2012

Replication and Extension of Ellis, Ladany, Krengel, and Shult (1996); Clinical Supervision and Research from 1981 to 1993: A Methodological Critique

Margaret A. Schutt
Lehigh University

Replication and Extension of Ellis, Ladany, Krengel, and Shult (1996); *Clinical Supervision and Research from 1981 to 1993: A Methodological Critique*

by
Margaret A. Schutt

Presented to the Graduate and Research Committee
of Lehigh University
in Candidacy for the Degree of
Doctor of Philosophy
in
Counseling Psychology

Lehigh University
May 2012

Approved and recommended for acceptance as a dissertation in partial fulfillment of the requirements for the degree of Doctor of Philosophy/ Counseling Education.

_____
Date

_____
Arnold Spokane, Ph.D.
Dissertation Chair

_____
Accepted Date

Committee Members:

_____
Arpana Inman, Ph.D.
Associate Professor, Lehigh University
Committee Member

_____
Cirleen DeBlaere, Ph.D.
Assistant Professor, Lehigh University
Committee Member

_____
Janet Muse-Burke, Ph.D.
Associate Professor, Marywood University
Committee Member

ACKNOWLEDGEMENTS

There is neither enough space in this document nor words in the dictionary to appropriately acknowledge those who helped me reach my goals and complete this dissertation. I have traveled a very difficult road in the company of spectacular people who have made the journey easier, my load lighter, my perspective brighter, and my spirit stronger. I wish to acknowledge the following people for their instrumental contribution to the completion of this work:

My committee, Dr. Spokane, Dr. Inman, and Dr. DeBlaere: Thank you for your insight, flexibility, and support. Your guidance and feedback helped me develop my research and document into something of which I can be proud.

Mary Yotter, who has been more influential in the completion of my dissertation than anyone: You know absolutely everything, and you never hesitate to go out of your way to help students. You hold many roles for which you are never fully recognized, including director, consultant, cheerleader, stand-in mother, and friend. Most of us would never graduate if not for you. May God have mercy on the department, and the students, when you retire.

Janet, my supervisor, turned mentor, turned role model, turned dissertation committee member and friend: You have been a source of calming reason, sharp wisdom, and unequaled support during my most difficult times. You are the picture of what a psychologist should be. I am a better psychologist and a better person for having known you, and I am so thankful that you agreed to be on my committee.

Melanie, a great friend and fantastic psychologist who agreed to help me code my data before you even knew what that would entail. I am both genuinely sorry and

gleefully happy that you ignored my advice not to pursue a doctoral degree in psychology. Misery loves company, and there is no one with whom I'd rather share that misery. I look forward to consulting and commiserating with you as we move on in our professional lives.

My parents, whose unwavering belief in me often helped bolster my belief in myself, and whose always-protective inclinations kept me feeling loved—and often laughing. You have never failed to encourage me and support me, and have served as excellent examples of the person I want to be. I only hope that I can do the same for Madeline that you have done for me.

Jen, my twin and best friend, without whom I doubt I would have made it this far. You know me better than anyone else, and you always know exactly what to say—whether to cheer me up, calm me down, or make everything seem ok. I simply cannot imagine how I would survive without you. With so many miles between us, I cherish your daily calls that keep me sane, grounded, and out of bed. And so does the phone company, I imagine.

Stacy, my sister who has been a constant source of cheerleading and humor throughout every one of my endeavors: You fiercely defend and support me whenever I need it—regardless of whether I am in the right, thankfully. You know the intrinsic value of ANTM and other silly things, and keep me from taking myself too seriously. I am so thankful that you are in my life and that you call me sister (but usually Peg).

My grandmother, the original Margaret Schutt, whose character and strength made her someone to admire and whose advice was always timely, wise, and unapologetically inappropriate. I thank you for teaching me to have courage, and for

believing so strongly in my ability to complete my degree. I know that you are watching over me as I complete this chapter of my life.

Joshua, my wonderful husband and best friend, you gave me strength and confidence when I had none. When you told me that we were "in this together", you did not realize that I was secretly including my dissertation. I would not have been able to complete my work without your mathematical brilliance—and your patience for my lack thereof. Not a day goes by that I don't thank God for you and Madeline.

Madeline, my beautiful daughter: As I struggled through the process of completing my dissertation, I looked at you and felt such joy. No matter how difficult or insurmountable things seemed, they became minor when I saw you laughing and playing—you reminded of what is most important in this world. I defended my dissertation on your third birthday, and kept your smiling face in my mind to keep me centered. I love you with all my heart.

TABLE OF CONTENTS

Page

Title Page …… i

Copyright Page……... ii

Certificate of Approval……...…… iii

Acknowledgements…..... iv

Table of Contents……..…… vii

List of Tables……..……... viii

Abstract ……... 1

Chapter I: Introduction……..…….. 2

Chapter II: Literature Review ……..… 16

Chapter III: Method……. 46

Chapter IV: Results…….…….……. 60

Chapter V: Discussion…….……..… 77

References……. 120

Appendices…….……175

A: Articles Reviewed in Current Study…… 175

B: Detailed Calculation procedures for converting to eta squared (η^2)…… 181

C: Test-Specific Procedures Computations used by G*Power……. 182

D: Coding Chart…… 186

Vita…… 189

LIST OF TABLES

Page

Table 1: Journals Reviewed in Ellis et al (1996) and in the Current Study …… 157

Table 2: Extension Articles…….……..... 158

Table 3: Measures Utilized in the Research Articles Included in the Current Study… 160

Table 4: Statistical Variables for Included Studies…….170

Table 5: Prevalence of validity threats in Ellis et al. (1996) and the current study …. 171

Table 6: Top Most Salient Methodological Threats…….…….. 173

Table 7: Inference categories and subcategories…… 174

ABSTRACT

The two purposes of this study were to (a) replicate and extend the methodology of Ellis et al. (1996) by evaluating the supervision literature from 1994 through 2010 and (b) address areas of focus omitted in the study conducted by Ellis and colleagues. In the current study, supervision research articles published from 1994 through 2010 (inclusive) were reviewed and included in the study using the inclusion/exclusion criteria designated by Ellis et al. (1996). A total of 62 studies were evaluated according to 49 threats to validity (Cook & Campbell, 1979; Russell et al., 1984; Wampold, Davis, & Good III, 1990) and 8 statistical variables according to procedures of Ellis et al. (1996). The data revealed consistencies with the findings of Ellis et al. (1996), including similar occurrences of unchecked Type I and Type II error rates and low statistical power. Ellis and Ladany's (1997) identification of six "cardinal inferences" of the supervision literature were partially supported, while three additional inferences emerged.

CHAPTER 1

Introduction

Research is intrinsic to evolution in any field of study. Researchers pose new questions and challenge accepted "truths" in hopes of extending and moving a base of knowledge forward. Research published in peer-reviewed journals offers professionals an opportunity to read new findings and apply gained knowledge to their work. It also encourages professionals to question and analyze their "tried and true" methods so that they do not become complacent and thus not useful. Simply put, research is necessary for progression. Practitioners hope to glean useful knowledge from research findings that they may use to inform their practice, and therefore count on careful research execution by authors of the articles they read.

As with most scholarly fields, psychological research is based on or connected to prior research. Completed work establishes a foundation and relevance for ongoing academic pursuit. Additionally, and of particular importance to psychology, the *practice* of any field can be built on prior research. Too often, however, a single study is conducted in an area of research and is never subsequently confirmed, replicated, or extended. A finding that is demonstrated once but is never repeated afterward is an unstable base from which to derive conclusions. While the conclusions from these studies may be very interesting and valuable, they may lose impact and utility without further exploration. Appropriate repetition means that a previous result will have its scope confirmed and extended (Lindsay & Ehrenberg, 1993). Consequently, replication of research is considered to be a crucial aspect of the scientific method. Replication is typically conducted to verify the results of an earlier study (Cumming, 2008; Lamal,

1990), but replication can also validate, extend, and even elucidate limitations from a prior study. Multiple sources of error exist in research, including human errors of procedure, observation, recording, computation, or reporting (Cumming, 2008; Nelson, Rosenthal, & Rosnow, 1986). With so many areas vulnerable to error, it would be almost impossible for research to be flawless! And, if errors occur in any of the steps in scientific research, the results will be affected. Consequently, it is fair to say that the worth of a study is limited until it is replicated. Even the strongest of empirical investigations could benefit from replication and extension; confirmation of results would be as valuable as identification of limitations. This is not to say that new ideas and theories should be set aside in favor of replication research, of course. New research provides new perspectives, ideas, and excitement to the field. Replication and extension of studies serves to aid this process, solidifying and supporting the ideas that serve as a foundation to further inquiry.

Replication is one way in which a researcher can evaluate the findings reported in previous research. On a larger scale, meta-analysis can assess the research findings in many studies. Where replication advances the understanding of results reported in one study, meta-analysis offers the opportunity to do the same with multiple studies. Meta-analysis, created by Glass (1976), has been most extensively used in psychotherapy outcome literature (see Glass, McGaw, & Smith, 1981). Meta-analysis essentially refers to a method of combining and comparing statistical results across studies, with the purpose being to estimate the true effect size of studies more powerfully than can be assessed in a single study (Glass et al., 1981). Additionally, a meta-analysis provides an opportunity to combine a great number of studies in a statistical manner that allows

discussion and interpretation of them together. This can provide researchers and readers a comprehensive understanding of the state of the literature/research that would otherwise be very difficult to assess. For example, if a researcher was studying all articles of a given topic on which no meta-analysis had been conducted, the researcher would have to find his/her own way to make comparisons and assessments of dozens or even hundreds of articles. Certainly, meta-analysis provides the means for assessing multiple studies more easily and more accurately than by going through one study at a time. But it also provides important data (e.g., power) that aid in the understanding of the state of research and the limits that need to be further addressed.

While the statistical methods of meta-analysis are comprehensive and can certainly stand on their own in analyzing multiple research studies, they are often included in a systematic review of research. Research variables (e.g., design and methodology) that affect the validity of research studies are also examined frequently. These aspects of research studies are as important to assess as are the statistical variables (i.e., power), because the quality and rigor of research affects the outcome and conclusions. As with meta-analysis, it is important to have a systematic approach to analyzing these variables. Where the researcher examining a hundred studies is well served by the clear, consistent procedure of meta-analysis, he or she would also greatly benefit from a systematic means of qualitatively assessing other study variables.

Identifying a system of assessment is intrinsically tied to consistent, unanimously agreed-upon identification and definition of the variables under study. Several researchers have offered specific criteria for assessing validity and quality in a research study (Cook & Campbell, 1979; Campbell, Stanley, & Gage, 1963; Shadish, Cook, &

Campbell, 2002). The most thorough and often referenced description of validity threats was proposed by Cook and Campbell, who identified 33 threats in total, including four main classes of validity with descriptions of the individual validity threats in each class. The first threat class is identified as *statistical conclusion validity* and includes seven threats: low statistical power, violation of assumptions of statistical tests, Type I error, unreliability of measures, unreliable treatment implementation, random irrelevancies in the experimental setting, and random heterogeneity of respondents. The second threat class is identified as *internal validity* and includes 13 threats: history, maturation, testing, instrumentation, statistical regression, selection, mortality, interactions with selection, ambiguity about the direction of causal influence, diffusion of treatments, compensatory equalization of treatments, compensatory rivalry by respondents receiving less desirable treatments, and resentful demoralization of respondents receiving less desirable treatment. The third threat class is identified as *construct validity* and includes 10 threats: inadequate preoperational explication of constructs, mono-operation bias, monomethod bias, hypothesis guessing within experimental conditions, evaluation apprehension, experimenter expectancies, confounding of constructs and levels of constructs, interaction of different treatments, interaction of testing and treatment, and restricted generalizability across constructs. The fourth threat class is identified as *external validity* and includes three threats**:** interaction of selection and treatment, interaction of setting and treatment, and interaction of history and treatment.

Russell, Crimmings, and Lent (1984) offered another set of validity threats, which they referred to as methodological threats to the validity of a study. The threats are divided into six internal and six external validity categories. The six threats to internal

validity are identified as the following: lack of adequate comparison group, no pretreatment assessment, inadequate sample size, variations or confounds in length of training across conditions, failure to randomly assign participants to conditions (non-randomization), and widely discrepant cell sizes. The six threats to external validity are identified as the following: restricted range of dependent variables, non-representative supervisee or supervisor sample, lack of follow-up assessment, use of role play or audiotaped client statements to assess supervised change, exclusive reliance on self-report data, and overly brief training period.

In addition to the methodology and analysis threats described, there is the possibility of problems with the hypothesis itself. As the guiding principle of the research, the hypothesis must be valid for the rest of the research to make sense. Relatedly, Wampold, Davis, and Good (1990) identified four threats to hypothesis validity that are most commonly found in research. The four threats to validity are: (a) inconsequential hypotheses, (b) ambiguous hypotheses, (c) non-congruence of research and statistical hypotheses, and (d) diffuse statistical hypotheses and tests. These threats specifically address inferences about the fit of the research hypothesis with theory and with statistical analyses (Ellis, Ladany, Krengel, & Schult, 1996).

Certainly, conducting research can become quite complex. Conducting 'good' research requires thorough understanding of the possible threats to validity, careful planning and execution of methodology, and appropriate application of statistical tests to analyze the results. Replication and meta-analysis offer further information about conducted research in order to clarify and extend the knowledge base. Practitioners can then use research conclusions to inform their practice. Interestingly, meta-analysis is

conducted frequently on therapy and the therapeutic process, but rarely is conducted on supervision. For example, this researcher used *Psychinfo* to conduct a very quick search of literature during the past 30 years. The search revealed at total of 99 articles that used meta-analysis to assess psychotherapy or aspects of the psychotherapy process. Conversely, the search resulted in only 3 articles that used meta-analysis to investigate supervision practice. Admittedly, this is not a comprehensive reflection of the literature published or conducted during the past 30 years, but it is an indicator of the needs of supervision research. Like counselors, supervisors refer to research to understand the supervisory process and inform their work. It is therefore important that supervision research receive review and analysis.

Supervision Research

Supervision is considered to be integral to the counseling psychology profession (e.g. Bernard & Goodyear, 1992; Borders & Brown, 2005; Callahan, Almstrom, Swift, Borja, & Heath, 2009; Ladany & Ellis, 1997; Watkins, 1998). A descriptive and inclusive definition of supervision—one that reflects the weight and responsibility the role carries—is offered by Bernard and Goodyear (1992):

> "…an intervention that is provided by a senior member of a profession to a junior member or members of that same profession. This relationship is evaluative, extends over time, and has the simultaneous purposes of enhancing the professional functioning of the junior member(s), monitoring the quality of professional services offered to the clients she, he, or they see(s), and serving as a gatekeeper for those who are to enter the particular profession." (p. 4).

Certainly, supervisors juggle a number of roles, and each role can significantly affect the development of the trainee and the therapy outcome for clients. Supervision has been purported to be one of the primary methods through which counselors are trained (Ladany, Friedlander, & Nelson, 2008; Watkins, 1998; Westefeld, 2009). With the realization of the importance of supervision, the number and scope of research studies in the supervision area has arguably been increasing since Bernard introduced his model for supervision (1979) and Bordin (1983) first applied his concept of the working alliance to the process of supervision. According to Inman and Ladany (2008), supervision research has gained momentum since the publication of the first Handbook of Psychotherapy Supervision. The authors report that:

> …the 1980s had a total of 185 articles addressing psychotherapy supervision (97 theoretical, 28 dissertations, and 60 empirical), whereas the 1990s saw an increase in publications by approximately 60% (i.e., a total of 291 articles—190 theoretical, 29 dissertations, and 72 empirical; p. 500).

Certainly, psychologists recognize the value supervision provided to the therapist and clients, if not the supervisor himself or herself. Inman and Ladany (2008) also found that there was only a 4% increase in psychotherapy-based supervision articles from 2000 until the writing of their review. Inman and Ladany (2008). Review of the supervision literature since 1983 reveals a broad array of topics, participants, purposes, and methods (e.g., Inman & Ladany, 2008; Ellis & Ladany, 1997). Inman and Ladany (2008) identified a pattern of research interests in this area; specifically, they found that the investigators in the 1980s produced research about topics including supervision models, supervisee variables, parallel process in supervision, and the impact of psychotherapy

supervision on client outcome, with a large amount discussing theoretical and conceptual issues in supervision (p. 500). Inman & Ladany also note a shift in pattern in the 1990s as a consequence of supervision being identified as intrinsic to psychologist training as per the American Psychological Association's Committee on Accreditation (COA, 1996; 2000). Subsequently, research topic areas began to expand and encompass a span of variables and topics. There are now supervision studies that explore specific relationship issues in supervision, such as conflict in supervision (e.g., Moskowitz & Rupert, 1983; Nelson & Friedlander, 2001), critical incidents in supervision (e.g., Ellis, 1991), positive and negative experiences in supervision (e.g., Allen, Szollos, & Williams, 1986; Ramos-Sanchez, Esnil, Goodwin, Riggs, Touster, Wright, Ratanasiripong, & Rodolfa¸ 2002; Worthen & McNeill, 1996), working alliance (e.g., Ladany & Lehrman-Waterman, 1999), supervision theory and process (e.g., Bernard, 1979; Ellis & Ladany, 1997; Heppner & Roehlke, 1984; Ladany et al., 2008; Loganbill et al., 1982; Shanfield et al., 1993), parallel process (Doehrman, 1976; McNeill & Worthen, 1989;), and role conflict (e.g., Olk & Friedlander, 1992).

Supervisor behaviors have also been investigated including such topics as supervisor disclosure (e.g., Ladany, Walker, & Melincoff, 2001) successful and unsuccessful supervisor behaviors (e.g., Dressel, Consoli, Kim, & Atkinson, 2007) and supervisor style (e.g., Dow, Hart, & Nance, 2009; Ladany et al., 2001). Trainee and supervisee variables have been explored in such areas as trainee disclosure and nondisclosure (e.g., Ladany, Hill, Corbett, & Nutt, 1996; Ladany & Melincoff, 1999), trainee anxiety and conceptual level (Birk & Mahalik, 1996), impact of supervision on self-efficacy (e.g., Cashwell & Dooley, 2001), and counselor experience (Ladany,

Marotta, & Muse-Burke, 2001). Group supervision has been investigated regarding trainee feedback (e.g., Coleman, Kivlighan, Jr., & Roehlke, 2009), hindering phenomena (e.g. Enyedy, Arcinue, Puri, Carter, Goodyear, & Getzelman, 2008), multicultural group supervision (e.g., Gainor & Constantine, 2002). Culture and diversity, an area that has grown significantly in the past 15 years, includes such areas as the impact of race and culture on supervisory process (e.g., Ancis & Ladany, 2010; Bhat & Davis, 2007; Constantine, Warren, & Miville, 2005; Hilton, Russell, & Salmi, 1995), and multicultural counseling competence (e.g., Gloria, Hird, & Tao, 2008; Inman, 2006; Ladany, Inman, Constantine, & Hofheinz, 1997a). The topic of marital and family therapy supervision includes such constructs as best and worst experiences (e.g., Anderson, Schlossberg, & Rigazio-DiGilio, 2000), and supervision practices (e.g., Carlozzi, Romans, Boswell, Ferguson, & Whisenhunt, 2001). For a more complete and thorough examination of this literature, see Inman and Ladany (2008).

Certainly there are more topics explored in supervision than there is room to describe here. Clearly, however, supervision research is an area that has been enthusiastically explored in recent years. Given the number of extant research articles, it is even more important than ever to be able to summarize and integrate these findings so that useful conclusions and comparisons can be made.

Rigor in Supervision Research

There is no question that the research must be accurate so that the practice is effective and appropriate. For counselor supervision, this is especially critical. If counseling supervision is based on faulty research or unfounded conclusions, it is fair to assume that counselor trainees could develop skills and practices that are no better than

having no supervision at all. More concerning, the practices may even be unwittingly harmful in nature; therefore, the mental health of clients quite literally depends on solid supervision research.

> If we can build a strong supervisory relationship, assess the supervisee's needs and level of development, and not be afraid to offer constructive criticism as well as praise, then I believe the supervisory process will be improved for all concerned—the trainee, the supervisor, and ultimately the client. (Westefeld, 2009, p. 315)

Ellis, Ladany, Krengel, and Schult (1996) examined supervision literature from 1981 to 1993. They stated two purposes: (a) to assess the status and scientific rigor of clinical supervision research from 1981 through 1993, and (b) to determine the extent to which supervision researchers have responded to the suggestions of the most recent comprehensive methodological review (Russell et al., 1984). The researchers examined each study against three sets of criteria previously described, which include: the 33 threats to validity proposed by Cook and Campbell (1979), the 12 methodological threats described by Russell and colleagues (1984), and the 4 hypothesis threats identified by Wampold and colleagues (1990). Ellis et al. (1996) performed a meta-analysis on the supervision literature that revealed significant concerns about the usefulness of the previous supervision studies.

The results of their study are sobering. Ellis and colleagues (1996) found threats to validity in every study they examined (see Table 1). In a follow-up study, Ellis and Ladany (1997) conducted a more focused review of the same literature base and organized the findings according to the inferences under investigation in each study.

They reported that 9 of the 16 most salient threats were to statistical conclusion validity construct validity, internal validity, hypothesis validity and external validity (Ellis & Ladany, p.41). In review of the numerous threats identified in the studies, the authors presented seven "most plausible rival explanations for data and results" (p. 458). These explanations are as follows: (a) Experimentwise Type I (incorrectly rejecting the null hypothesis in favor of the alternative) error was found in 72% of the studies (b) In 62% of the studies, the measures used in the studies were not psychometrically sound. Additionally, 83% of the studies were conducted with measures that were not developed for a clinical supervision context (c) Experimentwise Type II error (the false acceptance of the null hypothesis) was found in 50% of the studies. This is due to the fact that 91% of the investigators did not attempt to systematically control Type I or Type II error rates. Consequently, they were unlikely to detect true effects (d) In 43% of the studies samples were nonrandom or not representative of the target population. Therefore, incorrect inferences were drawn regarding the hypothesis because the sample did not reflect it. (e) Nonrandom assignment to treatment conditions occurred in 40% of the studies reviewed, which skews the data and leads to incorrect conclusions. (f) In 26% of the studies, clear inconsistencies were identified among the purpose, hypotheses, design–method, and analyses, leaving the results largely unusable.

Overall, the review conducted by Ellis and colleagues is disconcerting. The problems evident in the supervision literature leave practitioners wondering what information to trust and where to look for accurate and useable information on their roles as supervisors. Based on their findings, Ellis and colleagues (1996) provided recommendations to researchers for good research design, including a sample design for

reference. They urged researchers to attend to the validity threats that were ubiquitous in the literature and take steps to avoid the mistakes made by previous researchers.

Since Ellis et al. (1996), no large-scale examination of the supervision literature has been published. In 2008, Ellis, D'Iuso, and Ladany published a thorough review of one of the areas investigated by Ellis et al. (1996): supervision assessment, measurement, and evaluation of clinical supervision. In their chapter in *Psychotherapy supervision: theory, research, and practice* (2008), the authors describe their examination of the supervision assessment literature published between 1995 and 2007 (after the period examined in the last review, Ellis et al., 1996), employing the same methodology as before. While Ellis et al. (2008) did find improvement in the standards used by researchers for establishing psychometric properties of measures, they unfortunately also found continued flaws in the research. In their words, they discovered that "…researchers and editors continue to use or endorse substandard procedures to construct and test the validity of new and existing measures for clinical supervision" (Ellis et al., 2008, p. 496).

Given that fourteen years have passed since Ellis et al.'s (1996) analysis, and considering the findings of Ellis, D'Iuso, and Ladany (2008) in their examination of supervision assessment literature, it seems particularly timely and important to conduct a replication of Ellis and colleagues' work (1997) to assess if researchers have applied the recommendations put forth by those authors. Ellis, D'Iuso, and Ladany (2008) identified ongoing problems with and recommendations for supervision assessment research, which provides researchers and practitioners valuable information about conducting research as well as using current research. Examination of the remaining areas of supervision

research is an appropriate and important next step in understanding the quality of published supervision research.

Purpose of the Current Study

Supervision research is essential for the growth and development of supervision practice. It is important to ascertain the degree to which authors heed the findings and recommendations from prior studies (Ellis et al., 1996; Ellis & Ladany, 1997). Initially, Ellis et al. (1996) examined the degree to which researchers applied the recommendations put forth by Russell et al. (1984). They found that authors did not follow the recommendations but instead continued making mistakes in research design and interpretation. The authors' conclusions and recommendations about the state of supervision research, based on thorough and meticulous examination, is a call to the profession that changes in supervision research are necessary.

The present study is an examination of the supervision literature published since Ellis et al.'s (1996) methodological critique. This study replicates the methodology employed by Ellis and colleagues (1996), which includes conducting a power analysis and examining the methodology of the supervision literature published subsequent to Ellis et al.'s (1996) methodological critique. The current study also extends the work of Ellis and colleagues (1996) by expanding the search to include sixty-eight journals (many of which were not included in the original study simply because they were not yet in publication). In addition, the researcher utilized the methodology outlined by Ellis and Ladany (1997) to organize the reviewed research studies.

Hypotheses

Two hypotheses guided the current study. They are as follows:

H_1: It was hypothesized that the literature published from 1994 through 2010 would reflect improvement from the research studies reviewed by Ellis et al. (1996) and Ellis and Ladany (1997). Specifically, it was hypothesized that the literature would reveal a more careful approach to study design with attention to minimizing threats to validity, methodology, and hypotheses. This improvement was hypothesized to occur either due to researchers reading and employing recommendations from Ellis et al. (1996) and Ellis and Ladany (1997) and/or due to increased sophistication with research design that may occur naturally as the topic of supervision areas is explored and refined over time.

H_2: Ellis and Ladany (1997) identified six major themes (which they called "cardinal inferences") in their review of the supervision literature. It was hypothesized that the supervision literature examined in this study would support the *Cardinal Inferences* presented by Ellis and Ladany (1997). The authors derived these inferences from the 144 studies examined by Ellis et al. (1996) and the additional 13 studies added by Ellis and Ladany (1997). The authors report these as the major themes in clinical supervision research, and as such, it was hypothesized that they would maintain a presence in the current review. The current study investigated the fit of these inferences to the current literature. The inferences were modified to reflect the progression and focus of the recent supervision literature. In addition, the topics of multicultural competence, supervisor training, and use of technology in supervision emerged as new inferences. Consistent with Ellis and colleagues (1996), this researcher identified the central inference of each research study analyzed.

