias (see Cooper, 1984). Rosenthal (1984) also provides a method for estimating the number of unpublished studies (that average null results) necessary to threaten the overall conclusion drawn by the reviewer.

Rosenthal (1984) discusses several ways in which loss of information might be a problem in meta-analysis. The first question centers on the danger of trying to summarize a research domain by a single value such as a mean effect size, while overlooking other relevant moderator variables. In response, a meta-analyst can use as a coded moderator variable almost any dimension on which studies vary. Which dimensions that bear analysis will be pertinent to the particular research domain in question and specific hypotheses on those elements salient in determining differences in outcome. (Cooper, 1984; Rosenthal, 1984).

Another concern dealing with the potential loss of information is that meta-analyses gloss over details. Rosenthal (1984) responds to this by noting that in computing any statistics, details are glossed over. "To summarize means to gloss over details" and that "it is the data analyst's job in the individual study, and
the meta-analyst's job in meta-analysis to gloss well (p. 126)." Rosenthal goes on to state that "providing the reader with all the raw data of all the studies summarized avoids this criticism but serves no useful review function (p. 126)."

Problems of heterogeneity are also cited as potentially troublesome for meta-analysis. One part of this criticism questions the appropriateness of pooling, in a single meta-analysis, studies that have widely diverse independent and dependent variables. Are these differences not great enough to limit the possibilities of useful comparison? Glass and Smith (1978) responded to this question with their observation that it is an "apples and oranges" question, and that indeed apples and oranges are appropriate to mix if we wish to generalize to fruit. He states further that it is valuable to be able to make general statements about fruit, as well as to draw any distinctions between apples and oranges when appropriate. This is the reason for coding study characteristics when performing a meta-analysis.

There is often a problem of independence to contend with. This occurs when the same subjects within a study generate several effect size estimates
and several tests of significance. Some authors have chosen to treat each result from a different dependent variable as though it were a separate study (Glass et al., 1981; Smith et al., 1980). However, Rosenthal (1984) suggests that "the various dependent variables employed in a study should all be examined for clues as to the types of dependent variable that seem most affected and least affected by the independent variable of interest" (page 31) and that "because the various methods of combining probability levels require independent research results, we recommend that in the final overall analysis each sample of subjects or other units normally contribute only a single effect size and a single significance level to the total (page 32). Rosenthal's "mean effect size procedure" (1984) offers a method for averaging the effect sizes for all the dependent variables of each type included in a study.

Another criticism of meta-analysis is that it can lead to exaggeration of significance levels. Rosenthal (1984) reports that "when the null hypothesis is false and, therefore, ought to be rejected, it is indeed true that adding observations increases statistical power. This serves to increase its accuracy and decrease the probability of type II errors. He goes on to note that
when the null hypothesis is really true, adding studies does not lead to increased probability of rejecting the null hypothesis and that adding studies does not increase the size of the estimated effect.

In conclusion, the final chapter is far from being written on meta-analysis. Meta-analysis has not gone unchallenged (Strube, Gardner, and Hartmann, 1985; Wilson, 1985; Wilson and Rachman, 1983). The majority of these commentaries, however, express healthy skepticism rather than dogmatic resistance and have served the useful purpose of increasing critical self-examination among meta-analysts (Strube & Hartmann, 1983). It is most likely that controversies over meta-analysis will continue into the future.

Rather than viewing meta-analysis as a panacea, Rosenthal (1984) cautions that "our procedures are not perfect, we can use them inappropriately, and we will make mistakes" (p. 17). Nevertheless, when used appropriately, meta-analysis provides a more structured and systematic way of reviewing the literature than does the traditional narrative review. Meta-analysts, like their traditional narrative counterparts, make somewhat arbitrary decisions on classifying results, variables, and other aspects of studies With meta-
analysis, these decision rules are explicit and public.

While the controversy is still evident, there is a trend toward greater acceptance (Michelson, 1985; Rosenthal, 1984). Whether this may accurately be termed a "revolution in the making" (Rosenthal, 1984, page 16) there is reason to believe that it does signify an important breakthrough for evaluating research (Kazdin, 1985). That breakthrough may be seen as the resulting self-analysis of research methods, and methods of accumulating data in the social sciences. For this alone, we owe a debt to meta-analysis.
LIST OF REFERENCES