CHAPTER II

Literature Review

The general process of supervision essentially refers to the practice of directing and inspecting. Directing involves teaching, coaching, and modeling expected behaviors and practices, and inspecting involves critical and careful examination of said behaviors and practices. In most environments, supervisors follow a prescribed means of teaching, cover a specified amount of information and material, and have concrete means of measuring the success or weakness of their supervision of the trainee. The supervision of psychotherapy, however, is more complex and includes more variables than can be generically assessed with one concrete measure. One's practice of psychology is highly individual, and emerges from multiple variables that include such things as theoretical orientation, training, personal history, and personal style. Subsequently, supervision of psychotherapy is compound, as it integrates the variables of the trainee and the variables of the supervisor (as well as the variables of the clients). But the complexity of the supervisory relationship does not detract from its utility and importance. Ladany and colleagues (2008) assert that psychotherapy supervision is the principal educational vehicle through which people learn to become therapists. Since the goal of supervision is to help the therapist become more effective and skilled, the supervision process is arguably one of the most critical aspects of a counselor trainee's professional development. As such, it is important to study and understand supervision as thoroughly as possible so that supervisors can employ the most effective styles and strategies. Given the importance of the role of supervision, it is clear that research on psychotherapy supervision is essential to the practice of psychotherapy.

Assessing the Quality of Research

The threats to the validity of research, research design, and hypothesis development have been clearly articulated and published in the psychology literature (Campbell, 1957; Campbell, Stanley, & Gage, 1963; Cook & Campbell, 1979; Russell et al., 1984; Shadish, Cook, & Campbell, 2002). Therefore, one might assume that all researchers have read and adequately attended to/controlled for these possible threats. The question then is whether or not research published in scholarly, peer reviewed journals can be assumed to be "good". According to Ellis et al. (1996), the answer is, unfortunately, no. In fact, published literature contains multiple problems, including incorrect inferences, methodological and statistical flaws, and data misinterpretations (Ellis et al., 1996; Ellis & Ladany, 1997; Holloway, 1982). Many researchers recognize this problem, and supervision literature contains appeals to researchers for increased scientific rigor in supervision and training research. Prior to 1996, 32 reviews of supervision and counselor training articles were published in scientific literature. Ironically, none of these reviews actually evaluated the methodological or scientific rigor of the studies. According to Ellis et al. (1996), this is problematic because it "…may lead to equating or outweighing findings of excellent research with poor research, exacerbating theoretical ambiguity in the field, and/or drawing inaccurate inferences and conclusions" (p. 35). Since the goal of supervision is to help the supervisee/trainee become more effective and skilled, the supervision process is arguably one of the most critical aspects of a counselor trainee's professional development (Ladany & Ellis, 1996; Ladany, Friedlander, & Nelson, 2005; Olk & Friedlander, 1992; Watkins, 1998). As the

practice of quality supervision is informed and guided by research conclusions, these conclusions and inferences must be based on solid methodology.

In order to assess accurately the rigor of research, it is necessary to identify a universally accepted means of defining "good research." The contributions of Campbell, Stanley, Cook and Shadish (Campbell, 1957; Campbell & Stanley, 1963; Cook & Campbell, 1979, Shadish, Cook, & Campbell, 2002) to the experimental design literature are considered the most influential in the field. Campbell (1957) first defined the concepts of internal and external validity, which were expanded by Campbell and Stanley (1963). In 1979, Cook and Campbell further expanded these two types of validity into four components: statistical conclusion validity, internal validity, construct validity, and external validity. Cook and Campbell identified individual threats to each component of validity, reaching a total of 33 potential threats. The identification of threats to validity should assist researchers in anticipating areas where their experimental designs may threaten validity of their studies. And if researchers cannot control for these threats, then they should at least be able to discuss and defend their design choices to the research community. These validity components and their associated threats are described below.

*Statistical conclusion validity***.** Statistical conclusion validity refers to "the appropriate use of statistics to infer whether the presumed independent and dependent variables covary" (Shadish, Cook, & Campbell, 2002, p. 37). To continue, "statistical conclusion validity "…concerns two related statistical inferences that affect the covariation component of causal inferences" (Shadish et.al., 2002, p. 42). The first inference is whether the presumed cause and effect covary; an incorrect conclusion that cause and effect covary when they do not will result in a Type I error, while an incorrect

conclusion that they do not covary when they actually do will result in a Type II error. The second inference regards how strongly presumed cause and effect covary. In this inference, it is possible to overestimate or underestimate the magnitude of covariation. (Shadish et al., 2002). Cook and Campbell (1979) identified 7 threats to statistical conclusion validity.

The first threat is *low statistical power*, which is the probability of detecting a true effect and is determined by sample size, per comparison alpha, and population effect size. The second threat is *the violation of assumptions of statistical tests*, for example, heterogeneity of variances. The third threat is *Type I error*, which is an erroneous statistically significant effect or multiple statistical comparisons with no adjustment of the alpha level. The fourth threat is *unreliability of measures*, referring to reliability coefficients below .80 or unknown reliability in a supervision context. The fifth threat is *unreliable treatment implementation*, such as supervision interventions given differently to trainees. The sixth threat *is random irrelevancies in the experimental setting* such as third-variable problems in setting in the assessment of an aspect of supervision over the course of a semester. The seventh threat is *random heterogeneity of respondents* such as third-variable problems in participants (e.g., not controlling for experience level; Cook & Campbell, 1979).

Internal validity. Internal validity refers to whether the covariation of independent and dependent variables resulted from a causal relationship or whether it was simply by chance (Shadish et al., 2002). Internal validity directly addresses whether an experimental treatment/condition makes a difference or not, and whether there is sufficient evidence to support the claim. Campbell and Stanley (1963) stated that internal validity refers to

inferences about whether "…the experimental treatments make a difference in this specific experimental instance" (Campbell & Stanley, 1963, p.5). These authors also asserted that although ideally speaking a good study should be strong in both internal and external validity, internal validity is indispensable and essential. In contrast, the question of external validity is never completely answerable (Campbell & Stanley, 1963).

The first threat identified is *history*, which refers to the effects that occur as a result of events happening between pretest and posttest. The second threat is *maturation*, which is the effect occurring as a result of participants maturing or becoming more experienced between pretest and posttest. The third threat is *testing*, where study effects are influenced by participant familiarity with a test given multiple times. The fourth threat is *instrumentation*, which refers to ceiling or floor effects. The fifth threat is *statistical regression*, where pretest-to-posttest changes are due to regression to the mean. The sixth threat is *selection*, specifically nonrandomization of the sample. The seventh threat is *mortality*, which refers to differential dropout of participants among treatment conditions. The eighth threat identified is *interactions with selection*, where the selection of the sample (e.g., nonrandomization) interacts with other threats, such as maturation, history, or instrumentation. The ninth threat is *ambiguity about the direction of causal influence*, where it is unclear whether the independent variable influences the dependent variable or the dependent variable influences the independent variable. The tenth threat is *diffusion of treatments*, where participants in the control group learn about the interventions in the experimental groups. The eleventh threat is *compensatory equalization of treatments*, where supervisors attempt to equalize participants in less desirable treatments. The twelfth threat is *compensatory rivalry by respondents receiving*

less desirable treatments, where the control group participants change their behavior positively in response to the experimental group's more positive treatment. The thirteenth threat is *resentful demoralization of respondents receiving less desirable treatment*, where the control group participants change behavior negatively as a result of feeling demoralized in comparison with the experimental group's advantage (Cook & Campbell, 1979).

Construct validity. Construct validity addresses the generalization from "…the samples of persons, settings, and times achieved in a study to and across populations about which questions of generalization might be raised" (Shadish et al., 2002, p. 38). It is the validity of "…inferences about the higher order constructs that represent sampling particulars" (Shadish et al., 2002, p. 38).

The first threat is *inadequate preoperational explication of constructs*, such as inadequately defining key constructs. The second threat is *mono-operation bias*, which can occur by assessing a construct with only one measure. The third threat is *monomethod bias*, which occurs when one construct is assessed using only one method. The fourth threat is *hypothesis guessing within experimental conditions*, where participants guess what experimenters want them to do and behave accordingly. The fifth threat is *evaluation apprehension*, which essentially results in participants behaving in a socially desirable manner. The sixth threat is *experimenter expectancies*, where rater awareness of the research hypotheses biases their ratings. The seventh threat is *confounding of constructs and levels of constructs*, such as dichotomizing a continuous variable. The eighth threat is the *interaction of different treatments*, an example of which may be exposure to two treatments that results in a synergetic effect. The ninth threat is

an *interaction of testing and treatment*, such as participants reacting to pretesting. The final threat is *restricted generalizability across constructs*, meaning that there were too few potential constructs assessed that were affected by a treatment (Cook & Campbell, 1979).

*External validity***.** External validity addresses the generalization from "…operations to constructs, with emphasis on cause and effect constructs" (Shadish et al., 2002) It is essentially the generalizability of the treatment/condition outcomes to other conditions/settings/situations, and considers whether the same result of a given study can be observed in other situations. Campbell and Stanley (1963, p. 5) defined external validity as asking, "…to what populations, settings, treatment variables, and measurement variables can this effect be generalized?" Campbell and Stanley (1963) assert that external validity can never be conclusively reached; regardless of the number of cases that prove external validity, it only takes one disconfirming case to weaken external validity.

The first threat is *interaction of selection and treatment*, which results in limited generalizability of the experimental effect to as well as across other samples of people. The second threat is the *interaction of setting and treatment*, which results in limited generalizability of the experimental effect to as well as across other settings. The third threat is *interaction of history and treatment*, which results in limited generalizability of experimental effect to as well as across other times (Cook & Campbell, 1979).

Hypothesis validity. In addition to the methodology and analysis threats described, there is the possibility of problems with the hypothesis itself. As the guiding principle of the research, the hypothesis must be valid for the rest of the research to make sense.

Wampold, Davis, and Good (1990) designated the term "hypothesis validity" to refer to, "…the extent to which research results reflect theoretically derived predictions about the relations between or among constructs" (p. 360). Therefore, a study with adequate hypothesis validity will "inform theory," whereas a study with inadequate hypothesis validity creates ambiguity and uncertainty about the relationship between constructs (p. 360). The authors identified four threats to hypothesis validity that are most commonly found in research. Their first threat to hypothesis validity *is inconsequential research hypotheses* (p. 361). The authors suggest that for any given theory, multiple implications can be made. The question for the researcher, then, is whether or not his or her hypothesis about the given theory is a "crucial issue"; is it central to proving the given theory? (p. 361). The authors state:

> ...For example if theory T2 implied I21, which was identical to I 11, then any experimental result corroborating T1 would also corroborate T2. The hypothesis validity of a study is strengthened when the number of tenable theories that have implications similar to I11 is small. Ideally, corroborating T1 should simultaneously falsify a large number of competing theories (p. 362).

The key is determining the "crucial question" (Wampold et al., 2001) about the theory. The process by which this question is determined involves examining existing literature and knowledge about the theory and asking an important unanswered question. Inconsequential hypotheses do not consider or address the current literature or knowledge base and therefore do not lead to a "…convergence of knowledge" (p.362). Wampold and colleagues cite Platt (1964) as advocating the use of multiple hypotheses as one means of combatting inconsequential hypotheses (Platt, 1964; Wampold et al., 2001).

The second threat to hypothesis validity described by Wampold and colleagues (1991) is *ambiguous research hypotheses*. For example, "…If the experimental expectation *X* is not specified sufficiently, it may well be impossible to determine whether the obtained results *D* are similar or dissimilar to what was expected" (p. 363). And since the hypothesis presents an unanswered question about a theory, an ambiguous hypothesis leads to data that can neither confirm nor disconfirm the theory under investigation. The authors describe such research hypotheses:

> Ambiguous research hypotheses are often stated in journal articles with phrases such as 'the purpose of the present study is to explore the relation between . . .' or 'the purpose is to determine the relation between. . . .' In one sense, such research cannot fail, because some relation between variables will be 'discovered,' even if the relation is null (i.e., no relation). In another sense, the research will always fail because the results do not falsify or corroborate any theory about the true state of affairs (p. 363).

Essentially, a hypothesis should be specific and now exploratory in nature; researchers will certainly find *something*, but that something may simply be a result of chance.

Wampold and colleagues' (2001) third threat to hypothesis validity is *noncongruence of research and statistical hypotheses.* Simply put, the statistical hypothesis must correspond to the research hypothesis if any meaning is to be made from the results. When the research and statistical hypotheses are incongruent, even persuasive statistical evidence (small alpha levels, high power, and large effect sizes) will not allow valid inferences to be made about the research hypotheses (Wampold et al., 2001, p. 363).

The fourth and final threat to hypothesis validity identified by Wampold and colleagues (2001) is *diffuse statistical hypotheses and tests.* Wampold and colleagues (2001) describe three ways in which diffusion of statistical tests occurs. *First, use* of multiple statistical tests can result in theoretical ambiguity. This occurs when a research hypothesis is tested by many statistical tests (i.e., broken down into multiple statistical hypotheses). The authors state this as problematic because the results of the statistical tests may not be consistent and therefore interpretation of the group of results is not clear. For example, two tests may result in two conflicting results—how does a researcher interpret the evidence? Additionally, there is a problem in controlling for Type I and Type II error; control of Type I will lead to greater Type II error, so it is impossible to tell which results may be due to one type of error. Second, the use of Omnibus tests can threaten validity. According to the authors, According to the authors, omnibus tests are problematic because they "…contain effects, contrasts, or combinations that do not reflect solely the research hypothesis" (Wampold et al., p.364). The authors contend that the most focused test would be a multivariate planned comparison (Wampold et al., p. 365). Third, the inclusion of extraneous independent variables can threaten validity. While these variables are often added to increase generalizability, it also "…inflates the number of hypotheses tested, increasing the diffusion of the statistical tests" (Wampold et al., p. 366).

Methodological Threats. Russell, Crimmings, and Lent (1984) emphasized the importance of supervision and the apparent lack of formalized supervision training as of the writing of their paper. Their intent was to organize and clarify the knowledge to date within the supervision field. The authors presented an overview of supervision, which

includes theory, techniques, and literature review. In the course of their literature review, they identified a total of 12 methodological threats to quality supervision research, which consist of six threats to internal validity and six threats to external validity. Ellis and colleagues (1996) and Ellis and Ladany (1997) utilized these categories in their evaluation of the supervision literature. The threats to internal validity include the following: (a) lack of adequate comparison group, (b) no pretreatment assessment, (c) inadequate sample size, d)) variations or confounds in length of training across conditions, (e) failure to randomly assign participants to conditions (non-randomization), (f) widely discrepant cell sizes (suggesting that the homogeneity of variance assumption may have been violated). The threats to external validity include the following: (a) restricted range of dependent variables, (b) non-representative supervisee or supervisor sample, (c) lack of follow-up assessment, (d) use of role play or audiotaped client statements to assess supervised change, (e) exclusive reliance on self-report data, and (f) overly brief training period (Russell et al, 1984, p. 644).

Ellis and colleagues (1996)

The primary purpose of the investigation conducted by Ellis and colleagues (1996) was to assess the status and scientific rigor of clinical supervision research published from 1981 to 1991. Reasons to review and examine literature include identifying gaps in the literature, avoiding reinventing the wheel, extending current knowledge, identifying seminal works, identifying opposing views, identifying the derivation and statistical testing of overall factors/effect size parameters in related studies, generalizing to the population of studies, and simply dealing with the large amount of articles published each year.

Ellis and colleagues sought to "…provide quantitative operational definitions of accepted standards regarding the clinical supervision studies (e.g., sample size, effect size, statistical power, and per comparison and experimentwise error rates)" and "…to aggregate the quantitative data both across statistical tests and studies in order to allow comparison with previous statistical reviews" (Ellis et al., 1996, p. 36). The second purpose of the study was to ascertain the extent to which supervision researchers have responded to the suggestions of the most recent comprehensive methodological review by Russell et al. (1984). In this work, Russell and colleagues identified 12 threats to research studies, and it was the hypothesis of Ellis et al. that there would be a significant reduction of these threats in the literature subsequent to the publication of Russell et al.'s study. Ellis and colleagues utilized statistical variables of effect size, statistical power, and per comparison and experiment-wise errors to evaluate studies.

The overall findings in the study pointed to serious flaws in the research methodology of the supervision literature. The authors found violations of the methodological threats put forth by Russell et al. (1984). Ellis and Ladany replicated and extended their 1996 study one year later. It provided a more in-depth view of the studies that had been examined. The research was divided into categories more specific to the way in which supervision is understood and described each of the studies included according to their value and flaws. The value of this assessment is eye-opening; many published studies were so riddled with flaws as to be almost unusable to the reader. For example, they described a particular set of 13 studies as "…so seriously methodologically flawed… that trustworthy inferences could not be made from the results" (Ellis & Ladany, 1997, p. 473).

Ellis, Ladany, Krengel, and Schult initially conducted their meta-analysis of supervision literature in 1996. In 1997, Ellis and Ladany published an extension and replication that was extremely thorough in it its description of the studies used. Most valuable in the study is the findings about research shortcomings and the subsequent recommendations for future research. Their conclusions are particularly disconcerting, as most of the 144 studies examined were seriously flawed. This leaves a reader an exceedingly difficult task of trying to identify what conclusions are worthwhile and what conclusions are useless. Since the outcome of their examination suggested that supervision literature leaves much to be desired, it is important to assess the subsequent literature to ascertain if research quality has improved according to the standards and recommendations put forth by Ellis and Ladany. In replicating the study, the researcher adhered to the methodology utilized by Ellis et al. (1996).

Methodological evaluation variables. Ellis and colleagues (1996) evaluated each study in terms of 49 potential threats to the validity of the results. Included in these 49 threats were 4 classes of validity (and threats to each) identified by Cook and Campbell (1979), and the 12 methodological threats identified by Russell and colleagues (1984). Ellis and colleagues also identified supplemental evaluation criteria, which they organized into four sections. The first section addressed whether investigators did the following: explicitly (or implicitly) tested theory or models, presented explicit (or implicit) research hypotheses, used psychometrically sound measures, tested developmental inferences, or acknowledged the limitations of their research. The second section classified the type of research design: Experimental (randomization and manipulated independent variable), quasi-experimental (nonequivalent groups and

manipulated independent variable), ex post facto (nonequivalent groups and independent variable not manipulated), empirical case study, or scale development. The third section addressed whether there were inconsistencies among any of the following: stated purpose, research hypotheses, method, design, procedure, or data analyses (Ellis et al., 1996).The fourth section identified the most salient validity threats for each study.

The sample utilized by Ellis and colleagues (1996) included 2017 potential supervision articles, which they reported were identified through *Psychological Abstracts* and related databases (e.g., Educational Resources Information Center; ERIC), as well as from previous reviews (i.e., the ancestry approach; Cooper, 1989) and a systematic search of periodicals that routinely publish research on clinical supervision. The final sample of articles was published in six different journals, which included the following: *The American Journal of Psychiatry*, *The Clinical Supervisor*, *Counselor Educations and Supervision, Journal of Counseling Psychology, Professional Psychology: Research and Practice*, and *Psychological Reports.* Consistent with Bernard and Goodyear (1992), supervision was defined as an intensive interpersonally-focused relationship in which one or more persons are designated to facilitate the development of therapeutic competence in the other person or persons.

Ellis and colleagues (1996) identified multiple validity threats in every supervision study they examined. The threats to statistical conclusion validity were as follows: The average sample size per test was half that typically found in the counseling psychology literature; 80% or more of the 144 studies were judged to have inflated Type I or Type II error rates or unreliable dependent or independent measures; and 60% or more of the studies had data that did not violate statistical assumptions, did not evidence

irrelevancies in experimental setting, or did not unreliably implement the treatments. Regarding internal validity, selection bias was identified in 77% of studies and ambiguity of causal direction was identified in 69% of the studies. The threats to construct validity included monomethod bias in 79% of the studies, confounding of the construct with limited levels of the construct in 69% of the studies, and inadequate preoperational explication of the constructs in 69% of the studies. The researchers found that all threats to external validity were found in more than 82% of the studies.

Applying Wampold and colleagues' (1990) threats to hypothesis validity, Ellis and colleagues found more problems. In 83% of the studies, the authors identified inconsequential hypotheses and in 80% of the studies, ambiguous hypothesis were identified. In a startling 99% of the studies the authors identified diffuse statistical hypotheses and tests. Further, research hypotheses were explicated in 20% of the studies but left implicit in 37% of the studies. Another surprising finding is that 85% of the studies were conducted with measures that were psychometrically inadequate for a clinical supervision context. In 73% of the studies, a mismatch existed among the purpose, hypotheses, design, method, procedure, and statistical analysis. In the methodology of the studies, Ellis and colleagues found that 7% of the studies evidenced variations or confounds in length of training, non-representative supervisee or supervisor samples, use of role play or audiotaped client statements to assess change, or overly brief training length. Additionally, 78% of the studies had inadequate sample sizes and 66% relied exclusively on self-report data (Ellis et. al., 1996).

Ellis and Ladany (1997)

Ellis and Ladany (1997) replicated and extended the 1996 study by Ellis and colleagues. The purpose of this study was to reanalyze the original data (144 studies on supervision) and examine them in a more useful paradigm. No statistical analyses were performed in this study due to the overlap with Ellis et al. (1996). The authors sought to understand the state of research in each of the main areas of supervision research and organize the reviewed studies accordingly so as to be more easily understood and utilized by readers. The authors agreed upon six main categories of research into which they could place all studies, giving them the opportunity to examine research rigor and needs in each of the six identified categories. The authors referred to these categories as the "six cardinal inferences" of supervision literature. They include the following: 1) Inferences about the supervisory relationship (with subcategories of inferences about social influence theory, client-centered conditions, Strong's (1968) Social Influence Theory, role conflict and ambiguity, structure of the supervisory relationship, 2) inferences entailing matching in supervision, 3) inferences regarding supervisee development, 4) inferences relating to supervisee evaluation, 5) inferences about client outcomes in supervision, and 6) inferences about supervisees: new measures.

First cardinal inference: Inferences about the supervisory relationship. Ellis and Ladany (1997) assert that the onus of the supervisory relationship has been attributed either to the supervisor, the supervisee, or a mutual collaboration of both partners (p. 462). Ellis (1991) found that trainees rated the supervisory relationship as the most important component of a positive supervisory experience. Majcher and Daniluk's (2009) qualitative and longitudinal study of 6 counseling psychology doctoral students

emphasizes the critical role the relationship between supervisor and trainee plays in development.

Ellis and Ladany (1997) included the following subcategories regarding the supervisory relationship: (Ellis & Ladany, 1997). The subcategories regarding role expectations and structure of the supervisory relationship are represented in the current literature review through investigations about types of supervision; specifically, this refers to group supervision (Riva, Cornish, & Erickson, 1995; Wilbur, Roberts-Wilbur, Hart, & Morris, 1994), peer supervision (Benshoff, 1993), and practicum class supervision (Prieto, 1996). The role expectations, styles of supervision, and structure of supervision varies between these types of supervision. Consequently, matching between supervisor and supervisee(s) becomes more complex and requires further scrutiny.

Second cardinal inference: Inferences entailing matching in supervision. Ellis and Ladany (1997) reviewed multiple studies that investigated inferences regarding the impact of the matching supervisees and supervisors on attributes (such as sex, race, cognitive style and theoretical orientation) on supervision process and outcome. Ellis and Ladany included the following subcategories*: inferences regarding Bernard's Discrimination Model, inferences about individual differences (specific to race, gender, theoretical orientation, environmental setting, reactance, and cognitive style)* and *inferences regarding supervisee needs.*

Ellis and Ladany (1997) examined studies that explicitly tested Bernard's (1979) Model, hence the title of the first subcategory. In the literature reviewed in this study, Bernard's (1979) Model was applied to the concept of supervisor "style." For example, in Bernard's (1979) model, supervisors adopt different roles during the course of

supervision, including teacher, therapist, consultant, and colleague. The supervisor will shift these roles as appropriate and necessary to address the trainee's needs and presentation. The supervisee also juggles various roles, including student, therapist, and trainee (Olk & Friedlander, 1992). In the current literature, the idea of supervisor roles, or 'styles', is applied to understanding the match between supervisee and supervisor. Fernando and Hulse-Killacky (2005) examined the styles of directive teacher, supportive teacher, counselor, and consultant.

Third cardinal inference: Inferences regarding supervisee development. It is commonly supposed that counseling trainees will move through stages of development from prepracticum through internship and professional status (Bear & Kivilighan, 1994; Ellis & Ladany, 1997; Russell et al., 1984). Ellis and Ladany (1997) reported that investigators have tested inferences regarding ego development, conceptual development, several models of supervisee development, and generic supervisee development and experience level. In their investigation, Ellis and Ladany (1997) included subcategories of *inferences regarding ego development, inferences regarding conceptual development,* and *inferences regarding models of supervisee development.*

Fourth cardinal inference: Inferences relating to supervisee evaluation. Ellis and Ladany (1997) contend that supervisee evaluation is invaluable to supervision outcome, and found it "…unfortunate that only 10 investigations attempted to assess aspects of supervisee evaluation" from the years of 1981 to 1993 (p. 483). Even fewer have been published in subsequent years, and only one met criteria for this review. Havercamp (1994) conducted an investigation on the use of self-monitoring for supervisor evaluation of counseling trainees.

Fifth cardinal inference: Inferences about client outcomes in supervision. The process and efficacy of supervision affects client outcome. Issues such as supervisory match, supervisee development, countertransference, and parallel process all have an impact on the client. Researchers have published in the area of countertransference, but articles are largely descriptive/educational (e.g., Shafranske & Falender, 2008) and theoretical (e.g., Tobin & McCurdy, 2006), in nature. Countertransference does seem to lend itself to qualitative investigations, of which several have been conducted (e.g., Ladany, Friedlander & Nelson, 2005; Ladany, Marotta, & Muse-Burke, 2001; Zaslavsky, Nunes, Eizirik, & Nurse, 2005).

Ellis and Ladany (1997) identified one subcategory, that of *Parallel Process.* Hora (1957) cited in McNeill & Worthen, (1989) defined the parallel process as "…an unconscious identification with the client, and that supervisees involuntarily assume their client's tone and behavior to convey to the supervisor emotions experience while working with the client" (p. 329). McNeill and Worthen (1989) include in the definition "…vestiges of the supervisory relationship that may manifest themselves in a reciprocal manner in the therapeutic setting and are not limited to aspects of transference or countertransference" (p. 329).

Sixth cardinal inference: Inferences about supervisees: New measures. Watkins (1998) identifies the need for "valid, reliable supervision measures" (p. 94) as the first of ten key needs in psychotherapy supervision. Ellis and Ladany (1997) identified 7 measures that were specifically developed to assess supervisee variables, two of which they evaluated in their review. The first measure, *The Role Conflict and Role Ambiguity Inventory* (RCRAI) by Olk and Friedlander (1992), is "…a self-report measure that

assesses role difficulties (role conflict and role ambiguity) in supervisory relationships (past and present)" (Ellis & Ladany, 1997, p. 489). The second measure is the short form of the Barrett-Lennard Relationship Inventory (Schacht, Howe, & Berman, 1988).

In 2008, Ellis, D'Iuso, and Ladany published a chapter in *Psychotherapy Supervision: Theory, Research, and Practice* (Eds.) in which the authors reviewed the state of assessment of clinical supervision. The authors replicated the methods and procedures used in Ellis et al. (2006) and Ellis and Ladany (2007) in their investigation of research on clinical supervision measures. In their search, they utilized the following inclusion/exclusion criteria: (a) the main focus of the study was clinical supervision or the supervisory process; (b) the article was empirically based and was published in the literature since their last review (after 1997); (c) the article focused on measures or methods of assessing clinical supervision, supervisors, supervisees, and /or group supervision; (d) the article needed to describe the development of the measure and its psychometric properties, not just theoretical framework; and (e) the article presented further psychometric data about an existing measure for clinical supervision (Ellis et al., p. 478). Their sample included six articles that described and included an assessment scale or measure for supervision that met their inclusion criteria. Other articles about assessment were published during this time frame, but they did not meet the criteria for inclusion (see for example Miller, Korinek, & Ivey, 2006). Their included articles were as follows: Herbert, Ward, and Hemlick (1995), Lehrman-Waterman & Ladany (2000), McHenry and Freeman (1997), Meier (2000), Vespia, Heckman-Stone, and Delworth (2002), and White and Rudoph (2000).

Based on the results of their systematic review of the articles describing the development of a measure(s), the authors categorized each measure as either *recommended* or *not recommended*. Specifically, the psychometric properties on which the criterion for this categorization was based included the following: (a) reliability coefficients exceeding .80, (b) scale discrimination validity, scale scores intercorrelated less than approximately .7 and items not correlating highly on more than one scale/factor, (c) scores demonstrating acceptable properties, (d) scale scores cross-validated in at least one additional sample, (e) samples being reasonably large and representative of the target population and context, (f) presence of evidence provided for convergent and divergent construct validity of scores, (g) appropriate use of confirmatory statistical procedures, and (h) sufficient information and data provided to evaluate the psychometric properties of the measure (Ellis et al., p. 479). Of the six studies reviewed, only one measure, the *Evaluation Process within Supervision Inventory* by Lehrman-Waterman and Ladany (2000), was identified as *recommended*. The remaining five measures evaluated in this study were placed in the *not recommended* category due to excessive flaws in design and methodology. Their conclusions were as follows: (a) The *Supervisory Styles Inventory* and the *Supervision Questionnaire—Revised* (Herbert et al.,1995) were not recommended due to study methodology threats (such as small sample size), inadequate reliability, data not fitting with hypothesized structure, and insufficient discriminant validity, (b) The *Supervisor Emphasis Rating Form—Revised* (McHenry & Freeman, 1997) was not recommended due to methodological threats (such as small sample size), statistical validity threats (violation of assumptions of statistical tests and use of unreliable measures), and lack of cross-validity data, (c) The *Group Supervisory*

Behavior Scale (White & Rudolph, 2000) was not recommended due to external validity threats (no demographic data about the participants, no random selection, homogenous sample) and construct validity threats (low criterion validity coefficients), (d) The *Supervision Utilization Rating Form* (Vespia et al., 2002) was not recommended due to threats to the following: hypothesis validity (research hypotheses not stated), methodological validity (scales derived from importance ratings with no statistical tests), and statistical conclusion validity (inadequate statistical power, type II error rates were uncontrolled, violation of assumptions of statistical tests), (e) *Meier's 11 New Scales of Trainee Development* (Meier, 2000) were not recommended due to hypothesis validity threats (no theoretical basis, rationale, or hypothesis to provide a context for the scale scores), discriminant validity threats (most scale scores demonstrated interdependence), and provision of no psychometric or validity data other than assessment of change score data and internal consistency reliability, (f) The *Assessment Interview Skill Deployment Inventory, the Global Impressions of the Diagnostic Interview—Revised,* and *the Seminar Process Evaluation Form—Revised* (Rudolph et al., 1998) were not recommended due to an overall lack of data. The only data reported on the first two measures, which are rating protocols, was interrater reliability, and no data were reported for the third measure.

Extension with Ellis and Ladany's (1996) Cardinal Inferences

Ellis and Ladany (1997), in their follow-up to Ellis and colleagues (1996), created their six cardinal inferences out of the supervision literature they reviewed. As one would expect, the passage of time has seen evolution in the practice and understanding of counseling supervision. The literature reflects recognition of changes in environment, population, ideology, and technology. In reviewing the supervision literature from 1994

through 2011, three particular areas emerged with enough popularity to justify the addition of three categories, or "inferences," to the six proposed by Ellis and Ladany (1996). These new "cardinal inferences" have been labeled as the following: *inferences about culture, inferences about the use of technology in supervision,* and *inferences about supervision training.*

Inferences about culture. Ethnicity, culture, and multicultural counseling competence are topics that have swiftly gained popularity in the literature over the past fifteen years (e.g., Bhat and Davis, 2007; Dressel, Consoli, Kim, & Atkinson, 2007; Gloria, Hird, & Tao, 2008; Inman, 2006; Miller & Ivey, 2006; Nilsson & Anderson, 2004; Mori, Inman, & Caskie, 2009; Sue et al., 1992; Utsey, Gernat, & Hammar, 2005). The APA Multicultural Guidelines for multicultural education, training, research, practice, and organizational change (2003) define *culture* as the "…belief systems and value orientations that influence customs, norms, practices, and social institutions, including psychological processes (language, care taking practices, media, educational systems) and organizations (media, educational systems)" (p. 380). The Guidelines assert that all individuals are cultural beings, navigate their environments utilizing their own worldviews that include a set of beliefs, values, and traditions (APA, 2003, p. 380). Additionally, the guidelines state that our lifestyles are influenced by the historical, economic, ecological, and political forces on a group (APA, 2003, p. 380).

The increase of immigrant and culturally diverse people into the United States population ensures increase in counselor interactions with clients different from themselves. These changes in client demographics present challenges in counselor practice. Counseling clients representing this diversification requires a broadening of the

counselors' understanding of clients' cultural and social contexts and of knowledge regarding effective therapeutic interventions. In many ways, theoretical shifts must occur in models of human development and psychological well-being in order to incorporate what may be new beliefs, principles, ideologies, and perspectives that imminent from the diverse clients that require our services. In 1985, Katz stated the need for the profession to recognize that counseling is neither value-free nor disconnected from social, political, and historical realities and the need to identify effective methods of training and assessing cultural competence for those deliver psychological services. Sue and Sue (1999) point out that mental health professionals prefer to view themselves as "moral, just, fair-minded and decent," making it difficult them to recognize any potential harm that the cultural encapsulation to which Katz (1985) alludes, may create for clients. Currently, the construct of multicultural competence is most influenced by the triad model of awareness, knowledge, and skills (Sue et al., 1992). Specifically, multicultural counseling competence has been defined as counselors' awareness (attitudes and beliefs), knowledge, and skills in working with individuals from a variety of cultural groups (Sue et al., 1992).

The current literature review reflects numerous investigations regarding individual differences specific to race and racial identity (e.g., Bhat & Davis, 2007; Constantine et al., 2005; Inman, 1996; Gatmon, Jackson, Koshkarian, Martos-Perry, Molina, Patel, & Rodolfa, 2001). Racial identity is most typically defined in terms of the Racial Identity Models authored by Helms (1990). Helms developed a White Racial Identity Model to delineate stages of a White person's understanding of herself or himself as a White person as related to people of color. The model describes White Racial

Identity in six stages, from the lowest stage designated as *Contact* (characterized by ignorance or obliviousness to the sociopolitical implications of race as it is defined in this country) through the highest stage designated as *Autonomy* (characterized by internalization of nonracist White perspective wherein benefits of racism are rejected). Helms also created an African American Racial Identity Model consisting of four stages, from the lowest stage designated as *Pre-Encounter* (characterized by Euro-American frame of reference wherein persons act or think in ways that devalue African-Americans) through the highest stage designated as *Internalization* (characterized by a sense of inner security with one's own culture/race/ethnicity; Helms, 1990). Helms applied these Racial Identity Models to the therapy process and described three distinct types of interactions, or dyadic types, that can occur between two people with regard to these models (Helms, 1994). The first is a *regressive dyad*, an interaction in which the client's stage of racial identity is higher than that of the counselor. The second dyadic type is the *progressive dyad*, an interaction in which the counselor's stage of racial identity is higher than that of the client. The third and last dyadic type is the *parallel dyad*, an interaction in which the client and the counselor share similar racial attitudes.

Supervision researchers have utilized Helms' modes and applied them to the supervision relationship and process. Research has addressed the impact of White racial identity attitudes of counselor trainees and dyadic interactions (e.g., Constantine, Warren, & Miville, 2005; Utsey & Gernat, 2002). Ladany, Brittan-Powell, and Pannu (1997b) applied Helms' Racial Identity Models to the relationship and racial interaction between supervisee and supervisor from the perspective of supervisees. They utilized Helms' three dyadic types but divided the parallel dyad into parallel high (Supervisor and Trainee in

Phase II) and Parallel low (Both in Phase I). The strongest supervisory working alliance resulted from parallel high supervisory relationships, with progressive relationships having the second strongest reports of working alliance. Participants in a parallel low relationship demonstrated weaker bonds, with the regressive relationship reported as the weakest (Inman & Ladany, 2008; Ladany & Inman, 2008; Ladany et al., 1997).

Researchers have explored many different cultural issues specific to supervisors, including multicultural supervisory behaviors (Dressel, Consoli, Kim, & Atkinson, 2007), effects of supervisor's race (Hilton, Russell, & Salmi, 1995), self-reported multicultural supervision competence (Gloria, Hird, & Tao, 2008), multicultural framework for counselor supervision (Ladany, Inman, Constantine, Hofheinz, 1997) and supervisor cultural responsiveness /unresponsiveness in cross-cultural supervision (Burkhard, Johnson, Madon, Pruitt, Contreras-Tradych, & Kozlowski, 2006). Researchers have also investigated the trainee experience in regards to culture, including white counselor trainee reactions to racial issues (Utsey, Gernat, & Hammar, 2005). Research regarding the supervisory relationship and process as it regards culture include studies about cross-racial supervision (Schroeder, Andrews, & Hindes, 2009), spirituality and gender (Miller & Ivey, 2006), ethnicity, gender, and sexual orientation (Gatmon, Jackson, Koshkarian, Martos-Perry, Molina, Patel, & Rodolfa, 2001), and supervision incidents (Toporek, Ortega-Villalobos, & Pope-Davis, 2004). Additionally, minority supervisee experience has gained increased attention in the literature (Bhat & Davis, 2007; Mori et al., 2009; Nilsson & Anderson, 2004; Nilsson & Duan, 2007).

A particularly important issue is that of multicultural counseling competence and how best to assess and train therapists to become multiculturally competent counselors

(e.g., Constantine, Warren, & Miville, 2005; Gainor & Constantine, 2002; Gloria et al., 2008; Inman, 2006; Ladany et al.,1997a; Ladany et al., 1997b). Several measures of multicultural competence have been created to assess trainee competence, including the Multicultural Awareness-Knowledge-Skills Survey--Counselor Edition (MAKSS-CE; D'Andrea, Daniels, & Heck, 1991), Multicultural Counseling Knowledge and Awareness Scale (MCAS; Ponterotto, Gretchen, Utsey, Rieger, & Austin, 2002), Multicultural Counseling Inventory (MCI; Sodowsky, Taffe, Gutkin, & Wise, 1994), and the Cross-Cultural Counseling Inventory—Revised (CCC-I; LaFromboise, Coleman, & Hernandez, 1991).

Inferences about the use of technology in supervision. This is certainly the age of reliance on technology for communication: we utilize text and email as much as we used to rely on talking on the phone. We take it for granted that others are connected to social networking or that they are adept at using instant messaging and videoconferencing. Meeting with people in person is almost obsolete, as we can substitute it with any in an array of technological options. The process of supervision is not unaffected by these developments. Vaccaro and Lambie (2007) observe that, despite increase in popularity and use, supervision via email has received almost no formal investigation. Clingerman and Bernard (2004) investigated the use of e-mail as a supplemental modality for clinical supervision, studying the patterns of e-mails between practicum students and their supervisors. A significant decline in number of emails occurred as the practicum progressed, which has implications for the use of e-mail in supervision. Butler and Constantine (2006) examined the efficacy of a 12-week web-based peer supervision program for school counselor trainees. The authors found that participants in the program

had significantly higher collective self-esteem and case conceptualization skills than those who did not participate. Gainor and Constantine (2001) compared in-person to web-based supervision in their investigation of supervision satisfaction and multicultural case conceptualization. Results showed that in-person multicultural supervision was more effective in developing abilities in multicultural case conceptualization.

Inferences about supervision training. In their review of supervision literature, Ellis and Ladany (1997) did not discuss literature that directly addressed supervisor training—methods for training, the impact of training, etc. This is largely because the literature of that time was primarily descriptive and narrative in nature. Competency did not have as big a footprint in the literature as it does now. Researchers have become increasingly interested in counselor competency and the impact of training on counselor competency (Ladany, Friedlander, & Nelson, 2005; Milne & James, 2002; Milne, 2010). Training and supervision practices have been investigated in supervision of clinical, counseling, and school psychology (Keller, Protinsky, Lichtman, & Allen, 1996; Page, Pietrzak, & Sutton, 2001; Romans, Boswell, Carlozzi & Ferguson, 1995; Ward, 2001), Marriage and Family therapy (Anderson, Schlossberg, & Rigazio-DiGilio, 2000; school counseling (Kahn, 1999; McMahon & Patton, 2001), and rehabilitation counseling (Schultz, Ososkie, Fried, Nelson, & Bardos, 2002). Specific supervision training programs have also received attention (Sundin, Ögren, & Boëthius, 2008).

Ethical training is intrinsic to supervision training and has gained attention in the supervision literature (Falender & Shafranske, 2004; Sherry, 1997). Ladany, Lehrman-Waterman, Molinaro, and Wolgast (1999), in their investigation of psychotherapy supervisor ethical practices, investigated the adherence of psychotherapy supervisors to

ethical guidelines, finding a correlation with the supervisory working alliance and supervisee satisfaction.

Replicating and Extending Ellis et al. (1996*)*

As the practice of quality supervision is informed and guided by research conclusions, these conclusions and inferences must be based on rigorous and relevant methodology that is appropriate to the complexity of the process under investigation. The purpose of the current study is to replicate and extend the work of Ellis et al. (1996) and Ellis and Ladany (1997) in order to evaluate the supervision literature from 1994 through 2009. It is important to ascertain the degree to which authors incorporated the findings and recommendations from prior studies and reviews in subsequent studies. Ellis et al.(1996) and Ellis and Ladany (1997) initially examined the degree to which researchers applied the recommendations from Russell et al. (1984) and found that authors did not follow the recommendations but instead continued making mistakes in research design and interpretation. Ellis and colleagues contended that this is problematic because it may lead to "…obfuscation of excellent research by poor research, exacerbation of theoretical ambiguity in the field, and creation of inaccurate inferences and conclusions" (Ellis et al., 1996, p. 44). The authors' conclusions and recommendations about the state of supervision research, based on thorough and meticulous examination, is a call to the profession to make changes in supervision research. Ellis et al. (1996) published their investigation 12 years after Russell et al. (1984) published their recommendations for research. As no such meta-analytic studies have been published since, it seems particularly timely and important to conduct a replication of Ellis et al. (1996), to assess whether researchers have applied the recommendations put forth by those authors.

CHAPTER III

Method

Study Search Procedures

A search of the literature was conducted using ERIC, APA, and PsychINFO databases as well as historical literature reviews (Inman & Ladany, 2008). The search was limited to articles published in North American Journals (both the United States and Canada). Aside from Canadian journals, international journals were not included. The journals were searched using "supervision" as the only keyword; the search was not limited further so that relevant articles were not missed. In addition, only articles published in English were considered. The researcher reviewed abstracts in 68 peer-reviewed journals encompassing the years 1994 through 2010 (inclusive). The 68 journals were selected due to their association with psychology, psychiatry, therapy, counseling, rehabilitation, and/or education, with the expectation that research on supervision of psychotherapy may be included in any of these areas. Of the 68 journals reviewed, 765 articles about supervision from 28 journals were identified. Most of the journals included in this study were not in publication during the time of the original study. As the intent of this review is to examine the state of published research, unpublished research such as dissertations, conference papers, technical reports, or rejected manuscripts were not included.

Inclusion-Exclusion Criteria

Utilizing the inclusion criteria described by Ellis et al. (1996), the primary researcher reviewed the abstracts from the 765 articles and reduced the total number of articles to 108.

The inclusion criteria as identified by Ellis et al. (1996) were as follows: *First criterion*: The article must meet the definition of clinical supervision according to Bernard and Goodyear (1992), who defined supervision as an intensive interpersonally focused relationship in which one or more persons are designated to facilitate the development of therapeutic competence in the other person or persons. *Second criterion*: The article must be data-based and published in a refereed professional journal during the specified time period under study. *Third criterion*: Individual counseling/psychotherapy must be addressed as an integral part of the study. *Fourth criterion*: The types of supervision to be included in the study include individual supervision of individual, marriage, couples, and family therapy; group supervision of individual therapy; empirically based case studies; or postgraduate supervision (Ellis et al., 1996). Fields of therapy included counseling psychology, clinical psychology, school psychology, psychiatry, psychiatric nursing, counseling (e.g., school, community mental health, and rehabilitation), and social work (Ellis et al., 1996). *Fifth criterion*: Types of supervision to be excluded from the study included supervision involving pre-practicum or microskills training, supervision of group therapy, speech pathology, teacher supervision, anecdotal case studies, and unpublished manuscripts (Ellis et al., 1996). The sample excluded research on supervision assessment, measurement, and/or evaluation due to the review conducted by Ellis, D'Iuso, and Ladany (2008). Additionally, the researcher chose not to include supervision in psychiatric nursing because the tasks and training of psychiatric nurses vary significantly from counseling and clinical psychotherapy tasks and training. Comparison, then, was not deemed useful.

When the 108 articles had been collected, the primary researcher and coders reviewed each article in its entirety to ensure that each met the criteria set forth by Ellis et al. (1996), and articles that did not meet criteria were removed by consensus. First, a thorough examination of the articles revealed that 13 articles did not meet one or more of the inclusion criteria of Ellis et al. (1996) and/or the researcher's established criteria. Barnett-Queen and Larrabee (2000) included in their sample participants that were not necessarily in supervisor/supervisee roles. Hilton, Russell, and Salmi (1999) included participants who were undergraduate "counselors" with no counseling experience, and Knight (2001) included bachelor's level counselors. Peleg-Oren, Macgowan, and Even-Zahav (2007) included supervisors who did not meet the criteria/definition of supervisors and also included bachelor-level participants. Schoenwald, Sheidow, and Chapman (2009) used a sample combining caregivers and paraprofessionals, and so the supervision did not meet criteria for typical psychotherapy supervision. White and Russell (1995) and Dressel, Consoli, Kim, and Atkinson (2007) presented Delphi poll studies that had no analyzed data. Raimrez (2003) included paraprofessionals and service staff with supervisors, and therefore does not meet the criteria/definition of supervisor. Nyman, Nafziger, and Smith (2010) discussed supervision outcomes as a secondary purpose of the study, and also gathered no actual data about supervision. Additionally, four articles were excluded from this study because they used international samples: Gabbay, Kiemle, and Maguire (1999) used a sample from England; McMahon & Simons (2004) used a sample from Australia; Milne (2010) used a sample from England; and Schectman and Wirzberger (1999) used a sample from Israel.

Second, articles that were considered largely or primarily qualitative in nature were removed. The original intention of this investigation was to include those studies, but it became apparent that this was not a tenable undertaking. Qualitative research employs a very different methodology from quantitative research and has its own approach and language for addressing rigor and quality. Ellis et al. (1996) included some qualitative research in their review, but they unfortunately did not make changes in their methodology to appropriately address the data. Qualitative research adheres to a different set of standards than quantitative research (Cresswell, 1998). The goals, expectations, and reasons for conducting qualitative research are unlike those of a quantitative study (Cresswell, 1998). As such, the data and language used are difficult to compare to quantitative research. Researchers (presumably) put significant time and consideration into the design of their studies, and to evaluate only one portion of the design is not fair to the researchers—especially if the quantitative and power aspect of the study is lacking. While it may be of interest in the current study, it would put those studies in an unfair light and possibly lead readers of this study to assume that those research articles were problematic in their entirety.

Additionally, the discussion of research findings in the combined quantitative/qualitative designs comingle the data and draw inferences from a combination of the two sets of data. Consequently, it is very difficult to evaluate the legitimacy of the conclusions without thoroughly examining both types of data. For example, even if the quantitative methodology is perfect, conclusions still cannot be verified without inspection of the qualitative data. With the decision to exclude these studies, 16 studies were excluded (Burkard, A.W., Johnson, A.J., Madson, M.B., Pruitt,

N.T., Contreras-Tradych, D.A., Kozlowski, J. M., et al., 2006; Carter, Enyedy, Goodyear, Acinue, & Puri, 2009; Chui, 2010; deMayo, 2000; Dennin & Ellis, 2003; Enyedy, Arciune, Puri, Carter, Goodyear, & Getzelman, 2003; Fortune & Abramson, 1993; Gainor & Constantine, 2002; Havercamp, 1994; Ladany, Marotta, & Muse-Burke, 2001; Nelson & Friedlander, 2001; Peace & Sprinhall, 1998; Sells, Goodyear, Lichtenberg & Polkinhorne, 1997; Utsey, Gernat, & Hammar, 2005; White & Russell, 1995; Yourman & Farber, 1996).

Third, articles focusing on the development of a supervision measure or assessment were removed. Upon review of Ellis et al.'s (2008) review of supervision measurements, in which the authors replicated Ellis et al. (1996), it was decided that inclusion of measurement articles would be redundant. With this decision, 14 studies were excluded (Henggeler, Schoenwald, Liao, Letourneau & Edwards, 2002; Herbert, Ward, & Hemlick, 1995; Lehrman-Waterman & Ladany, 2001; Lochner & Melchert, 1997; Long, Lawless, & Dotson, 1996; Lovell, 1999; Miller, Korinek & Ivey, 2004; Miller et al., 2006; Nilsson & Dodds, 2006; Sells, Goodyear, Lichtenberg, & Polkinghorne, 1997; Sumerel & Borders, 1996; Thielsen & Leahy, 2001; Vespia, Heckman-Stone, & Delworth, 2002). Additionally, one supervision assessment article published after Ellis et al. (2008) was considered but then excluded because it contained a sample from Germany (Zabock, Drews, Bodansky, & Dahme, 2009).

After the researcher and coders completed the review, a total of 62 articles remained for inclusion in the current study.

Statistical Variables

The quantitative analyses utilized in the current study followed Ellis et al.'s (1996) methodology using the literature identified in the search of research from 1994 through 2010. The analysis was extended in this study to include the types of power analyses recommended by Cohen (1962). This additional analysis was not performed in Ellis et al. (1996), which was a limitation of the study. The reason for this additional analysis was to facilitate further comparison to other power analyses (e.g., Cohen, 1962; Haase & Solomon, 1982; Rossi, 1990). As in Ellis et al. (1996) and Ellis and Ladany (1997), statistical variables were aggregated across statistical tests and across studies to survey the general quality of the statistical methods employed in the study articles by assessing the prevalence of type II errors. This allowed the researcher to compare the articles evaluated in this study to previous empirical reviews (e.g., Ellis et al., 1996; Ellis & Ladany, 1997; Haase, Ellis, & Ladany, 1989; Rossi, 1990).

For each statistical test, in each study article, the researcher calculated the following statistical variables: sample size (N); sample effect size (η^2, partial eta squared); the minimum effect size one would expect to obtain given the sample size and an a-priori power of 80% ($\eta^2_{min(N)}$); the post hoc statistical power ($P_{(\eta 2)}$) as per Ellis et. al.'s method, the post hoc statistical power for small, medium and large effects as per Cohen's methods ($P_{(Small)}$, $P_{(Med)}$, $P_{(Large)}$); the per-comparison Type II error rate (α_{PC}) and the experiment-wise Type I (α_{EW}) and Type II ($\beta_{EW(\eta 2)}$) error rates.

Descriptive Discussion

Effect size measures the degree to which a phenomenon is present in a population (Cohen, 1988). Effect sizes typically fall into one of two categories (Ellis, P., 2010). The

first category measures the differences between groups, the "d family" (e.g., odds ratios, relative risks, Glass's delta), and the second measures the strength of a relationship, the "r family" (e.g., eta-squared, Pearson correlation coefficients, R^2s, beta coefficients, Cramer's Vs, and omega squares). Effect sizes can be reported in a variety of forms, and the articles reviewed presented the results in different ways (when they present them at all). Before the effect sizes were compared, the researcher converted them into a common metric. As per Ellis and colleagues (1996), the values were converted to eta-squared. Eta-squared measures the strength of association or magnitude of the effect, and "…embodies the notion of the proportion of dependent variable variance accounted for by categorical, independent effects" (Haase et al., 1989, p. 408). The formula for eta-squared depends on statistical variables that are not always reported in statistical studies (eta-squared is a function of the between sum-of-squares and the total sum-of-squares). Cohen (1965) derives an equivalent formula for eta-squared that is a function of the F statistic and the appropriate degrees-of-freedom for the effect, but this is only true for the one-way ANOVA design (Kennedy, 1970) and yields the positively biased "partial" eta-squared otherwise. Ellis and colleagues (2010) note that while partial eta-squared is positively biased, it should not have a serious impact on the other measures computed with eta-squared (Pierce et al., 1994). Following Pierce et al.'s suggestion, the researcher independently computed partial-eta-squared values based on other reported statistics even if effect sizes were already reported. The researcher also recorded the effect sizes reported in the study for comparison purposes.

Generally speaking, the reviewed articles reported the outcome of each statistical test encountered in the form of traditional test statistics (e.g., t-values, F-values, etc.),

degrees of freedom, and sample size. Using this information, the researcher was able to calculate the effect size used by G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) to compute the power for that test. G*Power uses Cohen's effect measures (Cohen, 1988) that that were defined specifically for the particular test being performed. So, for example, t-tests which test the significance of the mean were measured with the "d" and "dz" effect measures, F-tests were measured with "f" and "f^2," correlations were measured with the "r" effect measure, and finally χ^2 tests were measured using the "w" effect size. The researcher was then able to calculate η^2 based on the test-specific effect measures.

Calculation of Effect Size

Most formulae can be found in (Cohen, 1988), the G*Power user-manuals (Haase & Ellis, 1987; Haase, 1991). Both Haase & Ellis (1987) and Haase (1991) were instrumental in understanding how to calculate effect sizes in a multivariate setting and how to interpret the degrees of freedom of multivariate tests. For the formulas used in the statistical analysis, see Appendix B.

Statistical Power

Statistical power is the probability that the test will reject a false null hypothesis, or in other words, that it will not make a Type II error (incorrectly accepting the null hypothesis; i.e., a false negative). As power increases, the chances of committing a Type II error decrease. Power is equal to one minus the Type II error probability. Power can be used to calculate the minimum sample size required to accept the outcome of a statistical test but can also be used to calculate the minimum effect size that is likely to be

detected in a study using a given sample size. Power was calculated by means of G*Power (Faul et al., 2007) and Cohen's (1988) formulas.

Calculation of Power

G*Power is a power calculator that allows a user to specify the type of test they are going to perform followed by the type of power calculation they would like to perform; then, it presents the user with a graphical screen to type in the test parameters that are pertinent to the power calculation. These test parameters typically include the effect size, the sample size and degrees of freedom.

In order to facilitate greater transparency and ease of replication, however, the researcher used G*Power in its "generic" mode, which removes the forced choice program options and allowed the researcher to pre-compute the generic parameters used to calculate power and then enter those values into G*Power. These parameters include: the non-centrality parameter (λ), the type I error rate (α), and the relevant degrees of freedom (df1, df2).

Conceptually, the power calculation is computing one minus the cumulative probability derived from the distribution function of a test statistic if H1 were true and its critical value was the value derived from the cumulative probability of the distribution function of the test statistic if H0 were true given α. In order to determine the distribution function of the test statistic if H1 were true (rather than H0), one must shift the distribution of the traditional test-statistic to the right along the x-axis re-centering it on what its central value would be if H1 were true. The non-central Student's-t, F and χ^2 distribution functions are used for this task, and they are all parameterized by a non-centrality parameter along with the traditional distribution function parameters: value of

the random variable, Type I error rate, and degrees of freedom. The non-centrality parameter is a function of the sample size (N), effect measure (ME) and other test-specific parameters. As either the sample size or effect size increase, the value of λ increases. As λ increases, the non-central distribution is shifted further to the right and the power increases. Student's-t distribution, the F distribution and the χ^2 distribution all have associated with them non-central distributions. G*Power is able to compute the power for all tests that use the non-central Student's-t distribution, the F distribution and the χ^2 distributions. Note that the researcher did not use G*Power to compute the power of the Tukey's Range Test because it uses a Studentized-Range distribution which G *Power does not include. Therefore it was calculated separately.

The exact procedure to compute the minimum detectable effect size given the sample size and assuming 80% power ($\eta^2_{min\ (N)}$) is embedded within G*Power and is not known to the researcher. The output of G*Power's procedure was a value for λ. After G*Power found a value for λ, the researcher was able to solve for ME by inverting the equation she had to compute λ for the post-hoc tests.

Per Comparison and Experiment-wise Error Rates

Type II error is the probability of accepting a false null hypothesis. As per Ellis et al. (1996), the per-comparison Type II error rate was calculated as $\beta_{PC(\eta 2)} = (1 - P_{(\eta 2)})$ and the experiment-wise Type II error rate (also called "family-wise" rate) was computed as $\beta_{EW(\eta 2)} = 1 - (1 - P_{(\eta 2)})^s$, where s is the number of statistical tests in the experiment. The experiment-wise Type I error rate was computed in the same fashion where the per-comparison alpha value was set to 0.05 unless specifically defined otherwise by the study authors.

Detailed Statistical Coding Procedure

Two raters extracted the pertinent statistical data from each study. The primary researcher recorded the data from all 62 studies, and then a second rater repeated the procedure to ensure accuracy. The researcher used a spreadsheet (Microsoft Excel 2010) to simplify the calculation process. Each column contains the parameters for a single test. Different studies were broken out by different "tabs" in the spreadsheet. The tests from each study were grouped by family ID and test ID. The family ID was used to group tests for the experiment-wise error rate calculations. The test ID was used to uniquely identify a test. Follow-up tests that controlled for the Type I error rate in some manner were given the same test ID. The value of the test reason could either be "<main>" or "<peripheral>." This allowed the researcher to differentiate between the power of exploratory and other tests, and the power of tests used to directly support the main hypotheses (Cohen 1962). The next question regarding whether the author controlled for family-wise error rates was a study-wide summary statistic (i.e., only one occurrence of "<yes>" was required) to count the number of studies where the author explicitly reported having controlled for family-wise error rates.

The following items varied depending on the test type: Formulae for converting test parameters and statistics into η^2, the effect measure used by G*Power for the particular test, and the non-centrality parameter calculation. Additional parameters required by *G*Power* to perform the power calculations were conducted. The spreadsheet was designed in such a way that when the appropriate test was selected, the appropriate formulae were automatically loaded into the correct cells in the spreadsheet. Notable fields include the following:

1) the name of the test statistic (i.e., "F", "*t*", etc.)
2) the value of the test statistic
3) df1 and df2, which defined the degrees of freedom (df2 was non-zero for various F tests)
4) *n*1 and *n*2 defined the sample sizes (n2 was non-zero for comparison tests)
5) Reported effect value: Some authors reported effect sizes that were used by the researcher to cross-check her own calculations.
6) Number of dependent variables: The non-centrality parameter calculation required this information from multivariate tests (e.g., MANOVA and MMR)
7) rho: the effect size calculations for pair-wise tests (e.g., pair-wise t tests) require an estimation of the correlation between the two variables.
8) m: Various calculations for repeated-measures tests need to know the number of times the measure was repeated.
9) s: This is a parameter used in degrees-of-freedom, effect size and non-centrality parameter calculations of multivariate tests (e.g., MANOVA). It is commonly denoted as s and can be found for instance in (Haase, 1991) and (Haase and Ellis, 1987).
10) Tail: *t*-tests and correlation tests require the number of tails (1 or 2) in order to adjust alpha.

Methodological Evaluation Variables

The methodology of each study was evaluated in the same manner as Ellis et al. (1996), using Cook and Campbell's (1979) 33 threats to validity, the four threats to hypothesis validity submitted by Wampold et al. (1990), and Russell et al.'s (1984) 12

methodological threats. A sample coding sheet can be found in Appendix D. In order to ensure accurate representation of each study, the researcher used only the data and calculations that were explicitly reported in each study. Any data not provided in the article remained unknown and uncalculated. As one intention of this study was to ascertain the state of published research, it would have been contradictory to attempt to gain the missing data either through calculations or contacting the authors. Expecting that most research consumers assume that all salient information is included in the published work, it is logical that the research articles evaluated in this study should stand on their own for evaluation.

Each study was assigned to one of Ellis and Ladany's (1997) cardinal inferences or one of the three new inferences hypothesized in this study. Ellis and Ladany's (1997) sixth cardinal inference, *Inferences about supervisees: New Measures,* was omitted. This category was revisited by Ellis, D'Iuso, and Ladany in 2008. In their chapter, *State of the Art in the Assessment, Measurement, and Evaluation of Clinical Supervision*, the authors review measurement articles published between 1997 and 2007 utilizing their previous methodology and criteria. The review is thorough and therefore not replicated here. Only one other measurement article (Zarbock, Drews, Bodansky, & Dahme, 2009) was published after that time and within the time frame of this study, but it did not meet criteria for inclusion. Therefore, given the publication of Ellis et al.'s 2008 review and with the lack of any other published article meeting criteria in this category, this cardinal inference category was omitted for this review. (Readers are encouraged to read this review for further information.)

In assigning the articles to inference categories, the raters focused on the reported purpose of the research studies and the associated statistical analyses. Some articles discussed or referred to more than one inference, but a primary inference always stood out. Specifically, the primary inference of a study was associated with the statistical analyses performed. If another inference was present, it was always clearly a secondary focus in the paper and often with little data or analysis. Therefore, careful examination of each article resulted in relatively clear identification of single inferences for each study. Inference categories were clearly and consistently identified by both raters.

Statistical Conclusion Validity

According to Cook and Campbell (1979), statistical conclusion validity refers to the validity of conclusions about the covariation of independent and dependent variables. Cook and Campbell (1979) identified seven threats to statistical conclusion validity, which are as follows: low statistical power, violation of assumptions of statistical tests, Type I error, unreliability of measures, unreliable treatment implementation, random irrelevancies in the experimental setting, and/or random heterogeneity of respondents.

Internal Validity

Internal validity pertains to questions of relations and causality. Cook and Campbell (1979) identified the following 13 threats to internal validity: history, maturation, testing, instrumentation, statistical regression, selection, mortality, interactions with selection, ambiguity about the direction of causal influence, diffusion of treatments, compensatory equalization of treatments, compensatory rivalry by respondents receiving less desirable treatments, and/or resentful demoralization of respondents receiving less desirable treatment.

Construct Validity

Cook and Campbell (1979) asserted that construct validity is concerned with "…generalizations about higher-order constructs from research operations" (p. 38). They identified 10 threats to construct validity, which include the following: inadequate preoperational explication of constructs, mono-operation bias, monomethod bias, hypothesis guessing within experimental conditions, evaluation apprehension, experimenter expectancies, confounding of constructs and levels of constructs, interaction of different treatments, interaction of testing and treatment, and restricted generalizability across constructs.

External Validity

External validity "…refer[s] to the approximate validity with which conclusions are drawn about the generalizability of a[n observed] causal relationship to and across populations of persons, settings, and times" (Cook & Campbell, 1979, p. 39). Cook and Campbell identified the following three threats to external validity: interaction of selection and treatment, interaction of setting and treatment, and interaction of history and treatment.

Hypothesis Validity

In accordance with the procedure of Ellis et al. (1996), this study will utilize Wampold et al.'s (1990) four threats to the hypothesis validity of a study. The four include: inconsequential hypotheses (e.g., the extent to which hypotheses both corroborate one theory and falsify many others), ambiguous hypotheses (e.g., no clear specification of hypotheses or of the conditions under which hypotheses will fail or succeed), non-congruence of research and statistical hypotheses (i.e., incorrect statistical

procedures or statistics not adequately testing the research hypotheses), and diffuse statistical hypotheses and tests (i.e., any combination of the following three: multiple statistical tests per hypothesis, omnibus tests and subsequent follow-up or post hoc tests, or statistical analyses including extraneous independent variables not specified in the hypotheses (Ellis et al., 1996).

Russell et al.'s (1984) Methodological Threats

Consistent with Ellis et al. (1996), each study was evaluated according to Russell et al.'s (1984) 12 methodological threats for supervision research. The threats are divided into six internal and six external validity categories. Russell et al.'s (1984) six threats to internal validity are as follows: lack of adequate comparison group, no pretreatment assessment, inadequate sample size, variations or confounds in length of training across conditions, failure to randomly assign participants to conditions (non-randomization), and widely discrepant cell sizes (suggesting that the homogeneity of variance assumption may have been violated). Russell et al.'s six external validity threats include the following: restricted range of dependent variables, non-representative supervisee or supervisor sample, lack of follow-up assessment, use of role play or audiotaped client statements to assess supervised change, exclusive reliance on self-report data, and overly brief training period.

Coding Procedures

Interrater Reliability and Kappa

Interrater reliability was achieved through consensus estimates of interrater reliability, which is when two judges come to exact agreement on how to use a rating scale to score behaviors (Stemler, 2004). While other methods of achieving interrater

reliability exist, it was determined that the consensus method was most appropriate for this study. Consensus estimates are most useful when data are nominal in nature and different levels of the rating scale represent qualitatively different ideas (Stemler, 2004).

In the current study, two raters coded 20 articles and then the primary researcher computed Kappa. The raters reached 90% reliability on all of the 49 validity threats, which was considered to be sufficient. If raters reach a point where they agree on how to interpret a rating scale, then scores given by the two raters may be treated as equivalent, and then the remaining set of data to be coded can be split between the raters (Stemler, 2004). In the present study, the results of the initial coding set demonstrated that the two raters agreed on how to interpret the coding criteria to an adequate degree (above 80% for all categories). It was therefore determined that interrater reliability was adequate, so the raters divided the remaining 42 articles into two groups of 21. Articles outside of the training set were randomized by assigning 42 uniformly distributed random variables to the remaining 42 articles and sorting the articles based on the value of the random variable. Each rater coded one set of 21 articles, and then audited the codes of the 21 articles rated by the other coder.

Raters

Following the methodology of Ellis et al. (1996), each study in both the replication and the extension was rated in random order by two judges. Any resulting differences in ratings were resolved by consensus. To prevent rater drift and ensure interrater reliability over 80%, interrater agreement between the two raters was periodically checked. While 90% was used by Ellis and colleagues, the literature supports

the use of 80% as the standard and was used in the current study (Tinsley & Tinsley, 1987).

Three individuals served as raters in the current study. Following is a description of each rater. The first two raters conducted coding of methodological validity threats. The third rater and the first rater conducted computation and coding of the statistical variables. The primary researcher/rater is a White, married 35 year-old female counseling psychology Ph.D. student in counseling psychology at a university in Pennsylvania. She is an identical twin with no other siblings and is from a middle class background and was born and raised just outside of Philadelphia, PA, in a relatively urban setting. She was raised Catholic but no longer practices, though she considers herself to be spiritual. Her mother is a first-generation German American, whose mother was from Hungary and father was from Germany. Her father's parents were from Philadelphia; three generations back, his family immigrated to Philadelphia from Wales. Her research background includes attitudes towards gay men, multicultural counseling competency, supervision, and parallel process. She has utilized both qualitative and quantitative methodology. She has worked as a master's level therapist in three residential treatment centers, two state hospitals, three college counseling centers, two partial hospitals, and three outpatient clinics. She has worked extensively with low SES minority clients, adolescents, and trauma.

The second rater is a White, married, 28 year-old female counseling psychology Psy.D. student at a small private university in Pennsylvania. She is from a lower middle class background in rural Western Pennsylvania. She is the middle of three daughters and is the only one in her family to attend college. She was raised in a protestant church,

though her family did not attend often and she has not continued to follow the religion. She does consider herself to be spiritual. Her parents' families have lived in and around the same area in Western Pennsylvania for multiple generations, and it is commonly held that ancestors came from Germany. However, there is no further data available about this. Her research background includes investigations about counselor experiences and safety training for counselors, and she has primarily used quantitative analysis. She has worked at two different outpatient centers, one college counseling center, and a mobile therapy setting. She has worked extensively with low SES clients and families.

The third rater only coded the statistical variables only. This coder is a White, married, 35 year old male senior quantitative finance analyst. He holds a bachelor's degree in electrical engineering and a bachelor's degree in computer engineering. He also holds a master's degree in financial mathematics. He is from a middle class background in rural West Virginia. He has one younger sister and three younger stepbrothers. He was raised Baptist, and his grandfather and uncle are both Baptist ministers. He no longer follows the religion and considers himself to be agnostic. His parents' families have lived in the same area for generations but believe their ancestors came from England. He conducts quantitative research and statistical analysis in his work.

Raters were trained in rating procedures by the primary researcher until a minimum of 80% was reached on all rating variables. Training included instruction regarding Cook and Campbell's (1979) 33 threats to validity, Russell et al.'s (1984) 12 methodological threats for supervision research, and Wampold et al.'s (1990) four threats to the hypothesis validity of a study. Inter-rater reliability was established through a process of reviewers coding the same subset of studies and then comparing coding

assignments. Ellis (2010) defines inter-rater agreement as the number of agreements divided by the sum of agreements plus disagreements. Coding trial runs were repeated until inter-rater agreement reached scores close to one, indicating that coding definitions were sufficiently clear. Disagreements were resolved by discussion.

Those conducting any study will influence the outcome of the study itself. Rosnow and Rosenthal (1996) acknowledge that a typical scenario for a replication/extension team is that it consists of colleagues and/or a group of faculty along with their pre- and/or post-doctoral students. Such groupings will result in a higher likelihood of "intercorrelations" between members, which in turn is likely to result in an overall bias in coding. In the current study, these biases cannot be fully controlled simply because two of the researchers are in counseling psychology doctoral programs, and so their training alone brings with it ideology and theory about supervision and research. However, the researchers do not attend the same schools and are in different types of programs (Ph.D. and Psy.D.), and therefore the curriculum and ideology is somewhat different. Backgrounds of each coder have been recorded in detail (see above) so that possible bias can be identified. The third researcher was chosen so that evaluation of the studies would be informed by authority of sound statistical modeling and practice. Additionally, he has no specific training in psychology or psychological theory and therefore will be less likely to lean toward a particular perspective when performing selections of extension articles. A recurring critique by Ellis and Ladany (1996) was that researchers employed improper statistical analysis and misinterpreted their findings. It was essential that a member of the research team be skilled in statistical analysis in order to offset chance of missing statistical flaws in the research. A graduate student is unlikely

to have the expertise that a professional statistician would hold. A limitation of inclusion of this team member centers on limited experience reading and critiquing psychological studies. Therefore, in addition to training in coding procedures, he also received education about research methodology in psychology.

Code Assignments

The results of each study (effect types and sizes) and the study characteristics that affect the accuracy of the results (e.g., the sample size and the reliability of key measures) were coded. Raters coded the methodological evaluation variables (i.e., Cook & Campbell's (1979) threats to validity, Russell et al.'s (1994) methodological threats, and Wampold et al.'s (1990) hypothesis threats) by entering a "1" if the threat was present and a "0" if the threat was not present. Variables that were not applicable to each study were identified. Also, if there was not enough information to ascertain the presence or absence of a threat, it was noted.

CHAPTER IV

Results

For each study, the type of research design was recorded, as well as whether or not there was a clear statement of purpose, statement of hypotheses, statement of research questions, and acknowledgement of limitations. Only 6 of the 62 (9.4%) studies employed experimental methodology. The remaining studies employed 8 quasi-experimental designs, 35 ex post facto designs, and 13 simple survey designs. Ninety five percent (n = 61) of the articles reviewed explicitly stated the purpose of the study. Regarding statement of hypotheses, only 39 studies explicitly stated hypotheses (only two of which explicitly stated null hypotheses). This finding is an increase from the examination by Ellis and Ladany (1997), who found that only 24% of the articles had an explicitly stated hypothesis, which 38.5% left hypotheses implicit. In the current study, only 28 studies outlined research questions in actual question format. Regarding limitations, 58 studies acknowledged and described limitations of their studies. However, no researcher(s) identified all of the methodological threats identified by the current study.

Quantitative Evaluation

From an initial 108 studies, the inclusion/exclusion criteria (Ellis et al., 1996) reduced the sample to a total of 62 studies. Of those 62 studies, 51 included enough information about one (or more) of the statistical tests conducted to make it possible to compute the quantitative statistics. Of the 1,319 tests reported, 1,202 tests contained enough information to perform a power analysis. Table 4 presents the quantitative data averaged across the 1,202 adequately reported statistical tests. Means, standard

deviations, standard errors, and 95% confidence intervals about the median values are presented.

Some findings were noteworthy. The typical investigator conducted 13 statistical tests and discovered significance in 11 of them. The typical sample size was 107 participants. The median sample variance of the dependent variables accounted for by the independent variable was 6.5% translating to a "medium" effect. The median minimum detectable effect size was 5%, which suggests that, on average, as long the investigators were searching for "medium" effects, the tests that they conducted were of sufficient power to find them. This is further evidenced by the post-hoc power for "medium" effects, which had an average value of 79.5% and a median of 89.7% which is in-line with Cohen's guideline (Cohen, 1988) of designing tests with at least an 80% power. The median post-hoc power for "small" effects, however, was only 18.1% indicating that the tests were quite underpowered if "small" effects were actually present in the population and not "medium" effects.

Each study was examined for evidence of 49 possible threats to validity, as outlined by Ellis et al. (1996). Raters reviewed each study and coded the presence of each of the 49 threats with a "yes" or "no." If there was not enough data to suggest evidence of a threat, that variable was coded as "not enough data." Also, if the threat did not apply to a given study, it was coded as "not applicable." The percentages, interrater agreement, and Kappas for the 49 methodological threats across the 62 studies are presented in Table 5. Out of the 62 studies examined, 50 studies contained 15 or more threats to validity (80%). Additionally, 32 of those studies contained 25 or more threats (52%), and 26 studies contained over 35 threats. The threats that were most salient in the articles

reviewed are presented in Table 6. These top threats include *evaluation apprehension* (94%), *irrelevance in experimental setting* (94%), *lack of adequate control group* (92%), *exclusive reliance on self-report data* (86%), *ambiguity of causal direction* (83%), *Instrumentation* (82%), *heterogeneity of participants* (81%), *monomethod bias* (72%), *nonrepresentative supervisee/supervisor population* (61%), *unreliability of dependent/ independent measures* (58%), *interaction of setting and treatment* (57%), *mono-operation bias* (44%), and *unreliability of treatment implementation* (42%).

The threats identified in this study are detailed below, along with the values reported by Ellis et al. (1996). Comparison with Ellis et al. (1996) should be made with several issues in mind: First, the sample size in the current study is half that of the sample examined in Ellis et al. (1996), which decreases the power of this study to replicate the same or similar findings if they are present in the literature. Second, the types of research designs included in the current study are different from those included in Ellis et al. (1996); specifically, the majority of the designs were ex post facto, and therefore not all threats applied to these studies.

Hypothesis Threats (Wampold et al., 1990). Wampold et al. (1990) identified four threats to hypothesis validity. The first threat is *inconsequential hypotheses*, or hypotheses that do not ask critical questions. This essentially refers to whether the hypothesis asks a question that is not already answered by the existing literature. Additionally, Wampold et al. (1990) state that provision of multiple hypotheses reduces the risk of inconsequential hypotheses. In the current study, 100% (n = 62) studies were found to ask a critical question. The second threat regards *ambiguous hypotheses*—specifically, are hypotheses clear and easily understood? In the current study, 70% (n =

48) of the studies had clearly stated hypotheses, while 30% (n = 19) were unclear. The third threat is *non-congruence of research and statistical hypotheses*. In the current study, 78% (n = 50) of the studies included congruent research and statistical hypotheses. The fourth threat is *diffuse statistical hypotheses and tests*, examining whether multiple tests were used to test the hypotheses. If a data set is tested multiple times using different tests, the likelihood of committing a Type I error increases. In the current study, 55% were identified with this threat.

Russell et al.'s (1984) methodological threats. Russell and colleagues (1984) identified 11 threats to methodological validity, which were divided into two categories: internal and external validity. The first threat to internal validity is *lack of an adequate control group*. Ellis et al. (1996) found this to be a threat in 68.67% of the studies reviewed. In the current study, this threat did not apply to 13 survey studies. Out of the remaining 51 studies, 92% (n = 47) had either an inadequate control group or no control group. The second threat, *lack of pretreatment assessment*, was identified by Ellis et al. (1996) to be a threat in 56.67% of the studies reviewed. In the current study, this threat did not apply to 13 survey studies. Out of the remaining 51 studies, 22 % (n = 11) did not have a pretreatment assessment while 78% (n = 40) did. The third threat, *inadequate sample size*, was identified by Ellis et al. (1996) in 22% of the studies examined. The fourth threat, *variations or confounds in length of training across conditions*, was identified by Ellis et al. (1996) as being present in 89.33% of the studies. In this study, however, these threats were not identified because they did not apply to any studies. The fifth threat, *non-random assignment to condition*s, was identified by Ellis et al. (1996) in 40.67% of studies. In the current study, this threat was only applicable to the

experimental designs and the quasi-experimental designs (*n* = 50). Of those, 35% (*n*= 28) did not employ appropriate random assignment, while 65% (n = 15) did. The sixth threat, *widely discrepant cell sizes*, was identified by Ellis et al. (1996) to be a threat in 70.67% of studies. In contrast, the current study found the threat to be present in 30% (n = 19) of the studies. The typical cell categories related to this threat were level of training (degree) and culture. The seventh threat, *restricted range of dependent variables*, was identified by Ellis et al. (1996) to be present in 35.33% of the studies. However, the raters in the current study found this threat to be very difficult to assess. In the studies where the measures were created, not enough information was provided about the measures to assess restriction of range. Even in the articles where authors used established measures, this information was not reported. Consequently, raters chose not to rate this threat due to lack of sufficient information to make an informed assessment. The eighth threat, *non-representative supervisee or supervisor population*, was identified by Ellis et al. (1996) in 91.33% of the studies. This finding was also high in the current study, as the threat was identified in 61% (n = 39) of the studies. These authors used a sample of convenience yet generalized findings beyond that sample. The ninth threat, *lack of follow-up assessment*, was identified by Ellis et al. (1996) in 38% of the studies. In the current study, this threat was not applicable to the 13 survey designs. Of the remaining 49 studies, this threat was present in 17% (n = 11) studies. The tenth threat, *use of role play or audiotaped client statements to assess supervised change*, was found by Ellis et al. (1996) to be present in 92.67% of studies. In the current study, this threat was identified in 80% (n = 51) of the studies. The eleventh threat, *exclusive reliance on self-report data*, was identified by Ellis et al. (1996) in 33.33% of the studies. Unfortunately, the current study found that this

threat has increased in occurrence with the newer research. It was found to be a problem in 86% (n = 55) of the studies. The twelfth threat, *overly brief training period*, was found to be present in 95.33% of the studies examined by Ellis et al. (1996). However, the threat did not apply to any research in the current study because training was not involved in the methods.

Threats to external validity. The three threats to external validity only applied to 14 studies because these studies employed a treatment/intervention. The first threat, *interaction of selection and treatment*, was identified in 36% (n = 5) of those 14 studies (3.33% found in Ellis et al., 1996). The second threat, *interaction of setting and treatment*, was identified in 57% (n = 8) of the 14 studies (10.67% in Ellis et al., 1996). The third threat, *interaction of history and treatment*, was identified in 36% (n = 5) of those 14 studies (17.33% in Ellis et al., 1996).

Threats to internal validity. There are 12 threats to internal validity. The first threat, *history*, was identified in 80.67% of the studies reviewed in Ellis et al. (1996). In the current study, the threat was not applicable to 14 studies. Of the remaining 50 studies, history was found to be a threat in 18% (n = 9) of those studies. In these studies, the time that passed between pretest and posttest may have affected results. The authors of these 9 studies did not address the issue of history at all, whether to deny a threat or identify a possibility of a threat. Consequently, with no information given at all, the raters reasoned that it could not be concluded that the researchers attended to this threat. It is not known if the same considerations were used by Ellis et al. (1996), which could account for the degree of difference in findings. The second threat, *maturation*, was not applicable to 78% (n = 50) of the studies. This is a decrease in the incidence identified by Ellis et al.

(1996), who identified this threat in 79.33% of the studies. In the current study, maturation was considered to be a threat in 36% of the 14 remaining studies (n = 5). In these studies, the participants were students that continued to take classes during the study. Consequently, their learning outside of the experiment could account for some of the data variance. The third threat, *testing*, was also not applicable to 50 of the studies. Testing was found to be a threat in only 1 the remaining 14 studies, (7%). Ellis et al. (1996) identified testing as a threat in 78.67% of the articles they reviewed. In these studies, the same test was given multiple times, thus creating the possibility that participants became familiar with the tests. The fourth threat, *instrumentation*, was identified in 51 of the studies analyzed in the current study (82%). This is consistent with Ellis et al.'s (1996) finding that 94.67% of studies contained this threat.

The fifth threat, *statistical regression to the mean*, was not applicable to 48 studies. Of the remaining 14 studies, 2 were considered to have this threat (14%). This seems to be consistent with the findings by Ellis et al. (1996), who identified this threat in 88.67% of the studies examined. The sixth threat, *interaction of selection and other threats*, was identified by Ellis et al. (1996) in 64% of the studies. The current study identified this threat in only 3% (n = 2) of the articles examined. The seventh threat, *differential attrition*, did not occur in any study—i.e., no participants dropped out in any of the studies reviewed (compared to the observation of 67.33% observed by Ellis et al., 1996). The eighth threat, *ambiguity of causal direction*, Ellis et al. (1996) identified 30% of studies as having this threat. The raters in the current study acknowledge that this threat is an intrinsic problem of ex post facto design and survey research, therefore all studies with those methodologies (n = 48) were identified with this threat. In addition, 5

of the remaining articles were also identified with this threat, concluding with a total of 83% (n = 53) studies. The ninth, tenth, and eleventh threats, *diffusion of treatments, rivalry by participants* and *resentful demoralization*, were neither addressed nor denied in any of the applicable (experimental and quasi experimental, n = 16) research articles. Ellis et al. (1996) identified the percentages of these threats as 99.33%, 98.67%, and 99.33% respectively.

Threats to construct validity. The first threat, *inadequate preoperationalization explication,* was identified as a threat in 31.33% of the studies examined by Ellis et al. (1996). In the current study, it was only identified in 13% (n = 8) of the studies. Overall, the authors did a thorough job of explaining the study variables. The second threat, *mono-operation bias,* was identified by Ellis et al. (1996) in 75.33% of the studies. In the current investigation, this threat was identified in 44% (n = 28) of the sample. These authors employed measures that utilized the same operation (i.e., likert scales). The third threat, *monomethod bias,* was identified in 20.67% of the studies examined by Ellis et al. (1996). In the current study, monomethod bias was a problem in 72% (n = 46) of the studies. Authors in these studies typically only used self-report measures. The fourth threat, *hypothesis guessing within treatments,* was evident in 79.33% of the studies examined by Ellis et al. (1996). In the current study, this threat was present in only 9% (n = 6) of the studies evaluated. It is reasoned that this is tied to the type of research designs employed, specifically ex-post facto designs, which may be less affected by this threat. The fifth threat, *evaluation apprehension,* was identified by Ellis et al. in 78% of studies. This threat was identified in 94% (n = 60) of the current articles. Only three articles addressed issues relating to social desirability or evaluation apprehension, yet all studies

demonstrated possible concerns with these issues. The sixth threat, *experimenter expectancies,* was found by Ellis et al. (1996) in 71.33% of the studies. In the current study, this threat was much less prevalent (5%, n = 3). Again, the issue of research methodology is reasoned to be contributory to this result. Ex post facto designs, which made up the majority of the sample included in this study (n = 50), seem to have been less affected –particularly survey designs. The seventh threat, *confounding of construct with levels of construct,* was identified in 28.7% of studies examined by Ellis et al. (1996). Consistently, this threat was identified in 30% (n = 19) of the studies reviewed in the current study. Typically, these confounds were related to educational degree. The eighth, ninth, and tenth threats did not apply to 50 of the articles reviewed because of their ex post facto designs. As only 12 articles remained for evaluation, comparison with Ellis et al. (1996) should be made with caution. The eighth threat, *interaction of treatments,* was found to be a problem in only 1 study (8%). The ninth threat, *interaction of testing and treatments,* was not identified in any of the 12 studies. The tenth threat, *restricted generalizability across constructs,* was identified in 5% (n = 3) of the remaining articles.

*Threats to Statistical Conclusion Validity***.** The first threat, *low statistical power,* was identified by Ellis et al. (1996) to affect 76.67% of the studies. Consistent with these findings, the current study identified this threat in present in 77% of the studies analyzed. The second threat, *violation of assumption of statistics,* was found in 60% of the studies reviewed by Ellis et al. (1996). The current study identified this threat in 5% (n = 3) of the studies. The third threat, *inflated error rate,* was identified in 14.67% of the studies examined by Ellis et al. (1996). This threat was not identified in any of the studies

reviewed in the current meta-analysis. The fourth threat, *unreliability of dependent or independent variable measures,* was found by Ellis et al. (1996) in 10.67% of the studies. In contrast, this threat was identified to be present in 58% (n = 37) of the studies evaluated in the current study. Some employed newly created measures that had no support data, some used old measures when newer or improved versions were available, and some used measures for which they did not report any data, leaving it to question. The fifth threat, *unreliability of treatment implementation,* was present in 80% of the studies included in Ellis et al. (1996). The current study identified this threat in 42% (n = 28) of studies. The sixth threat, *irrelevance in experimental setting,* was found in 6% of studies evaluated by Ellis et al. (1996). In the current study, it was present in a much greater 94% (n = 60) of the studies. Again, the differences in findings seem to be associated with the type of research designs employed; with ex post facto designs accounting for the majority of the studies in this analysis (13 of which are survey design), it opens opportunities for more irrelevancies in the experimental setting. Specifically, it is impossible to control an experimental setting for participants that complete surveys at home or online. The seventh threat, *heterogeneity of participants,* was identified by Ellis et al (1996) in 40% of the studies evaluated. In the current study, this threat was found in 81% (n = 52) of the studies. For example, mailed questionnaire packets were sent to a select group, and then were returned (presumably) by those most interested in the study. Therefore, the data may reflect simple individual differences in the responding groups that were irrelevant to the phenomena under investigation.

Quantitative Summary

A summary of the methodological threats can be found in Table 5, where they are compared to the findings by Ellis et al. (1996). Inconsistent findings between Ellis et al (1996) and the current study are worth noting. First, threats that occurred less frequently in the current study (i.e., improved) include the following: *threats to hypothetical validity, lack of pretreatment assessment, widely discrepant cell sizes, lack of follow-up assessment, history, hypothesis guessing within treatments, inadequate preoperationalization explication, mono-operation bias,* and *unreliability of treatment implementation.* We cannot infer a true improvement simply by comparing percentages, however. The percentages certainly provide information about the frequency of threats in general, but they do not provide the "why" behind the percentages. Instead, the percentages offer insight into areas of that require further examination. Specifically, percentages do not reflect the shift in employment of research design (for example, what seems to be a reliance on ex post facto design over all others). But this new information about the prevalence of threats in psychotherapy supervision research highlights areas that require further examination by researchers. Second, two threats, *variations or confounds in length of training across conditions* and *overly brief training period* were not applicable in the current study. Third, there were five threats, *unreliability of treatment implementation, restricted range of dependent variables, diffusion of treatments, rivalry by participants,* and *resentful demoralization* that were not coded in the current study because raters felt there was not enough information to assess it. Fourth, there were two threats, *exclusive reliance on self-report data* and *monomethod bias* that presented as threats more frequently than identified by Ellis et al. (1996). This is likely due, again, to the prevalence of ex post facto design. Some categories, (e.g.,

instrumentation and interaction of selection and other threats, experimenter expectancies, violation of assumption of statistics, unreliability of dependent or independent variable measures, and *irrelevance in experimental setting*) were found to be different between Ellis et al. (1996) and the current study. This is reasoned to be a direct result of the types of research methodology used in the research examined in this study. While, like Ellis and colleagues (1996), the majority of the studies in the current investigation were ex post facto in design (48 of 62), 13 of those were survey designs. This was not the case for Ellis and colleagues (1996); the 72% of their sample that utilized ex post facto design did not, in fact, include any survey designs.

Integrative Review

Ellis and Ladany (1997) organized their review according to the six "cardinal inferences" that the authors identified in the supervision articles they evaluated. The authors provided a brief overview of each study, specifically regarding threats to methodological and statistical threats to validity. This format will be used in the discussion of the articles examined in the current study. In the current study, the set of supervision articles published after Ellis and Ladany (1997) were examined in order to assess whether those inferences continued to be represented in psychotherapy supervision research. The second hypothesis of this study was that the supervision literature examined would support the cardinal inferences identified by Ellis and Ladany (1997). Of the six inferences identified by Ellis and Ladany (1997), only one inference was not supported. The cardinal inference category of *Inferences Related to Supervisee Evaluation* was not identified in any of the research articles examined. This is partially due to the exclusion of articles on instrument development. Additionally, it was expected

that new areas of interest, or inference, would be present simply due to the evolution of the field and the passage of time. Specifically, it was hypothesized that the areas of culture and multicultural competence, use of technology, and supervision training would emerge as strong themes in the literature. These themes were supported in the supervision research examined in this study.

While five of the six themes identified by Ellis and Ladany (1997) were supported in general by the current study, modifications to the inference categories were required in order to better fit and reflect the current study data. First, two category titles—*Inferences about the Supervisory Relationship* and *Inferences About Client Outcome* —were maintained but not all of their associated subcategories were supported, and so were modified accordingly (see discussion of each inference that follows). Second, Ellis and Ladany's (1997) inference category, *Inferences Regarding Supervisee Development,* contained research about the process of development for supervisees. The current study included multiple studies that addressed development, but also included other topics related to supervisees (i.e., supervisee perspectives about supervision and supervisee nondisclosure). Therefore, this researcher changed Ellis and Ladany's cardinal inference title to *Inferences Regarding Supervisees*, allowing inclusion of all inferences specific to supervisees. Third, Ellis and Ladany's (1997) second cardinal inference, *Inferences Entailing Supervisor Matching* was determined to be a subcategory of the new inference category, *Inferences about Culture* and so was subsumed by that inference. These modifications better reflected the included studies, and resulted in a total of 7 inferences: *Inferences about the Supervisory Relationship, Inferences Regarding the Supervisee, Inferences about Client Outcome, Inferences about Culture, Inferences about the Use of*

Technology in Supervision, *Inferences about Supervisor Training*, and *General Inferences about Supervision Practice* (see Table 7). An additional note: Since inclusion and exclusion criteria were utilized in examination of the supervision literature, not every supervision article published between 1994 and 2010 was included in this study. Consequently, it cannot be said that the included studies are representative of all supervision studies Therefore, the use of the word "cardinal" to refer to the inference categories is inappropriate—since inferences explain only a subset of the supervision literature, they should not be misconstrued as "cardinal" inferences of all supervision literature. Subsequently, for the purposes of this study and the discussion of findings, the results will be identified simply as inferences.

Before discussing the inferences and associated articles, a brief discussion about the problems associated with reliance on self-report measures is warranted. Since all of the articles included in this study contain some type of self-report measure, it seems useful to detail issues now and then simply refer back to them through the remainder of the integrative review. Self-report measures are inherently susceptible to social desirability contamination and evaluation apprehension. If the topic explored in the measure can cause embarrassment or defensiveness, for example, such as the topic of multicultural competency, then a participant may purposely respond in a socially desirable manner regardless of his or her true feelings. Also, if the participant is completing a measure where evaluation is possible or likely (for example, as part of a class), then the participant may alter his/her responses to meet whatever is deemed most desirable. In light of these concerns about self-report measures, additional measures of the investigated variables that are not self-report (such as observation) should be

employed to reduce possible alternative explanations of results. Having addressed the concerns about self-report here, I will not detail them again but will refer to the concerns about self-report measures where appropriate. Additionally, it is important to note that not every validity threat or methodological flaw in each study is discussed. Instead, the most salient threats are included in the descriptions.

First Inference: Inferences about the Supervisory Relationship

The category of *Inferences about the Supervisory Relationship* includes research about various aspects of the supervisory relationship and issues that affect the supervisory relationship. As found in Ellis and Ladany (1997), the supervisory relationship was found to be the most frequently examined. The authors identified five subcategories of this inference. The subcategories *Supervisory Working Alliance Model, Role Conflict and Ambiguity, and Structure of the Supervisory Relationship* were supported by the literature included in the current study. However, the subcategories *Client-Centered Conditions* and *Strong's (1968) Social Influence Theory* were not represented in the reviewed studies. Two additional subcategories emerged, which have been titled *Supervisor Style*, and *Ethics in the Supervisory Relationship*. Finally, the subcategory of *Parallel Process*, which Ellis and Ladany (1997) listed under Inferences about *Client Outcomes*, was included in this category instead.

Supervisory Working Alliance Model. Several studies examined the supervisory working alliance and/or issues relating to or affecting the working alliance. Ladany, Ellis, and Friedlander (1999) investigated the relationship of the supervisory working alliance with trainee self-efficacy and satisfaction. The use of trainees over practicing professionals, as well as the predominance of female trainees (n = 72), limits the

generalizability of this study. Also, the ex post facto design is limiting because researchers cannot randomly assign participants or predict the direction of causal inferences. Monomethod bias is present with exclusive use of self-report measures. Additionally, all information is from the supervisee's perspective, which introduces bias. Ramos-Sanchez, Esnil, Goodwin, Riggs, Osachy, Touster, Wright, Ratanasiripong, and Rodolfa (2002) conducted an exploratory national supervision study in an attempt to assess the relationship between supervisee developmental level, working alliance, attachment, and negative experiences in supervision. Problems associated with ex post facto research apply. Monomethod bias is present with exclusive use of self-report measures. The sample included 126 respondents, mostly white women, which severely limits generalizability. Of the mailed surveys, only 28% responded, which is less than the typically accepted 50% for mailed surveys. This may have resulted in the statistical conclusion validity threat of random heterogeneity of respondents. Also, the authors used three measures but did not discuss their reliability or validity—though they are established reliable and valid measures. Generalizability is severely limited by the lack of any information about the supervisors.

Two studies examined attachment in the working alliance. The first study by Ligiéro and Gelso (2002) examined the relationship between therapist attachment styles, countertransference behaviors, and working alliance. Monomethod bias is present with exclusive use of self-report measures. Problems with generalizability result from a small sample as well as from the use of trainees for both supervisee and supervisor groups though the study hypotheses is not about trainees. Though many supervisors were doctoral students, they are still students. Additionally, all participants were still attending

classes, which likely affected the way in which they behaved in supervision—and could have resulted in the threat of evaluation bias or socially desirable responding. The simple fact that therapists are still in training and that many supervisors were also still in training is a large concern—results may be more reflective of level of experience/training and not necessarily due to the phenomenon under evaluation. The authors did separate experienced versus less experienced in the analysis and found no significant differences, which may be due to how close the supervisees and supervisors were in training. Also, the authors used a new measure with little validity data. Finally, the problems associated with ex post facto research apply. The second study that examined attachment and the supervisory relationship was conducted by Riggs and Bretz (2006). In this study, 87 doctoral level psychology interns completed an online survey about attachment processes and supervision experience. Problems associated with ex post facto research apply. Additionally, the exclusive reliance on self-report is problematic, as well as the fact that the study only includes the supervisee's perspective of the relationship with the supervisor. There is no data on the supervisors at all, which makes conclusions about them difficult to support and almost impossible to generalize.

*Role Conflict and Ambiguity***.** One study examined role expectations in the supervisory relationship. Tromski-Klingshirn and Davis (2007) investigated supervisees' perceptions of their clinical supervision regarding the dual role of clinical and administrative supervisors. Researchers found that the majority of supervisees receiving clinical and administrative supervision from the same person did not view this dual sup role as problematic. However, the supervisees are mostly masters level (138 masters' to 5 doctoral degrees), and so their appreciation for and understanding of the supervision

process is different than that of doctoral students. Also, the majority of the participants were white females, which limits generalizability. A significant confound is present due to the varying amount of supervision received by the participants—supervisees received 0 to 5 hours per week of individual clinical supervision (*M* = 1.3 hours) and/or 0 to 4 hours per week of group clinical sup (*M* =.08). Additional confounds include the fact that only one state was included and no participants were post-license.

*Structure of the Supervisory Relationship***.** Wilbur and Roberts-Wilbur (1994) conducted a seven-year study about structured group supervision. The participants were all masters' degree students, the majority of which were white women, so results are not generalizable to people of color, men, doctoral students or practicing professionals. One major threat to validity is the fact that the two supervisors in the study were the creators of the group structure under investigation and therefore are biased in the evaluation. Also, the study took place over seven years, and it cannot be assumed that the authors never spoke about the study with each other; this would have affected the manner in which they conducted their supervision. Finally, the assessment was created for the study and has no reported validity or reliability data.

McCurdy and Owen (2008) investigated the use of a sand tray in Adlerian-Based Clinical Supervision. This experimental design had too many confounding variables for readers to make useful inferences from the data. The sample was small *(n* = 31), mostly white (93%). There were two supervisors, but no information was provided about them other than their training. This prevents any generalizability. Only self-reports by the supervisees were collected, and since the supervisees were also attending classes, internal

validity was threatened. In addition, socially desirable responding and/or evaluation contamination is likely due to the student population.

Ellis, Krengel, and Beck (2002) tested self-focused attention theory in clinical supervision and its effects on supervisee anxiety and performance. They clearly stated the study purpose, hypotheses, and research questions. Authors conducted a-priori power analyses to establish necessary sample size and included an appropriate sample. In fact, the only threat found is the possible unreliability of measures, as the reliability and validity are not reported for some methods in coding.

Sterner (2009) investigated the influence of the supervisory working alliance on supervisee work satisfaction and world-related stress in professional settings. The majority of the participants were white females, so results are not generalizable. The cell sizes are widely discrepant regarding doctoral versus masters' degrees and type/amount of supervision per week. Also problematic is that all data is based on self-report, entering the problem of social desirability responding and biased perception of the supervisees. There are no corresponding supervisor reports.

Ladany and Friedlander (2005) examined the relationship between the supervisory working alliance and trainee experience of role conflict and role ambiguity. The majority of the sample was white women, limiting generalizability. Also, the majority of the sample was advanced students at internship level, so results do not apply to all levels of trainees or to practicing professionals.

*Supervisor Style***.** In Ellis and Ladany (1997), the topic of supervisor style was included as a subcategory of *Inferences Entailing Matching*. However, the studies addressing supervisor style in the current study discuss the impact of style on the

supervisory relationship and the working alliance. Consequently, the subcategory fits better within the category of supervisory relationship. Several studies examined the way in which the supervisor's style and/or behavior affect the supervisory relationship. Ladany, Walker, and Melincoff (2001) examined the relationship of supervisory style to the supervisory working alliance and supervisor self-disclosure. The majority of supervisees and supervisors that participated were white women, which affects generalizability. The ex post facto design creates ambiguity about causal inference.

Fernando and Hulse-Killacky (2005) explored the relationship of supervisory styles to satisfaction with supervision and the perceived self-efficacy. There was inconsistency in methodology, as 29 surveys were directly administered by the first author and 54 were mailed to participants. The majority of participants were white women, limiting the generalizability. Also, the participants were from six different masters' degree programs in Florida, Iowa, Louisiana, North Carolina, New Jersey, and Pennsylvania. The problem is that it is not clear how many participants came from each state, and there is no rationale given for the choice of those states. Another problem is that the sample was self-selected, which may suggest random heterogeneity of the sample. The reliance on self-report opens the study to socially desirable responding, especially from those who were administered the measure directly and in person.

Hart and Nance authored and co-authored a series of three articles investigating styles of counselor supervision as perceived by supervisors and supervisees. Hart and Nance (2003) evaluated the preferences of supervisors and supervisees for 4 styles of counselor supervision (*directive teacher, supportive teacher, counselor,* and *consultant*.). Authors found that the styles of *teacher* and *counselor* are used predominantly. In 2004,

Henry, Hart, and Nance investigated the degree of agreement between supervisors and supervisees on topics and content discussed during supervision. Dow, Hart, and Nance (2009) investigated the level of agreement between supervisors and supervisee about the most important topics they discussed about styles of supervision. Throughout all three studies, specific methodological threats are apparent. First, all three studies drew their sample from the same university, and participants included in the studies were predominantly White women. Consequently, generalizability is limited. Second, the participants identified as supervisors 2nd year doctoral students) were theoretically not very different from the supervisees (final semester masters' students). In terms of clinical sophistication and knowledge, there may be differences, but developmentally they are all students not yet practicing as independent professionals. As such, they are influenced by their roles as students and may have very different perspectives from postgraduate professionals. Perhaps the largest concern with these articles is that they examine data collected over 9 years—no participant was involved more than the 10 sessions. However, the passage of time brings to question threats to internal and construct validity, which the authors do not address. Issues regarding history and maturation affect the researchers, and inconsistencies in the experimental setting and implementation threaten statistical conclusion validity. Additionally, random heterogeneity of the respondents, not only due to being at the same university but also being part of a specific time cohort, may threaten the statistical conclusion validity as well.

Johnson and Stewart (2008) used Bandura's (1997) model of competency development to evaluate the sources and level of supervisory self-efficacy among experienced Canadian psychology supervisors. The measure is unreliable and has no

supportive data. The reliance on self-report, the sample from self-selection, and predominance of PhDs affect the validity of the study. The measure was created for the study and had no accompanying validity or reliability data. Canadian antidiscrimination policies were reported as preventing questions about racial or ethnic origins, so it is unclear to what populations the findings could be generalized. Also, monomethod bias affects the interpretation of the results.

Ethics in the Supervisory Relationship. Ladany, Lehrman-Waterman, Molinaro, and Wolgast (1999) investigated the nature and extent of supervisor adherence to the ethical practice of psychotherapy supervision and supervisee satisfaction. Participants were students, so there is a possibility that students have different perspectives on ethics than experienced professionals. This limits generalizability to practicing professional. Also, reliance on self-report is problematic because the participants may not have an entirely accurate recall of what happened with their supervisors. Lack of inclusion of supervisor reports affects validity of conclusions.

Navin, Beamish, and Johanson (1995) surveyed field-based mental health supervisors in Ohio regarding ethical supervisory practices The survey was created for the study and authors did not provide any data on the validity or reliability of the instrument. The reliance on self-report threatens validity, and the design prohibits inference of causal direction of the independent variables. Additionally, this topic is vulnerable to socially desirable responding but no steps were taken to control for this.

Miller and Larrabee (1995) explored the incidence and effect of sexual intimacy in graduate education between faculty members and their students. The authors did not include men in the sample due to the low incidence levels of men having sexual contact

with supervisors, though this may be due to factors specific to gender. A problem with the interpretation of results is that the authors compare the data to studies conducted in 1986 and 1989, but do not take into account that issues related to sexual advances and conduct with supervisors were treated differently at that time. Self-selection may have led to the threat of heterogeneity of respondents.

Parallel Process. Ellis and Ladany (1997) placed the category of Parallel Process under *Client Outcomes*, but the article about parallel process investigated in the current study reflects inferences about the supervisory relationship more than client outcomes. Herron, Primavera, & Ramirez (1997) investigated the existence of parallel process from the viewpoint of supervisors and supervisees. The authors designed the *Parallel Process Survey* for this study, but provided no validity or reliability data on the measure. Reliance on self-report and monomethod bias is also validity threats. Parallel process can go unnoticed by those involved, and so observational data would have been useful. Also, the sample includes both psychodynamic and non-psychodynamic participants, but there are twice as many psychodynamic as non-psychodynamic participants. This likely skewed the data.

Second Inference: Inferences Regarding the Supervisee

Ellis and Ladany (2007) labeled this category as inferences about supervisee development. In the current review, research studies addressed supervisee development along with several other issues specific to supervisees. Consequently, this inference was renamed as inferences about the supervisee (in general), with inferences about *Supervisee Development* becoming a subcategory. The other subcategories of this inference category include *Disclosure and Nondisclosures, Perspectives about the Supervision and the*

Supervisory Relationship and *Supervisee Self-Efficacy.* Ellis and Ladany's (1997) review included the following subcategories : *Ego Development, Conceptual Development, Littrell et al.'s (1979) Model, Hogan's (1964) Model, Loganbill et al.'s (1982) and Sansbury's (1982) Models, Stoltenberg's (1981) Model, Stoltenberg and Delworth's (1987) Model,* and *Generic Supervisee Development and Experience Level.* While some developmental models are discussed in the current review, development is also discussed in general.

Disclosure and Nondisclosure. Ladany, Hill, Corbett, and Nutt (1996) explored nondisclosures in supervision by supervisees. The majority of the participants were white women, limiting generalizability. Also, reliance on self-report and monomethod bias both threaten validity. The ex post facto design resulted in ambiguity about direction of causal inference, but authors acknowledge this. Additionally, the response rate of 50% introduces questions about the characteristics of nonresponders.

Ladany and Lehrman-Waterman (1999) examined the content and frequency of supervisor self-disclosures and their relationship to supervisory working alliance. The participants were mostly white women, limiting generalizability. Limitations were acknowledged by the authors. On limitation is that the supervisees were asked to recall supervisor self-disclosures, and so their answers are affected by memory and by those disclosures most salient to them. Also, the problems associated with ex post facto design and reliance on self-report measures applies.

Perspectives about the Supervision and the Supervisory Relationship. Geller, Farber, and Schaffer (2010) investigated the ways in which trainees construct and use mental representations of their relationships with their supervisors. A departure from the

typical supervision literature, the authors address an interesting aspect of clinical supervision. However, there are no statistical tests performed on the data—only descriptive data is reported. Also, very little is said about the measure created for the study, and therefore it is unreliable. The majority of the participants were white women who reported using a primarily psychodynamic theoretical orientation. No other demographics were reported, which prevents generalizability. The homogeneity of the sample in terms of theoretical orientation and reported demographics makes it difficult to generalize findings.

Supervisee Self-Efficacy. Cashwell and Dooley (2008) investigated the impact of receiving supervision vs. not receiving supervision on counselor self-efficacy. They administered a self-report inventory to professional counselors in a community setting and doctoral level students in a university counseling lab setting. The sample was small (n = 33) and the majority were working professionals (n = 29). Since only 4 members of the sample were doctoral students, it is not possible to generalize any findings to the doctoral student population. There is an additional confound of varying amounts of supervision: 2 received supervision biweekly, 19 received supervision weekly, and one received supervision 6 times per month. Differences in amount of supervision could easily be presented as an alternative explanation of the results. Additionally, 3 counselors had two supervisors, one with a masters and one with a doctorate, which further confounds the results. Overall, there is simply too much variation in the provision of supervision to draw any conclusions.

Supervisee Development. Chagon and Russell (1995) conducted a study in which supervisors viewed videotape vignettes of counselors demonstrating the first three

developmental levels of Stoltenberg's (1981) Counselor Complexity Model. A total of 48 participants were included in the study: 21 men and 27 women served as supervisors, 17 first year doc students comprised the "no experience" supervisor category, 16 third and fourth year doc students comprised the "low experience" category, and 15 counseling psychologists working in either academic or counseling centers comprised the "high experience" category. No other data on the participants is available, so the possibility of confounding data or limited generalizability is present. Videotapes of supervision sessions were developed to match the first three developmental levels of Stoltenberg's model. Since it is analogue, the verbal and nonverbal behaviors may not have been as natural-looking as in a true session (and therefore compromises generalizability). Also categorizing participants into levels of experience by level of training without including other experience may be problematic. The sample is too small to provide generalizability; further, there is little information on the participants. Socially desirable responding and/or evaluation apprehension is likely an issue with the student population.

Birk and Mahalik (1996) examined counselor trainee conceptual level, type of supervision environment, and trainee anxiety as predictors of counselor developmental level. The sample was small (n = 29), was mostly female masters level trainees which limits generalizability. Each trainee saw one client for three session and received supervision from advanced doctoral students or faculty. The different supervisor education level is confounding. Also, the trainees were enrolled in classes during the study, which presents an alternate explanation for results. Additionally, two groups had supervision at their university, which affects evaluation apprehension.

Third Inference: Inferences about Client Outcome

There is little research on client outcome, and unfortunately the articles that did research this topic were found to be seriously flawed. Reese, Usher, Bowman, Norsworthy, Halstead, Rowlands and Chisholm (2009) investigated the impact of client feedback on trainees when used in the context of supervision. Multiple problems with the study prevent even limited generalizability of data. The sample size is small (n = 28) and was mostly white women, all in 2nd year of training in a masters level MFT program. Also, there is significant threat to internal validity due to unreliable measures. The authors created two 4-item measures, items of which were taken from subscales of other measures. There is no reported validity or reliability. In addition, these measures utilized a visual analog scale for responses. The visual analog scale requires that participants place a hash mark on each of the four analog scales that are 10 centimeters long, with scores on the left side of the scale indicating lower functioning and scores on the right indicating higher functioning. A ruler is then used to measure the distance from the left end of the scale to the client's hash mark. The measures for the four items are then summed to provide a total score, ranging from 0 to 40. There is no data to support this type of responding, and the authors do not discuss a rationale for using such a scale. Since there are no studies referenced that utilize such measures, it is impossible to generalize the findings. Perhaps the biggest problem is that there is absolutely no description or demographics of the clients—their individual characteristics and/or presenting problems create serious confounds for the study. The authors report that they address socially desirable responding by informing participants to answer honestly because scores would not be tied to the trainee's grade or evaluation. This is not an effective means for controlling social desirability.

Locke and McCollum (2001) investigated clients' views of live supervision and satisfaction with therapy. While an interesting research topic, the multiple threats to validity inhibit conclusions and generalizability. Participants include therapists, clients, and supervisors. There is absolutely no data on the participants in the study. Some supervisors made more intrusions than others, which may have been viewed positively or negatively by participants. There is no data about the working alliance between the therapists and clients or the therapists and supervisors, which of course could make a big difference in the results. There is also no information on how long the clients have been in therapy or how much experience the counselors have. Overall, the high number of confounding variables prevents any real conclusions from being drawn.

Callahan, Almstrom, Swift, Borja, and Heath (2009) explored the contribution of supervisors to intervention outcomes. A threat to validity is the use of archival data of self-report measures from 76 discharged clients and their therapists. A total of 40 trainees in clinical doctoral programs were included, and they were mostly white and female. There is no actual data on client problems, just the statement that clients had "…a range of common clinical presentations with a mean on Global severity index of 1.11". There is no data on the supervisors; the authors essentially infer supervisor characteristics from the data of clients. Very little can be drawn or generalized from this study.

Fourth Inference: Inferences about Culture

This new category includes research on all areas of culture, which includes race, ethnicity, sexual orientation, spirituality, age, and ability. It is because of this multidimensional view of culture that Ellis and Ladany's (1997) category of *Inferences Entailing Matching* was included as a subcategory here; the studies that address match in

supervision all investigate matching on one or more of the cultural dimensions described above. The subcategories that emerged from the data include the following: *Multicultural Competence, International Trainees/Students, Matching in Supervision, and Gender.*

Overall, there is a concern with socially desirable responding. The topic of multicultural competence contains issues about which people would prefer not to reveal. No studies discussed acknowledgement of this issue or control for social desirability.

*Multicultural Competence***.** Inman (2006) investigated the direct and indirect effects of marriage and family therapy trainees' perceptions of their supervisors' multicultural competence in supervision on the supervisory working alliance, trainees' multicultural competence, and perceived supervision satisfaction. Out of 650 solicited, 147 MFT trainees responded. This response rate of 22.6% is low (average response rate is about 50% for mail surveys), and introduces the possibility of random heterogeneity of respondents (i.e., what is different about the nonresponders?). The sample was mostly white (n = 103), female (n = 121), and masters' level (n = 90, with doctoral degree as the next most frequent at n = 37). The sample limits generalizability. The study findings are limited by reliance on self-report, as social desirability contamination could be a problem.

Ladany, Inman, Constantine, and Hofheinz (1997) examined supervisee multicultural case conceptualization ability and self-reported multicultural competence as functions of supervisee racial identity and supervisor focus. Participants were randomly assigned to one of 2 experimental conditions. In one condition, they were instructed by their supervisor to include issues pertaining to race in their case conceptualization, and in the other they did not. Participants then completed self-report instruments. There is a

possibility of socially desirable responding. Also, the authors admit that persons of color were grouped as one and may have different results with different ethnic groups.

Gloria, Hird, and Tao (2008) assessed the self-reported supervision practices, experiences, and multicultural competence of 211 White intern supervisors supervising predoctoral interns. There is no demographic or descriptive data about the participants, and so participant characteristics (such as sex, race, program of study, past multicultural training, etc.) are confounds. There is also no information about the supervisees. The response rate was only 17%, which is too low to allow any inferences about the results or about the nonresponders. Other threats include unreliability of measures, social desirability contamination, and monomethod bias.

International Trainees/ Students. Mori, Inman, and Caskie (2009) explored the relationship between international trainees' acculturation level and cultural discussion on supervision satisfaction, and how perceived cultural discussion may mediate the relationship between supervisor multicultural competence and trainee satisfaction with supervision. A total of 104 international trainees (84 female, 18 male, and two unknown; mean age of 30) were used as the analysis sample in this study. The authors report that due to low representation of international regions, participants from European countries and Canada were coded as a single category (n = 25) and participants from other regions (south Asia) were also grouped together (n = 71). The grouping, while understandable, eliminates the ability to understand the within group differences of these participants, which the authors acknowledge. Limitations of the study are addressed and discussed by the authors. Regarding methodology, the fact that the survey was online means it was not accessible to all, and 38 of 144 did not complete all of the surveys. The authors suggest

putting demographics at beginning of survey to get a better understanding of who does/doesn't complete. Also, not all scales used in the study had supporting data with an international sample. Issues associated with ex post facto research apply to the study.

Nilsson and Anderson (2004) conducted a study on supervision of international students. Not all instruments used were normed for this population, so it can be argued that the results were not valid. The sample was a subset of a larger sample in a study on training issues with students in APA-accredited professional psychology programs, and so there is no data on how the participants were recruited and there is little detail about how that study was conducted. Additionally, the sample was small (n = 42). Problems associated with ex post facto design are relevant.

Matching in Supervision. Gatmon, Jackson, Koshkarian, Martos-Perry, Molina, Patel, and Rodolfa (2001) explored ethnic, gender, and sexual orientation variables in supervision. The majority of the 289 predoctoral psychology interns was white (n = 219), yet the authors still drew conclusions about cultural match. This is a threat to validity as well as generalizability. Also, the researchers grouped persons of color together, which is problematic because it assumes all people of color have the same views. The reliance on self-report of only the supervisee introduces threats of socially desirable responding. Also, the biased view of the supervisee limits interpretation because there is no corresponding data from supervisors. Additionally, the researchers used an old measure to assess satisfaction with supervision, which is confusing because there are newer measures that assess the same constructs and have validity.

Constantine, Warren, and Miville (2005) investigated whether there are significant differences among progressive, parallel (i.e., both parallel-high and parallel-

low), and regressive white supervisor–white supervisee pairings. The researchers measured self-reported multicultural counseling competence and case conceptualization ability. The majority of the sample was female doctoral students in their 3rd year or beyond and their white doctoral practicum supervisors. The resulting sample had a small number of regressive racial identity dyads. This may be due to socially desirable responding and/or the varying levels of previous multicultural training for both the supervisees and supervisors.

Utsey and Gernat (2002) examined white racial identity attitudes and the ego defense mechanisms used by white counselor trainees in racially provocative counseling situations. Participants were 145 white counselor trainees (majority females with masters' degrees) from small universities, so generalizability is limited. There is a concern with comparing doctoral students to master's students because it is unclear if there is a discrepancy in amount of multicultural training. Therefore the participants' educational backgrounds should be carefully documented. There is a significant concern about socially desirable responding due to the topic of the study, and no steps were taken to address this.

Nilsson & Duan (2001) explored the supervision experiences in 69 U.S. racial/ethnic minority supervisees working with white supervisors. The participants included 33% (n = 23) self-described as Hispanic, Latino, or Latina; 23% *(n* = 16) as African American or Black; 19% *(n* = 13) as multiracial, 16% (n = 11) as Asian American or Pacific Islander, 6% (n = 4) as Arab American, and 3% (n = 2) as American Indian or Alaska Native. The majority were women (n = 49) were women. The major concern about this study is that no information about the supervisors was collected.

Therefore, it is impossible to tell what characteristics or behaviors or training of the supervisors may have affected the results. Also, ex post fact design makes interpretation of the causal direction of influence impossible.

Ladany, Brittan-Powell, and Pannu (1997) explored the influence of supervisor racial identity interact and racial matching on the supervisory working alliance and supervisee multicultural competence. Ex post facto design prohibits making causal inferences, and the data is exclusively supervisee perceptions. This limits interpretations of the data because the supervisees recall and focus may not fully portray the phenomenon.

Bhat and Davis (2007) investigated the role of race, racial identity attitudes, and working alliance in counseling supervision using data obtained from supervisors. The biggest concern with this study is that the supervisors evaluated the racial identity of the supervisees. There is no measure validated for this purpose, and second-hand assessment of someone's racial identity is difficult to trust—the supervisors would be affected by their own racial identity status and social desirability. The majority of the participants was white, female, and had masters' degrees. The only data provided on the supervisees is their race, which was mostly white. This affects generalizability and interpretation.

Gender. Walker, Ladany, & Pate-Carolan (2007) investigated gender-related event in psychotherapy supervision from the perspective of female trainees. The majority of the supervisees and supervisors were white, but no information is provided about dyad composition with the Persons of Color in the sample. The supervisors were not included in the study but were described by the supervisees. The sample was self-selected and was therefore subject to heterogeneity of sample validity concerns. The measure used was

created for the study, but no validity data is presented. Validity threats associated with ex post facto design and reliance on self-report applies.

Wester, Vogel, and Archer (2004) investigated whether male psychology interns would deal with their socialized restricted emotionality in supervision by using either the turning against other or the turning against self-defensive style. The authors also included perception of power in the analysis. Several issues regarding the sample threaten validity. First, the sample included 103 males, majority white, which limits generalizability. Second, the selection of participants presents a problem because they are all still students, consequently their perception of power may be lower than that of post graduates. Third, participants are self-selected, so men with higher levels of RE may not have chosen to participate. Fourth, culture seriously impacts restricted emotionality and is not addressed at all in this study. Differences exist between cultures and ethnicities regarding male expression of emotionality, but this was not addressed. Fifth, 51 participants were interns at veterans hospitals, which tend to have a higher number of men, and this may have affected the results (i.e., perhaps the topic of restricted male emotionality is addressed more frequently in these settings. Validity threats associated with ex post facto design and reliance on self-report applies.

Szymanski (2005) investigated whether feminist supervision practices were related to one's own feminist identity and various beliefs regarding feminism in general. The sample included 135 clinical supervisors (94 female and 41 male, 84% white). Self-report may have created socially desirable responding. Surveys were sent to all APA divisions related to counseling, and division 17 for Advancement of Women was included. There is no data on how many responded from this division, so it could be that

more responded from this division due to interest. This would affect the rating of the phenomena and limit generalizability. This also leads to possible threat of random heterogeneity of respondents.

Fifth Inference: Inferences about the use of technology in supervision

Only two articles fell into this category, yet this researcher found it to be an important topic that is best kept as its own category. The use of technology has changed the way people communicate on a daily basis, and there is no data to suggest that the counseling and supervision world is any different. For example, many clients use the internet to investigate their mental health symptoms and diagnoses, and many mental health providers utilize email to confirm or cancel appointments or communicate with clients between appointments. One can only expect that technology will play a role in clinical supervision as well.

E-mail. Clingerman and Bernard (2004) investigated the use of email as a supplemental modality for clinical supervision. Unfortunately, the design of the study was lacking. The study was quantitative, but their data would have been better served by qualitative examination. The sample was too small, only 19 trainees, and they emailed as part of a class assignment. Authors assigned a one-word category to emails exchanged from supervisee to supervisor, and the largest category was "personalization", which included self-discovery, reflection, and personal growth. However, the sample consisted of students who emailed for a class assignment, so one would not be surprised that their emails most often reflected personalization because they wanted to demonstrate this to their professor (social desirability). They were limited by one email to the professor, with one response, per week, so that further limits what one might ask. A further confound is

that they were in class, and that it was a practicum class—which suggests that they were receiving other supervision at their practicum site and therefore might have considered class as an ancillary supervision. Consequently, there is ambiguity about the direction of causal influence—i.e., was the email format cause for the type of emails sent, or was it due to format restrictions or other supervision? These issues were not addressed by the researchers and not acknowledged as limitations.

Online Discussion. Butler and Constantine (2006) investigated a 12 week web-based peer supervision group to investigate to what degree the group increases collective self-esteem and written case conceptualization ability. Participants included 48 school counselor trainees, 24 (19 women, 5 men, all white) in the web-based peer supervision group and 24 (18 women, 6 men, all white) in the comparison group (no peer supervision). The sample was small and not generalizable. Perhaps the biggest confound is that the students were also taking classes and receiving other supervision, which could provide alternate explanations for results about their self-esteem and conceptualization ability.

Sixth Inference: Inferences about Supervisor Training

While the initial pool of 765 articles included many articles regarding supervisor training, only three research studies met criteria for inclusion in the current study. Stevens, Goodyear, and Robertson (1998) examined changes that might occur in supervisors as they progress from novice to expert. Unfortunately, threats to validity are pervasive. The sample was small (n = 60) and consisted mostly of white females. The sample had a range of supervision experience, from none (n = 12) to over 10 years (*n* =17), with the remaining 31 identified as somewhere in between. This presents a

problem because half of the sample has an "in-between" amount of experience, though the exact amount is not described. This is confounding; what if most of the "in-betweeners" have only 1 or 2 years of experience, for example? It certainly limits how one would interpret the findings. The sample demographics, all living in southern California metro area and mostly white females, limits generalizability. Regarding the methodology, all but 3 participants were administered the experimental procedures in groups that included 2 to 13 participants. They viewed a video of a counseling session by a newly assigned trainee—after viewing this tape, supervisors were asked to list their thoughts about supervision in anticipation of meeting with the trainee. There is a possible effect of the way in which the participants were grouped (who was in their group and did that influence their responses). The measure asked participants to identify the focus of the session from one of five possible topics. This is a major limitation because it misses the opportunity to explore each statement more fully, forcing it into a category. Self-efficacy was measured by only one question about how capable they felt about supervising that one client—and the authors actually the discussed correlates with self-efficacy over just that one question.

Scott, Ingram, Vitanza, and Smith (2000) surveyed APA accredited programs in counseling (60%), clinical (45%), and combined professional scientific programs (29%) regarding training of supervisors. The results combine Psy.D.s and Ph.D.s, so it is not possible to discern differences in training between these two types of programs. The self-report by the program directors is problematic because there is a possibility of socially desirable responding that doesn't reflect actual training practice. The sample is not representative due to the low return rate of 48%—of 256 programs contacted, only 123

were included in the final sample. Additionally, there is no data about the programs/schools that responded (i.e., what part of the country, large or small, etc.).

Seventh Inference: General Inferences about the Practice of Supervision

This category contains articles that surveyed different clinical populations about supervision and supervision practices. The following subcategories emerged: *Group Supervision Practices, Supervision Practices of Academic Faculty, Supervision of Community College Counselors, Supervision of Marriage and Family Counselors, Supervision of School Counselors, Supervision of Substance Abuse Counselors,* and *General Perceptions of Clinical Supervision.* Problems associated with ex post facto studies, monomethod bias, and reliance on self-report measures apply to all studies in this category.

Group Supervision Practices. Riva and Erickson (1995) conducted a survey of group supervision supervisors at predoctoral internships. Only the supervisors were surveyed, with no information on the supervisees or their perceptions. There was no hypothesis and no description of the measure used, so there is no way to know what was actually assessed. The majority of participants were white and male, so it is not generalizable. Also, the characteristics of the group leaders are not detailed or accounted for in the study.

Supervision Practices of Academic Faculty. Tyler, Sloan, and King (2000) conducted a national survey of psychotherapy supervision practices of academic faculty. The return rate was 50% (n = 149), but only 57 were used in the study. This low return rate is a threat to validity. The sample was mostly male (n = 41). The small sample,

heterogeneity of the sample (all from APA Division 12), and self-selection prevents generalizability.

Supervision of Community College Counselors. Coll et al. (1995) conducted a survey of clinical supervision of community college counselors. The sample included 60 community college counselors, with 75% masters' degrees and the remaining 25% were equal parts doctoral degrees, educational specialist degrees, and bachelor's degrees. The mix of degree type is a concerning confound, because it results in different practices and different understandings of the importance of supervision.

Supervision of Marriage and Family Counselors. Lee, Dunn, and Nichols (2005) compared AAMFT approved supervisors with masters' and doctoral degrees. One major problem is that the data of 'relevant items' were pulled from a national survey conducted by Lee et al. (2004). The authors did not discuss anything about the Lee study or the methodology. Additionally, there is no discussion of how many items or what types of items were included in the present study. The sample included equal number of males and females, but was mostly white and so generalizability is limited. There is no discussion about how the sample was recruited, which threatens validity.

Carlozzi, Romans, Boswell, Ferguson, and Whisenhunt (1997) surveyed directors of training of programs accredited by the Commission on Accreditation for Marriage and Family Therapy Education (CACREP) and the Council for Accreditation of Counseling and Related Educational Programs (COAMFTE) regarding training and supervision practices. The major problems associated with this study include unreliability of measures and self-selection of the sample. Also, only training directors reported, so it could be that this did not represent the actual practice.

Anderson et al. (2000) surveyed family therapy trainees about best and worst supervision experiences. The sample included trainees from programs holding COAMFTE accreditation. The majority of the sample was female and white, and a high percentage of the sample worked in college counseling centers. These issues limit generalizability. The measure was created for the study and no supporting data is reported, so it is an unreliable measure. There is also a concern with self-selection and question about the characteristics of nonresponders.

Kanno and Koeske (2010) surveyed MSW students about supervision and satisfaction with their field placements. The procedure is problematic because the survey was completed by MSW students while they were in class. The authors report that participation was optional, but when given in class it seems unlikely that many people would refuse. Therefore, there is a strong argument for evaluation apprehension and socially desirable responding. There is an additional problem that all participants are from the same school and that the measure, created for the study, has no reliability or validity.

Supervision of Rehabilitation Counselors. Schultz, Ososkie, Fried, Nelson, and Bardos (2002) surveyed counselors about supervision practices in public rehabilitation counseling settings. A confound is presented by the fact that counselors were employed by the Division of Vocational Rehabilitation in only two states, and therefore are a heterogeneous group. This limits generalizability. The majority of the participants were white females with master's degrees, also limiting generalizability. Also, the measure was unreliable.

Supervision of School Counselors. Studer & Oberman (2006) surveyed school counselor trainees about supervision practices provided to school counselors. The response rate was low (only 37%) and the sample was mostly white females with masters' degrees. Therefore, generalizability is limited. There is very little information about the measure, and it has no validity or reliability. Kahn (1999) conducted a survey of 119 supervisors about priorities and practices in field supervision of school counseling students. The authors followed up with structured telephone interviews with 12 participants, though they do not explain or support this procedure. The majority of the sample was white and female, and all were from Pennsylvania, so the generalizability is limited. A confound was that 74.5 % of the supervisors were providing supervision for the university where they had received their training. Page, Pietrzak, and Sutton (2001) conducted a national survey of school counselor supervision. The participating counselors were surveyed regarding their current supervision, desire for clinical supervision, and rating of supervision goals, yet the majority of the participants were not currently receiving supervision. This seriously affects the utility of the findings, as the authors apply what these unsupervised counselors *want* to what supervised school counselors experience in supervision.

Supervision of Substance Abuse counselors. Culbreth (1999) surveyed clinical substance abuse counselors about clinical supervision practices. The strongest part of this article is that the authors referenced Ellis et al. 1996! Problems with the study include a very low response rate of 35% and a sample with too wide a range of education. Different levels of education result in different views of supervision, and therefore confound the results. Another confounding variable is whether the counselor is in recovery or not,

which may affect view of supervision. Socially desirable responding is a possibility, especially since the results show participants do not care about supervisor recovery status while the literature suggests the opposite.

General Perceptions of Clinical Supervision. McCarthy, Kulakowski, and Kenfield (1994) surveyed licensed psychologists from one Midwestern state to assess the nature of clinical supervision for experienced practitioners. A majority of the participants were female, white, and had a Ph.D., which limits generalizability. Additionally, the measure created for the study has no reported validity or reliability data.

Borders, Cashwell, and Rotter (1995) conducted a survey of supervisors of counselor licensure applicants in two states, and results indicated that state boards' supervision regulations do have some impact on the practice of supervision. A problem is that all levels of education were combined (doctoral, masters' and specialist), so it is impossible to tell from the write-up what happened with each group. Additionally, all participants were from South Carolina and Missouri. South Carolina was chosen because it is the only state to have license for supervisor of counselor licensure applicants, which therefore makes the results even less generalizable. No reliability or validity data is provided for the measure used in the study.

Romans, Boswell, Carlozzi, and Ferguson (1995) surveyed training directors of counseling, clinical, and school psychology programs accredited by the APA on training and supervisory practices and perceptions of various modalities of supervision. The main concerns are the majority of white participants, the low school response rate, and socially desirable responding.

CHAPTER V

Discussion

The purpose of the current study was to replicate and extend the findings of Ellis et al. (1996), who critically examined the state of psychotherapy supervision research published between 1981 and 1994. The current study, following Ellis et al.'s (1996) methodology, evaluated the state of psychotherapy supervision research published from 1995 through 2010. The researcher sought to ascertain the level of methodological rigor of the recent studies, as well as investigate the amount of attention paid by researchers to controlling threats to the validity of their research. Consistencies as well as differences were identified regarding validity threats between Ellis et al.'s (1996) finding and the current study (See Table 5 for the comparison between Ellis et al., 1996, and the current study). Consistent findings of high levels of threats to validity are most pertinent in this study. As such, the most salient threats identified in both studies include the following validity categories of Russell et al.'s (1984) threats: *lack of an adequate control group, nonrandom assignment to condition*s, *non-representative supervisee or supervisor population, use of role play or audiotaped client statements to assess supervised change, evaluation apprehension, confounding of construct with levels of construct,* and *inflated error rate.* Each of these threats was identified in a high percentage of studies in both Ellis et al. (1996) and Ellis and Ladany (1997).

It is essential to emphasize that the intent of this study was not to devalue current literature, but instead to highlight areas in current psychotherapy supervision research that require further attention and improvement. The research examined in the study reflects attention to some, but not all, of the possible threats to the validity of research

investigations. The criteria utilized in the current study for identifying threats to validity (i.e., Cook & Campbell, 1979; Russell et al., 1984; Wampold, Davis, & Good III, 1990) are therefore particularly valuable. These criterions can provide a "checklist" of sorts for researchers to consult as they design and execute research ideas.

The first hypothesis of the current study was that the supervision research published from 1994 through 2010 would show improvements from the studies investigated by Ellis et al. (1996). Specifically, it was hypothesized that the literature would reveal a more careful approach to study design with attention to minimizing threats to validity, methodology, and hypotheses. This improvement was hypothesized to occur either due to researchers reading and employing recommendations from Ellis et al. (1996) and Ellis and Ladany (1997) and/or due to increased sophistication with research design that may occur naturally as the topic of supervision areas is explored and refined over time. In general, the quantitative findings are consistent with those reported by Ellis et al. (1996). Specifically, methodological flaws were identified in every study, to varying degrees. These most salient threats identified in the studies include the following: *evaluation apprehension* (94%), *irrelevance in experimental setting* (94%), *lack of adequate control group* (92%), *exclusive reliance on self-report data* (86%), *ambiguity of causal direction* (83%), *Instrumentation* (82%), *heterogeneity of participants* (81%), *monomethod bias* (72%), *nonrepresentative supervisee/supervisor population* (61%), *unreliability of dependent/ independent measures* (58%), *interaction of setting and treatment* (57%), *mono-operation bias* (44%), and *unreliability of treatment implementation* (42%). In the studies where these threats were evident, the researchers often did not appear to attempt to control for these threats or discuss the

possible ramifications of these threats. It is hoped that the findings here will further emphasize to researchers the need for careful research design and execution, and further support applications of the recommendations put forth by Ellis and colleagues (1996).

The second hypothesis of the current study centered on the six *Cardinal Inferences* of psychotherapy supervision research introduced by Ellis and Ladany (1997). It was hypothesized that these inferences would continue to be major themes in the supervision literature published from 1995 through 2010. Only one of the inferences, *Inferences Related to Supervisee Evaluation,* was not supported. This is partially due to the exclusion of articles on instrument development. As expected, the evolution of the field of supervision led to research interest in new areas. In particular, the areas of culture and multicultural competence, use of technology, and supervision training appeared prominently in current research. These were consequently identified as new inferences. In general, findings of the current study indicate that the basic themes intrinsic to supervision research from1981 through 1994 continue to guide the field. And as would be expected, the field has continued to grow and expand with the passage of time.

Limitations

Quantitative Limitations

The first limitation of the current study is that a large percent of the original set of articles was eliminated from the study due to use (either primarily or mostly) of qualitative methodology. As a result, the true state of clinical supervision research cannot be fully assessed from these findings, and therefore the results of this investigation cannot be generalized to all published clinical supervision research.

The second limitation regards the bias in the statistical data. Ellis and Ladany (1997) identified three main issues affecting their statistical data, and all three of these issues affect the current study as well (p. 146). First, not all authors presented complete statistical information for the tests performed in their studies, which affected this researcher's ability to compute quantitative data for these studies (n = 51). Second, most authors reported complete statistical data for significant tests but reported little or no data for non-significant tests. Therefore, results of analyses are based only on tests that had completely reported data. The third statistical bias issue regards the "file-drawer problem" (Rosenthal, 1979). In most large-scale literature reviews, there is a heavy reliance on published studies. According to Rosenthal (1979), this results in overestimated significance of published results because studies that do not show significant results are not likely to be published. Consequently, distribution of effect sizes are biased, skewed, or completely cut off, creating a serious base rate fallacy. Rosenthal notes that, "…for any given research area, on cannot tell how many studies have been conducted but never reported" (Rosenthal, 1979, p. 638). He states that, "…the sobering lesson is that small numbers of studies that are not very significant, even when their combined *p* is significant, may well be misleading in that only a few studies filed away could change the combined significant result to a non-significant one." (p. 640). While it is assumed that unpublished research has more flaws than published research, the results of the current study still do not represent the scope of supervision research conducted during the time period evaluated.

The third limitation is that the exact sample size used for a statistical test was not always clear. Ex post facto research (*n* = 50), in particular, was affected by imprecise

sample size; while the total number of participants was always stated, the exact number of participants that responded to a particular piece of the survey was much less clear. In the case where *F* tests were performed, the researcher observed that the degrees of freedom reported with the statistic sometimes implied a slightly different sample size than what would be implied by the stated sample size. An example can be found in Nilsson and Anderson (2004). The authors performed a multiple regression in the first step of a step-wise regression procedure. The study authors reported the following *F* statistic: "F (1, 38) = 6.04, p < .02". Given the stated degrees of freedom and information on the kind of test performed, the sample size can be inferred as follows:

$$n = df2 + (df1 + 1)$$
$$= 38 + (1 + 1)$$
$$= 40$$

However, the authors' reported sample size was 42. Since it was not known whether there were unreported reductions to the degrees of freedom, the researcher chose to use the reported sample size (n = 42) to compute the non-centrality parameter and the reported degrees of freedom to compute the sample size. The impact is that power is potentially overstated by a small amount since the non-centrality parameter is an increasing function of sample size and statistical power is an increasing function of the non-centrality parameter. In the previous example the post-hoc power using the sample effect size implied by the *F* statistic and the stated sample size (n = 42) was 91%. If one uses the sample size implied by the degrees of freedom (n = 40) and the *F* statistic, then the post-hoc power is 90%. It is important to note that in cases where the degrees of freedom implied a drastically different sample size from the reported sample size, that

study authors tended to clearly report the smaller sample size in table footnotes. Therefore, readers have an accurate picture of the data and results.

The fourth limitation has to do with the coding training set. From conclusions reached by Ellis and Ladany (1997), it was recognized that not all validity threats applied to all types of research designs. Consequently, the researcher identified a-priori which types of threats were not applicable to which studies as part of the coding training (for example, pre and posttests do not apply to survey design). When assembling the set of 20 articles to establish interrater reliability, it was not possible to identify the type of research designs ahead of time since the selection was random. Therefore, the number of studies utilizing of each type of design was not equally present in the training set—and considering the varying amounts of the designs in the study, they were not all equally likely to be included. As a result, the initial interrater reliability was not equally tested on all designs, which may have affected the results.

Qualitative Limitations

Limitations of the coding methodology require discussion. First, the coding is affected by random irrelevancies in the coding setting and by extraneous factors in the lives of the raters that may have influenced the ratings. For example, the first rater conducted the coding on a full-time daily basis, while also taking care of her 2 year old daughter for part of the day. The second rater conducted coding during hours when she was not at her pretdoctoral internship. The third rater conducted statistical coding when not at his full time employment, where he works with mathematical equations all day. Multiple uncontrollable variables—such as levels of stress, tiredness/alertness, pressure to code 'correctly', amount of time between coding sessions, and contamination/overlap

of work performed during the day, for example—all have an impact on coding results. Therefore, results should be interpreted with these issues in mind.

The second limitation regards the codes assigned to the 49 methodological threats. Ellis and Ladany (1997) found little or no variability on 6 of the threats investigated (p. 460) (instrumentation, statistical regression, diffusion of treatment, compensatory equalization, rivalry by participants, and resentful demoralization). Likewise, the current study found little variability in these threats. Ellis and Ladany (1997) suggested that this lack of variability may be explained by the fact that these evaluation criteria were not applicable to a large percentage of the research design and methodologies employed in the studies investigated (p. 460). This explanation is certainly applicable to the current study. Most of the studies were ex post facto in design (n = 50), 13 of which were survey design. Quasi-experimental design was the second most utilized design (n = 8).

Recommendations for future research

Literature review. In general, it seems most useful to recommend that researchers review literature specific to conducting research (e.g., Ellis et al., 1996; Russell et al., 1984). These articles address all areas of the research methodology, pointing out areas that researchers often forget about. It is not possible create a perfect study, and it is expected that designs will have flaws. However, researchers who understand the threats to study validity can take steps to control for them. Also, if it is not possible to control for some threats, then the authors can address it in their discussions for readers to use in interpreting and applying the results.

Measure selection. As seen repeatedly in this study, reliance on self-report is pervasive in supervision research. Further, reliance on self-report from only one

perspective of a phenomena is also ubiquitous (e.g., supervisee self-report about supervisory relationship without supervisor perspective). Results are difficult to interpret and generalize if no other data is available to support or contradict the self-report data. Additionally, it is important to report supporting data of the measures used in a study. Even if the measures are pervasive in the literature, it should not be assumed that all readers are aware of the psychometric properties. Without this data, readers cannot accurately assess or apply research findings to their own work.

Power analysis. It is very important for researchers to conduct an a-priori power analysis. Ascertaining the number of participants needed to reach a medium or large effect size allows researchers to gather enough participants to support their findings—or acknowledge that they will not be able to meet the participant requirement and adjust their studies accordingly. Researchers dedicate a great deal of time and energy to their research designs and implementation; identifying the appropriate sample size is a relatively easy thing to do in order to ensure that hard work concludes in usable results.

Replications and Extensions. Lindsay and Ehrenberg (1993) suggest that one reason for the low frequency of replication and extension articles is that replication is unexciting. However, replication is often necessary so that research findings are confirmed and refined. As was found in this study, researchers conduct interesting studies whose results are clouded by multiple limitations. Supervision research could be moved forward if researchers thoroughly review previous literature and build on it.

Previous research. All measures used in a study should be the newest and most reliable measures available that assesses the variables being investigated (for example, see Gatmon et al., 2001). Replication, revision, or modification of measures with

established validity/reliability can often be more useful than creating a new measure of the same construct—instead of furthering the knowledge base, it may unnecessarily complicate the evaluation of the construct(s).

Sample. Is the sample representative of the population under investigation? Are there confounds in the levels of training or experience? It is important to gather and report necessary information about the sample so as not to introduce confounds. Many researchers used a sample of convenience, which invariably hurt them in terms of threats to internal validity and external validity. This issue is particularly relevant to survey research, as without a representative sample, almost no conclusions can be drawn with any degree of confidence. Another issue with sample selection is the use of students as the primary sample. Admittedly, they are often easier to access and persuade to participate than working professionals in the field. But use of students adds alternative explanations to results, such as evaluation apprehension, desire to please supervisors/teachers, lower levels of counseling skill and conceptualization ability, etc. Students also are attending classes and learning while involved in the study, so if the study extends a few months, then a threat to internal validity occurs—i.e., how can one tell what is a result of the experiment versus what is a result of the student learning from their coursework? Additionally, when professionals read articles with student samples, they may feel as though they are far removed from a student mentality and thus have little in common with the sample—and the results.

Social desirability. Many studies utilized self-report data, which is particularly susceptible to socially desirable responding. This is an even larger threat with certain sample populations, such as trainees, and with certain topic areas, such as multicultural

competence. Researchers rarely addressed controlling for social desirability, yet it is impossible to overlook the possible contribution of social desirability to the results. While it is probably impossible to completely remove the possibility of socially desirable responding, it is important for researchers to do everything they can to control for it. This may include ensuring that research with trainees is completely separate from their schooling or education so that their responses are unlikely to be motivated by desire to please professors or fear of evaluation. In situations where this type of complete separation is not possible, administration of a social desirability scale may be appropriate. Additionally, using methods besides self-report (e.g., observations) can counter socially desirable responding.

Acknowledge limitations. There will be limitations in every study, but acknowledging those in detail and explaining why they occurred will go a long way to helping readers understand the study, understand how it can be used, and possibly replicate with the limitations corrected (e.g., Ladany et al., 1997; Szymanski, 2005). Seven of the research articles (11%) reviewed in this study did not acknowledge any limitations. Of the remaining 55 studies, 42 studies (76%) only acknowledged a small percentage of the threats that were actually present. When no limitations are acknowledged (see Ramos-Sanchez et al., 2002; Wester et al., 2004) it suggests that the researchers did not consider or attempt to control for limitations. Even if the study was perfectly designed and executed, consumers may question the results because there is no evidence of self-reflection or critical observation about the study design.

Qualitative Research. The review of supervision literature for inclusion in this study highlighted the popularity of qualitative research methodology. In fact, seventy-

seven articles containing qualitative methodology, either primarily or partially, were reviewed during the early phase of article collection—and these seventy-seven are only the supervision articles that appeared to meet Ellis et al.'s criteria for inclusion. Meaning, the actual number of studies including qualitative methodology in supervision and counseling is higher. Certainly this methodology offers insight into the process of supervision that cannot be reached by quantitative means alone, making it a very valuable research design. Additionally, it is often very enjoyable to read because of its narrative nature. Given the number of studies, it is apparent that an evaluation of the state of qualitative research in supervision is needed.

Several methods of conducting and coding qualitative research exist, many of which have been supported through research. For example, *Consensual Qualitative Research* (see Hill, Thompson, & Williams, 1997) and *Discovery-Oriented* research methodology (see Mahrer, 1988) are two methods supported by research and replication. However, as per this researcher's brief review of qualitative studies, some researchers choose to create their own systems of coding and analyzing qualitative data that may or may not result in an accurate reflection of the study data. Research guides practice, and poorly designed research leads to bad practice. A review of the published qualitative research is needed to assess the quality of published qualitative studies in clinical supervision so that practitioners and researchers can better understand and improve the research that guides supervision.

References

American Psychological Association (1996). Office of program consultation and accreditation guidelines and principles for accreditation of programs in professional psychology. Washington, DC.

American Psychological Association (2000). Office of program consultation and accreditation guidelines and principles for accreditation of programs in professional psychology. Washington, DC.

American Psychological Association. (2002). Ethical principles of psychologists and code of conduct. *American Psychologist, 57*, 1060 - 1073. doi: http://dx.doi.org/10.1037/0003-066X.57.12.1060

American Psychological Association. (2003). Guidelines for multicultural education, training, research, practice, and organizational change for psychologists. *American Psychologist*, *58*, 377 - 402. doi: http://dx.doi.org/10.1037/0003-066X.58.5.377

Abramowitz, M. & Stegun, I.A. (1972). *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. New York: Dover Publications.

Allen, G.J., Szollos, S.J., & Williams, B.E. (1986). Doctoral students' comparative evaluations of best and worst psychotherapy. *Professional Psychology: Research and Practice, 17*, 91 – 99. doi: http://dx.doi.org/10.1037/0735-7028.17.2.91

Alpher, V.S. (1991). Interdependence and parallel processes: A case study of structural analysis of social behavior in supervision and short-term dynamic psychotherapy. *Psychotherapy: Theory, Research, Practice, Training, 28*, 218 - 231. doi: http://dx.doi.org/10.1037/0033-3204.28.2.218

Ancis, J.R. & Marshall, D.S. (2010). Using a multicultural framework to assess supervisees' perceptions of culturally competent supervision. *Journal of Counseling & Development, 88,* 277 - 284. doi: http://dx.doi.org/10.1002/j.1556-6678.2010.tb00023.x

Ancis, J.R. & Ladany, N. (2001). A multicultural framework for counselor supervision. In L.J. Bradley & N. Ladany (Eds.), *Counselor supervision principles, process, and practice* (pp. 63 - 86). Ann Arbor, MI: Taylor & Francis.

Anderson, S.A., Schlossberg, M., & Rigazio-DiGilio, S. (2000). Family therapy trainees' evaluations of their best and worst supervision experiences. *Journal of Marital and Family Therapy, 26,* 79 - 91. doi: http://dx.doi.org/10.1111/j.1752-0606.2000.tb00278.x

Bahrick, A.S., Russell, R. K., & Salmi, S.W. (1991). The effects of role induction on trainees' perceptions of supervision. *Journal of Counseling & Development, 69,* 434 - 438. doi: http://dx.doi.org/10.1002/j.1556-6676.1991.tb01540.x

Barnett-Queen, T. & Larrabee, M.J. (2000). Sexually oriented relationships between educators and students in mental-health-education programs. *Journal of Mental Health Counseling, 22,* 68 - 84.

Bear, T.M. & Kivilighan, M. (1994). Single-subject examination of the process of supervision of beginning and advanced supervisee. *Professional Psychology: Research and Practice, 25,* 450 - 457. doi: http://dx.doi.org/10.1037/0735-7028.25.4.450

Benshoff, (1993). Peer supervision in counselor training. *The Clinical Supervisor, 11,* 89 – 102.

Bernard, J.M. & Goodyear, R.K. (1992). *Fundamentals of clinical supervision.* Needham Heights, MA: Allyn & Bacon.

Bernard, J.M. (1979). Supervisory training: A discrimination model. *Counselor Education and Supervision, 19,* 60 - 68. doi: http://dx.doi.org/10.1002/j.1556-6978.1979.tb00906.x

Bhat, C.S & Davis, T.E. (2007). Counseling supervisors' assessment of race, racial identity, and working alliance in supervisory dyads. *Journal of Multicultural Counseling and Development, 35,* 80 - 91. doi: http://dx.doi.org/10.1002/j.2161-1912.2007.tb00051.x

Birk, J.M. & Mahalik, J.R. (1996). The influence of trainee conceptual level, trainee anxiety and supervision evaluation on counselor developmental level. *The Clinical Supervisor, 14,* 123 -137. doi: http://dx.doi.org/10.1300/J001v14n01_09

Borders, L.D. & Brown, L.L. (2005). *The new handbook of counseling supervision* (2nd ed.). Mahwah, NH: Erlbaum. doi: http://dx.doi.org/10.1002/j.1556-6978.1994.tb00294.x

Borders, L.D. & Fong, M.L. (1994). Cognitions of supervisors-in-training: An exploratory study. *Counselor Education and Supervision, 33,* 280 - 293. doi: http://dx.doi.org/10.1002/j.1556-6978.1995.tb00209.x

Borders, L.D., Cashwell, C.S., & Rotter, J.C. (1995). Supem rvision of counselor licensure applicants: A comparative study. *Counselor Education and Supervision, 35,* 54 - 69. doi: http://dx.doi.org/10.1002/j.1556-6978.1995.tb00209.x

Bordin, E.S. (1983). A working alliance based model of supervision. *The Counseling Psychologist, 2,* 35 - 41. doi: http://dx.doi.org/10.1177/0011000083111007

Burkard, A.W., Johnson, A.J., Madson, M.B., Pruitt, N.T., Contreras-Tradych, D.A., Kozlowski, J. M., et al. (2006). Supervisor cultural responsiveness and unresponsiveness in cross-cultural supervision. *Journal of Counseling Psychology, 53,* 288 - 301. doi: http://dx.doi.org/10.1037/0022-0167.53.3.288

Butler, S. K. (2003). Multicultural sensitivity and competence in the clinical supervision of school counselors and school psychologists: A context for providing competent services in a multicultural society. *Clinical Supervisor, 22,* 125 - 141. doi: http://dx.doi.org/10.1300/J001v22n01_09

Butler, S.K. & Constantine, M.G. (2006). Web-based peer supervision, collective self-esteem, and case conceptualization ability in school counselor trainees. *Professional School Counseling, 10,* 146 - 152.

Callahan, J.L., Almstrom, C.M., Swift, J.K., Borja, S.E., & Heath, C.J. (2009). Exploring the contribution of supervisors to intervention outcomes. *Training and Education in Professional Psychology, 3,* 72 - 77. doi: http://dx.doi.org/10.1037/a0014294

Campbell, D.T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin, 54,* 297 - 312. doi: http://dx.doi.org/10.1037/h0040950

Campbell, D.T., Stanley, J.C., & Gage, N.L. (1963). Experimental and quasi-experimental designs for research. Boston, MA: Houghton Mifflin and Company.

Carlozzi, A.F., Romans, J.S.C., Boswell, D.L., Ferguson, D.B., & Whisenhunt, B.J. (2001). Training and supervision in counseling and marriage and family therapy

programs. *The Clinical Supervisor*, *15*, 51 - 60. doi: http://dx.doi.org/10.1300/J001v15n01_04

Carter, J.W., Enyedy, K.C., Goodyear, R.K., Arcinue, F., & Puri, N.N. (2009). Concept mapping of the events supervisees find helpful in group supervision. *Training and Education in Professional Psychology*, *3*, 1 - 9. doi: http://dx.doi.org/10.1037/a0013656

Cashwell, T.H., & Dooley, K. (2001). The impact of supervision on counselor self-efficacy. *The Clinical Supervisor, 20*, 39 - 47. doi: http://dx.doi.org/10.1300/J001v20n01_03

Chagnon, J. & Russell, R.K. (1995). Assessment of supervisee developmental level and supervision environment across supervisor experience. *Journal of Counseling & Development, 73*, 553 - 558. doi: http://dx.doi.org/10.1002/j.1556-6676.1995.tb01793.x

Chui, E.W.T. (2010). Desirability and feasibility in evaluating fieldwork performance: Tensions between supervisors and students. *Social Work Education, 29*, 171 - 187. doi: http://dx.doi.org/10.1080/02615470902912219

Clingerman, T.L. & Bernard, J.M. (2004). An investigation of the use of e-mail as a supplemental modality for clinical supervision. *Counselor Education and Supervision*, *44*, 82 - 95. doi: http://dx.doi.org/10.1002/j.1556-6978.2004.tb01862.x

Coleman, M.N., Kivlighan, D.M., Jr., & Roehlke, H.J. (2009). A taxonomy of the feedback given in the group supervision of group counselor trainees. *Group*

Dynamics: Theory, Research, and Practice, 13, 300 - 315. doi: http://dx.doi.org/10.1037/a0015866

Coll, K.M. (1995). Clinical supervision of community college counselors: Current and preferred practices. *Counselor Education and Supervision, 35*, 111 - 117. doi: http://dx.doi.org/10.1002/j.1556-6978.1995.tb00215.x

Constantine, M.G., Warren, A.K., & Miville, M.L. (2005).White Racial Identity dyadic interactio ns in supervision: Implications for supervisees' multicultural counseling competence. *Journal of Counseling Psychology, 52*, 490 - 496. doi: http://dx.doi.org/10.1037/0022-0167.52.4.490

Cooper, J.B. & Ng, K.(2009). Trait emotional intelligence and perceived supervisory working alliance of counseling trainees and their supervisors in agency settings. *International Journal for the Advancement of Counselling, 31*, 145 - 157. doi: http://dx.doi.org/10.1007/s10447-009-9074-4

Cook, T.D., & Campbell, D.T. (1979). Quasi-experimentation: Design and analysis for field settings. Boston: Houghton Mifflin.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology, 65*, 145 – 153. doi: http://dx.doi.org/10.1037/h0045186

Cohen, J. (1973). Eta-squared and partial eta-squared in fixed factor ANOVA designs. *Educational and Psychological Measurement, 33*, 213 - 218. doi: http://dx.doi.org/10.1177/001316447303300111

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.) New York: Academic Press.

Constantine, M.G., Warren, A.K., & Miville, M.L. (2005). White racial identity dyadic interactions in supervision: Implications for supervisees' multicultural counseling competence. *Journal of Counseling Psychology, 52,* 490 - 496. doi: http://dx.doi.org/10.1037/0022-0167.52.4.490

Cresswell, J.W. (1998). *Qualitative inquiry and research design: Choosing among five traditions*. Thousand Oaks, CA: Sage.

Culbreth, J.R. (1999). Clinical supervision of substance abuse counselors: *Current and preferred practices. Journal of Addictions & Offender Counseling, 20,* 15 - 25. doi: http://dx.doi.org/10.1002/j.2161-1874.1999.tb00137.x

Cumming, G. (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on psychological science, 3,* 286 - 300. doi: http://dx.doi.org/10.1111/j.1745-6924.2008.00079.x

D'Andrea, M., Daniels, J., & Heck , R. (1991). Evaluating the impact of multicultural counseling training. *Journal of Counseling & Development, 70,* 143 – 150. doi: http://dx.doi.org/10.1002/j.1556-6676.1991.tb01576.x

Dennin, M.K., & Ellis, M.V. (2003). Effects of a method of self-supervision for counselor trainees. *Journal of Counseling Psychology, 50,* 69 - 83. doi: http://dx.doi.org/10.1037/0022-0167.50.1.69

DeMayo, R.A. (2000). Patients' sexual behavior and sexual harassment: A survey of clinical supervisors. *Professional Psychology: Research and Practice, 31,* 706 - 709. doi: http://dx.doi.org/10.1037/0735-7028.31.6.706

Der Pan, P.J., Deng, L. F., & Tsai, S.L. (2008). Evaluating the use of reflective counseling group supervision for military counselors in Taiwan. *Research on*

Social Work Practice, 18, 346 - 355. doi: http://dx.doi.org/10.1177/1049731507313981

Dodenhoff, J.T. (1981) Interpersonal attraction and direct–indirect supervisor influence as predictors of counselor trainee effectiveness. *Journal of Counseling Psychology*, *28*, 47 - 52. doi: http://dx.doi.org/10.1037/0022-0167.28.1.47

Doehrman, M. (1976). Parallel processes in supervision and psychotherapy. *Bulletin of the Menninger Clinic, 40*, 3 - 104.

Dow, D.M., Hart, G.M., & Nance, D.W. (2009). Supervision styles and topics discussed in supervision. *The Clinical Supervisor, 28*, 36 - 46. doi: http://dx.doi.org/10.1080/07325220902832515

Dressel, J.L., Consoli, A.J., Kim, B.S., & Atkinson, D.R. (2007). Successful and unsuccessful multicultural supervisory behaviors: A Delphi poll. *Journal of Multicultural Counseling and Development, 35*, 51 - 64. doi: http://dx.doi.org/10.1002/j.2161-1912.2007.tb00049.x

Efstation, J.F., Patton, M.J., & Kardash, C.M. (1990). Measuring the working alliance in counselor supervision. *Journal of Counseling Psychology, 37*, 322 - 329. doi: http://dx.doi.org/10.1037/0022-0167.37.3.322

Ellis, M.V., Ladany, N., Krengel, M., & Schult, D. (1996). Clinical supervision research from 1981 to 1993: A methodological critique. *Journal of Counseling Psychology*, 43, 35 - 50. doi: http://dx.doi.org/10.1037/0022-0167.43.1.35

Ellis, M.V., Krengel, M., & Beck, M. (2002). Testing self-focused attention theory in clinical supervision: Effects of supervisee anxiety and performance. *Journal of*

Counseling Psychology, 40, 101 - 116. doi: http://dx.doi.org/10.1037/0022-0167.49.1.101

Ellis, M.B., D'Iuso, N., & Ladany, N. (2008). State of the art in the assessment, measurement, and evaluation of clinical supervision. *Psychotherapy supervision: Theory, research, and practice* (2nd ed.), Hoboken, NJ, US: John Wiley & Sons.

Ellis, M.V., Dell, D.M., & Good, G.E. (1988). Counselor trainees' perceptions of supervisor roles: Two studies testing the dimensionality of supervision. *Journal of Counseling Psychology, 35*, 315 - 324. doi: http://dx.doi.org/10.1037/0022-0167.35.3.315

Ellis, P.D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results.* Location: Cambridge University Press.

Ellis, M.V. (1991) Critical incidents in clinical supervision and in supervisor supervision: Assessing supervisory issues. *Journal of Counseling Psychology, 38*, 342 - 349. doi: http://dx.doi.org/10.1037/0022-0167.38.3.342

Ellis, M.V. & Ladany, N. (1997). Inferences concerning supervisees and clients in clinical supervision: An integrative review. In C.E. Watkins, Jr. (Ed.), *Handbook of psychotherapy supervision* (pp. 447 - 507), New York: Wiley.

Enyedy, K.C., Arcinue, F., Puri, N.N., Carter, J.W., Goodyear, R.K., & Getzelman, M.A. (2003). Hindering phenomena in group supervision: Implications for practice. *Professional Psychology: Research and Practice, 34*, 312 - 317. doi: http://dx.doi.org/10.1037/0735-7028.34.3.312

Falender, C.A. & Shafranske, E.P. (2007). Competence in competency-based supervision practice: Construct and application. *Professional Psychology: Research and Practice*, *38*, 232-240. doi: http://dx.doi.org/10.1037/0735-7028.38.3.232

Falender, C.A. & Shafronske, E.P. (2004). *Clinical supervision: A competency based approach.* Washington, DC: American Psychological Association. doi: http://dx.doi.org/10.1037/10806-000

Faul, F., Erdfelder, E., Lang, A.G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175 – 191. doi: http://dx.doi.org/10.3758/BF03193146

Fischetti, B.A. & Crespi, T.D. (1999). Clinical supervision for school psychologists: National practices, trends and future implications. *School Psychology International*, *20*, 278 - 288. doi: http://dx.doi.org/10.1177/0143034399203003

Fernando, D.M. & Hulse-Killacky, D. (2005). The relationship of supervisory styles to satisfaction with supervision and the perceived self-efficacy of master's-level counseling students. *Counselor Education and Supervision, 44*, 293 - 304. doi: http://dx.doi.org/10.1002/j.1556-6978.2005.tb01757.x

Fordham, A.S., May, B., Boyle, M., Bentall, R P., & Slade, P.D. (1990). Good and bad clinicians: Supervisors' judgments of trainees' competence. *British Journal of Clinical Psychology*, *29*, 113 - 114. doi: http://dx.doi.org/10.1111/j.2044-8260.1990.tb00856.x

Fortune, A., & Abramson, J. (1993). Predictors of satisfaction with field practicum among social work students. *The Clinical Supervisor, 11,* 95 - 110. doi: http://dx.doi.org/10.1300/J001v11n01_07

Friedlander, M.L., & Ward, L.G. (1984). Development and validation of the supervisory styles inventory. *Journal of Counseling Psychology, 31,* 541 – 557. doi: http://dx.doi.org/10.1037/0022-0167.31.4.541

Friedlander, M.L., Keller, K.E., Peca-Baker, T.A., & Olk, M.E. (1986). Effects of role conflict on counselor trainees' self-statements, anxiety level, and performance. *Journal of Counseling Psychology, 33,* 73 - 77. doi: http://dx.doi.org/10.1037/0022-0167.33.1.73

Friedlander, M.L. & Snyder, J. (1983). Trainees' expectations for the supervisory process: Testing a developmental model. *Counselor Education and Supervision, 22,* 342 - 348. doi: http://dx.doi.org/10.1002/j.1556-6978.1983.tb01771.x

Gabbay, M.B., Kiemle, G., & Maguire, C. (1999). Clinical supervision for clinical psychologists: Existing provision and unmet needs. *Clinical Psychology & Psychotherapy, 6,* 404 - 412. doi: http://dx.doi.org/10.1002/(SICI)1099-0879(199911)6:5%3C404::AID-CPP209%3E3.0.CO;2-B

Gainor, K.A., & Constantine, M.G. (2002). Multicultural group supervision: A comparison of in-person versus web-based formats. *Professional School Counseling, 6,* 104 - 111.

Gatmon, D., Jackson, D., Koshkarian, L., Martos-Perry, N., Molina, A., Patel, N., & Rodolfa, E. (2001). Exploring ethnic, gender, and sexual orientation variables in supervision: Do they really matter? *Journal of Multicultural Counseling and*

Development, 29, 102 - 113. doi: doi: http://dx.doi.org/10.1002/j.2161-1912.2001.tb00508.x

Geller, J.D., Farber, B.A., & Schaffer, C.E. (2010). Representations of the supervisory dialogue and the development of psychotherapists. *Psychotherapy: Theory, Research, Practice, Training, 47,* 211 - 220. doi: http://dx.doi.org/10.1037/a0019785

Glass, G.V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5,* 3 – 8.

Glass, G.V., McGaw, B., & Smith, M.L. (1981). *Meta-analysis in social research.* Beverly Hills, CA: Sage.

Gloria, A.M., Hird, J.S., & Tao, K.W. (2008). Self-reported multicultural supervision competence of White predoctoral intern supervisors. *Training and Education in Professional Psychology, 2,* 129 - 136. doi: http://dx.doi.org/10.1037/1931-3918.2.3.129

Goodyear, R.K. (1990). Gender configurations in supervisory dyads: Their relation to supervisee influence strategies and to skill evaluations of the supervisee. *The Clinical Supervisor, 8*(2), 67 - 79. doi: http://dx.doi.org/10.1300/J001v08n02_06

Greenspan, R., Hanfling, S., Parker, E., Primm, S., & Waldfogel, D. (1991). Supervision of experienced agency workers: A descriptive study. *The Clinical Supervisor, 9,* 31 - 42. doi: doi: http://dx.doi.org/10.1300/J001v09n02_04

Haase, R.F. (1991). Computational formulas for multivariate strength of association from approximate F and x^2 tests. *Multivariate Behavioral Research, 26,* 227 – 245. doi: http://dx.doi.org/10.1207/s15327906mbr2602_2

Haase, R.F., & Ellis, M.V. (1987). Multivariate analysis of variance. *Journal of Counseling Psychology, 34,* 404 - 413. doi: http://dx.doi.org/10.1037/0022-0167.34.4.404

Haase, R.F., Ellis, M.V., & Ladany, N. (1989). Multiple criteria for evaluating the magnitude of effects. *Journal of Counseling Psychology, 36,* 511 - 516. doi: http://dx.doi.org/10.1037/0022-0167.36.4.511

Haase, R.F., Waechter, D.M., & Solomon, G. S. (1982). How significant is a significant difference? Average effect size of research in counseling psychology. *Journal of Counseling Psychology, 29,* 58 – 65. doi: http://dx.doi.org/10.1037/0022-0167.29.1.58

Hart, G., & Nance, D. (2003). Styles of counselor supervision as perceived by supervisors and supervisees. *Counselor Education and Supervision, 43,* 146 - 159. doi: http://dx.doi.org/10.1002/j.1556-6978.2003.tb01838.x

Haverkamp, B. (1994). Using assessment in counseling supervision: Individual differences in self-monitoring. *Measurement and Evaluation in Counseling and Development, 27,* 316 – 324.

Hawkins, P., & Shohet, R. (2000). *Supervision in the helping professions* (2nd ed.), Cambridge, England: Open University Press.

Hedges, L.V. & Olkin, I. (1983). Clustering estimates of effect magnitude from independent studies. *Psychological Bulletin, 93,* 563 – 573. doi: http://dx.doi.org/10.1037/0033-2909.93.3.563

Hedges, L.V. & Olkin, I. (1984). Nonparametric estimators of effect size in meta-analysis. *Psychological Bulletin, 3,* 573 – 580.

Helms, J.E. (1990). *Black and White racial identity: Theory, research, and practice.* Westport, CT: Greenwood Press.

Henggeler, S.W., Schoenwald, S.K., Liao, J.G., Letourneau, E. J., & Edwards, D. L. (2002). Transporting efficacious treatments to field settings: The link between supervisory practices and therapist fidelity in MST programs. *Journal of Clinical Child and Adolescent Psychology, 31,* 155 - 167. doi: http://dx.doi.org/10.1207/153744202753604449

Henry, P.J., Hart, G.M., & Nance, D.W. (2004). Supervision topics as perceived by supervisors and supervisees. *The Clinical Supervisor*, *23,* 139 - 152. doi: http://dx.doi.org/10.1300/J001v23n02_09

Heppner & Roehlke (1984).Differences among supervisees at different levels of training: Implications for a developmental model of supervision. *Journal of Counseling Psychology, 31,* 76 – 90. doi: http://dx.doi.org/10.1037/0022-0167.31.1.76

Herbert, J.T., Ward, T.J., & Hemlick, L.M. (1995). Confirmatory factor analysis of the Supervisory Style Inventory and the Revised Supervision Questionnaire. *Rehabilitation Counseling Bulletin*, *38,* 334 - 349.

Hill, C.E., Thompson, B.J., & Williams, E.N. (1997). A guide to conducting consensual qualitative research. *The Counseling Psychologist, 25,* 517 - 572. doi: http://dx.doi.org/10.1177/0011000097254001

Hilton, D.B., Russell, R.K., & Salmi, S.W. (1995). The effects of supervisor's race and level of support on perceptions of supervision. *Journal of Counseling and Development, 73,* 57 - 563. doi: http://dx.doi.org/10.1002/j.1556-6676.1995.tb01794.x

Holloway, E.L. (1982). Interactional structure of the supervision interview. *Journal of Counseling Psychology, 29,* 309 - 317. doi: http://dx.doi.org/10.1037/0022-0167.29.3.309

Hora (1957). Contribution to the phenomenology of the supervisory process. American Journal of Psychotherapy, 11, 769 – 773.

Hsu, W. (2007). Effects of solution-focused supervision. *Bulletin of Educational Psychology, 38,* 331 - 354.

Iberg, J.R. (1991). Applying statistical control theory to bring together clinical supervision and psychotherapy research. *Journal of Consulting and Clinical Psychology, 59,* 575 - 586. doi: http://dx.doi.org/10.1037/0022-006X.59.4.575

Inman, A.G. (2006). Supervisor multicultural competence and its relation to supervisory process and outcome. *Journal of Marital and Family Therapy, 32*, 73 - 85. doi: http://dx.doi.org/10.1111/j.1752-0606.2006.tb01589.x

Inman, A.G. & Ladany, N. (2008). Psychotherapy supervision: Theory, research, and practice. In Hess, A.K., Hess, K.D., & Hess, T.H. (Eds.): *Psychotherapy Supervision: Theory, Research, and Practice* (2nd ed.), pp. 500 - 517. Hoboken, NJ: John Wiley & Sons Inc.

Inman , A.G., & Ladany, N. (2008). Developments in counseling skills training and supervision. In Brown, S.D. & Lent, R.W. (Eds): Handbook of Counseling Psychology, pp. 338-354. Hoboken, NJ: John Wiley & Sons, Inc.

Inman, A.G., Schlosser, L.Z., Ladany, N., Howard, E.E., Boyd, D.L., Altman, A.N., & Stein, E.P. (2011). Non-disclosures in doctoral-level advising relationships.

Training and Education in Professional Psychology, 5, 149 - 159. doi: http://dx.doi.org/10.1037/a0024022

Jaspen, N. (1965). The calculation of probabilities corresponding to values of z, t, F, and chi square. *Educational & Psychological Measurement, 25,* 877 - 880. doi: http://dx.doi.org/10.1177/001316446502500319

Johnson, E.A., Stewart, D.W. (2008). Perceived competence in supervisory roles: A social cognitive analysis. *Training and Education in Professional Psychology, 2,* 229 - 236. doi: http://dx.doi.org/10.1037/1931-3918.2.4.229

Kahn, B.B. (1999). Priorities and practices in field supervision of school counseling students. *Professional School Counseling, 3,* 128 - 136.

Kanno, H., Koeske, G.F. (2010). MSW students' satisfaction with their field placement: The role of preparedness and supervision quality. *Journal of Social Work Education, 46,* 23 - 38. doi: http://dx.doi.org/10.5175/JSWE.2010.200800066

Katz, (1985). The sociopolitical nature of counseling. The Counseling Psychologist, 13, 615 – 624.

Kauderer, S. & Herron, W.G. (1990).The supervisory relationship in psychotherapy over time. *Psychological Reports, 67,* 471 - 480.

Keller, J.F., Protinsky, H.O., Lichtman, M., & Allen, K. (1996). The process of clinical supervision: Direct observation research. *The Clinical Supervisor, 14,* 1996, 51 - 63. doi: http://dx.doi.org/10.1300/J001v14n01_04

Kennedy, J. J. (1970). The eta coefficient in complex ANOVA designs. *Educational and Psychological Measurement, 30,* 885 - 889. doi: http://dx.doi.org/10.1177/001316447003000409

Kivlighan, D.M., Angelone, E.O., & Swafford, K.G. (1991). Live supervision in individual psychotherapy: Effects on therapist's intention use and client's evaluation of session effect and working alliance. *Professional Psychology: Research and Practice, 22,* 489 - 495. doi: http://dx.doi.org/10.1037/0735-7028.22.6.489

Knight, C. (2001). The process of field instruction: BSW and MSW students' views of effective field supervision. *Journal of Social Work Education, 37,* 357 - 379.

Krause, A.A., & Allen, G.J. (1988). Perceptions of counselor supervision: An examination of Stoltenberg's model from the perspectives of supervisor and supervisee. *Journal of Counseling Psychology, 35,* 77 - 80. doi: http://dx.doi.org/10.1037/0022-0167.35.1.77

Ladany, N., & Friedlander, M.L. (1995).The relationship between the supervisory working alliance and trainees' experience of role conflict and role ambiguity. *Counselor Education and Supervision, 34,* 220 - 231. doi: http://dx.doi.org/10.1002/j.1556-6978.1995.tb00244.x

Ladany, N., Hill, C.E., Corbett, M.M., & Nutt, E.A. (1996). Nature, extent, and importance of what psychotherapy trainees do not disclose to their supervisors. *Journal of Counseling Psychology, 43,* 10 - 24. doi: http://dx.doi.org/10.1037/0022-0167.43.1.10

Ladany, N. & Ellis, M.V. (1997). Inferences concerning supervisees and clients in clinical supervision: An integrative review. In C. Edward, Jr. (Ed.), *Handbook of Psychotherapy Supervision,* (pp. 447 – 507). Hoboken, NJ.: John Wiley & Sons.

Ladany, N., Inman, A.G., Constantine, M.G., & Hofheinz, E.W. (1997a).Supervisee multicultural case conceptualization ability and self-reported multicultural competence as functions of supervisee racial identity and supervisor focus. *Journal of Counseling Psychology*, *44*, 284 - 293. doi: http://dx.doi.org/10.1037/0022-0167.44.3.284

Ladany, N., Brittan-Powell, C.S., & Pannu, R.K. (1997b).The influence of supervisory racial identity interaction and racial matching on the supervisory working alliance and supervisee multicultural competence. *Counselor Education and Supervision, 36*, 284 - 304. doi: http://dx.doi.org/10.1002/j.1556-6978.1997.tb00396.x

Ladany, N., Ellis, M.V., & Friedlander, M.L. (1999).The supervisory working alliance, trainee self-efficacy, and satisfaction. *Journal of Counseling & Development*, *77*, 447 - 455. doi: http://dx.doi.org/10.1002/j.1556-6676.1999.tb02472.x

Ladany, N. & Lehrman-Waterman, D.E. (1999).The content and frequency of supervisor self-disclosures and their relationship to supervisor style and the supervisory working alliance. *Counselor Education and Supervision, 38*, 143 - 160. doi: http://dx.doi.org/10.1002/j.1556-6978.1999.tb00567.x

Ladany, N., Lehrman-Waterman, D., Molinaro, M., & Wolgast, B. (1999).Psychotherapy supervisor ethical practices: Adherence to guidelines, the supervisory working alliance, and supervisee satisfaction. *The Counseling Psychologist*, *27*, 443 - 475. doi: http://dx.doi.org/10.1177/0011000099273008

Ladany, N., Walker, J.A., & Melincoff, D.S. (2001). Supervisory style: Its relation to the supervisory working alliance and supervisor self-disclosure. *Counselor Education*

and Supervision, 40, 263 - 275. doi: http://dx.doi.org/10.1002/j.1556-6978.2001.tb01259.x

Ladany, N., Marotta, S., & Muse-Burke, J.L. (2001). Counselor experience related to complexity of case conceptualization and supervision preference. *Counselor Education and Supervision, 40,* 203 – 219. doi: http://dx.doi.org/10.1002/j.1556-6978.2001.tb01253.x

Ladany, N. & Muse-Burke, J.L. (2001). Understanding and conducting supervision research. In L.J. Bradley & N. Ladany, (Eds.), *Counselor Supervision: Principles, process, & practice* (3rd ed., pp. 304 - 329). Philadelphia: Brunner-Routledge.

Ladany, N. (2005). Conducting effective clinical supervision. In G. P. Koocher, J. C. Norcross, & S. S. Hill (Eds.), *Psychologists' desk reference* (2nd ed., pp. 682 - 685). New York: Oxford University Press.

Ladany, N., Friedlander, M.L., & Nelson, M.L. (2005). Working through countertransference when super-vision is needed. In Ladany, N., Friedlander, M.L., & Nelson, M.L. (Eds.), *Critical events in psychotherapy supervision: An interpersonal approach.* (pp. 99 - 126). Washington, DC: APA. doi: http://dx.doi.org/10.1037/10958-005

Ladany, N., Friedlander, M.L., & Nelson, M.L. (2008). *Critical events in psychotherapy supervision: An interpersonal approach.* Washington, DC: American Psychological Association. doi: http://dx.doi.org/10.1037/10958-005

Lafromboise, T.D., Coleman, H.L., & Hernandez, A. (1991). Development and factor structure of the cross-cultural counseling inventory—revised. Professional

Psychology: Research and Practice, 22, 380 – 388. doi: http://dx.doi.org/10.1037/0735-7028.22.5.380

Lamal P.A. (1990).On the importance of replication. *Journal of Social Behavior & Personality, 5,* 31 - 35.

Larson, L. M., Suzuki, L. A., Gillespie, K. N., Potenza, M. T., Bechtel, M. A., & Toulouse, A. L. (1992) Development and validation of the Counseling Self-Estimate Inventory. *Journal of Counseling Psychology, 39,* 105 - 120. doi: http://dx.doi.org/10.1037/0022-0167.39.1.105

Lazar, A., & Mosek, A. (1993).The influence of the field instructor-student relationship on evaluation of students' practice. *The Clinical Supervisor, 11,* 111 - 120. doi: http://dx.doi.org/10.1300/J001v11n01_08

Lee, R.E., Nichols, D.P., Nichols, W.C., & Odom, T. (2004).Trends in family therapy supervision: The past 25 years and into the future. *Journal of Marital and Family Therapy, 30,* 61 - 69. doi: http://dx.doi.org/10.1111/j.1752-0606.2004.tb01222.x

Lehrman-Waterman, D., & Ladany, N. (2001). Development and validation of the Evaluation Process Within Supervision Inventory. *Journal of Counseling Psychology, 48,* 168 - 177. doi: http://dx.doi.org/10.1037/0022-0167.48.2.168

Lindsay, R.M., & Ehrenberg, S.C. (1993). The design of replicated studies. *The American Statistician, 47,* 217 – 228. doi: http://dx.doi.org/10.2307/2684982

Lipsey, M. W. (1990). Design sensitivity: Statistical power for experimental research. Newbury Park, CA: Sage.

Lent, R.W., Cinamon, R.G., Bryan, N.A., Jezzi, M.M., Martin, H.M., & Lim, R. (2009). Perceived sources of change in trainees' self-efficacy beliefs. *Psychotherapy: Theory, Research, Practice, Training, 46,* 317 - 327. doi: http://dx.doi.org/10.1037/a0017029

Ligiéro, D.P. & Gelso, C.J. (2002). Countertransference, attachment, and the working alliance: The therapist's contribution. *Psychotherapy: Theory, Research, Practice, Training, 39,* 3 - 11.

Lochner, B.T. & Melchert, T.P. (1997). Relationship of cognitive style and theoretical orientation to psychology interns' preferences for supervision. *Journal of Counseling Psychology, 44,* 256 -260. doi: http://dx.doi.org/10.1037/0022-0167.44.2.256

Locke, L.D. & McCollum, E.E. (2001). Clients' views of live supervision and satisfaction with therapy. *Journal of Marital and Family Therapy, 27,* 129 - 133. doi: http://dx.doi.org/10.1111/j.1752-0606.2001.tb01146.x

Loganbill, C., Hardy, E., & Delworth, U. (1982). Supervision: A conceptual model. *The Counseling Psychologist, 10,* 3 - 42. doi: http://dx.doi.org/10.1177/0011000082101002

Long, J.K., Lawless, J.J., & Dotson, D.R. (1996). Supervisory Styles Index: Examining supervisees' perceptions of supervisory style. *Contemporary Family Therapy: An International Journal, 18,* 589 - 606. doi: http://dx.doi.org/10.1007/BF02195719

Lovell, C. (1999). Supervisee cognitive complexity and the integrated developmental model. *The Clinical Supervisor, 18,* 191 - 201. doi: http://dx.doi.org/10.1300/J001v18n01_12

Mahrer, A.R. (1988). Discovery-oriented psychotherapy research: Rationale, aims, and methods. *American Psychologist, 43,* 694 - 702. doi: http://dx.doi.org/10.1037/0003-066X.43.9.694

Majcher, J.A. & Daniluk, J.C. (2009). The process of becoming a supervisor for students in a doctoral supervision training course. *Training and Education in Professional Psychology, 3,* 63 - 71. doi: http://dx.doi.org/10.1037/a0014470

Mallinckrodt, B., & Nelson, M.L. (1991).Counselor training level and the formation of the psychotherapeutic working alliance. *Journal of Counseling Psychology, 38,* 133 - 138. doi: http://dx.doi.org/10.1037/0022-0167.38.2.133

McCarthy, P., Kulakowski, D., Kenfield, J.A. (1994).Clinical supervision practices of licensed psychologists. *Professional Psychology: Research and Practice, 25,* 177 - 181. doi: http://dx.doi.org/10.1037/0735-7028.25.2.177

McCurdy, K.G. & Owen, J.J. (2008). Using sandtray in Adlerian-based clinical supervision: An initial empirical analysis. *The Journal of Individual Psychology, 64,* 96 - 112.

McHenry & Freeman (1997). A multimethod-multitrait validation study of Supervisor Emphasis Rating Form—Revised. *The Clinical Supervisor, 15,* 1 – 17.

McMahon, M. & Simons, R. (2004). Supervision training for professional counselors: An exploratory study. *Counselor Education and Supervision, 43,* 301 - 309. doi: http://dx.doi.org/10.1002/j.1556-6978.2004.tb01854.x

McNeill, B.W., Stoltenberg, C.D., & Romans, J.S. (1992). The integrated developmental model of supervision: Scale development and validation procedures. *Professional*

Psychology: Research and Practice, 23, 504 - 508. doi: http://dx.doi.org/10.1037/0735-7028.23.6.504

McNeill, B.W. & Worthen (1989). The parallel process in psychotherapy supervision. *Professional Psychology: Research and Practice, 20*, 329 – 333. doi: http://dx.doi.org/10.1037/0735-7028.20.5.329

Meehl, P.E. (1990).Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, 195 - 244.

Meier, S.T. (2000). Treatment sensitivity of the PE form of the Social Skills Rating Scales: Implications for test construction procedures. *Measurement and Evaluation in Counseling and Development, 33*, 144 – 156.

Miller, M.M.& Ivey, D.C. (2006). Spirituality, gender, and supervisory style in supervision. Contemporary Family Therapy: An International Journal, 28, 323 – 337. doi: http://dx.doi.org/10.1007/s10591-006-9012-0

Miller, G.M. & Larrabee, M.J. (1995). Sexual intimacy in counselor education and supervision: A national survey. *Counselor Education and Supervision, 34*, 332 - 343. doi: http://dx.doi.org/10.1002/j.1556-6978.1995.tb00199.x

Miller, M.M., Korinek, A.W. & Ivey, D.C. (2006). Integrating spirituality into training: The Spiritual Issues in Supervision Scale. *American Journal of Family Therapy, 34*, 355 - 372. doi: http://dx.doi.org/10.1080/01926180600553811

Milne, D., Aylott, H., Fitzpatrick, H., & Ellis, M.V. (2008).How does clinical supervision work? Using a "best evidence synthesis" approach to construct a basic model of supervision. *The Clinical Supervisor, 27*, 170 - 190. doi: http://dx.doi.org/10.1080/07325220802487915

Milne & James (2002). The observed impact of training on competence in clinical supervision. *British Journal of Clinical Psychology, 41,* 55 – 72. doi: http://dx.doi.org/10.1348/014466502163796

Milne, D. (2010). Can we enhance the training of clinical supervisors? A national pilot study of an evidence-based approach. *Clinical Psychology & Psychotherapy, 17,* 321 - 328.

Mori, Y., Inman, A.G., & Caskie, G.I. (2009). Supervising international students: Relationship between acculturation, supervisor multicultural competence, cultural discussions, and supervision satisfaction. *Training and Education in Professional Psychology, 3,* 10 – 18. doi: http://dx.doi.org/10.1037/a0013072

Moskowitz, & Rupert, (1983). Conflict resolution within the supervisory relationship. Professional Psychology: Research and practice, 14, 632 – 641.

Navin, S., Beamish, P., & Johanson, G. (1995). Ethical practices of field-based mental health counselor supervisors. *Journal of Mental Health Counseling, 17,* 243 - 253.

Nelson, M.L., & Holloway, E.L. (1990).Relation of gender to power and involvement in supervision. *Journal of Counseling Psychology, 37,* 473 - 481. doi: http://dx.doi.org/10.1037/0022-0167.37.4.473

Nelson, M.L. & Friedlander, M.L. (2001).A close look at conflictual supervisory relationships: The trainee's perspective. *Journal of Counseling Psychology, 48,* 384-395. doi: http://dx.doi.org/10.1037/0022-0167.48.4.384

Nelson, N., Rosenthal, R., & Rosnow, R.L. (1986).Interpretation of significance levels and effect sizes by psychological researchers. *American Psychologist, 41,* 1299 - 1301. doi: http://dx.doi.org/10.1037/0003-066X.41.11.1299

Nilsson, Johanna E., Duan, Changming (2007).Experiences of prejudice, role difficulties, and counseling self-efficacy among U.S. racial and ethnic minority supervisees working with White supervisors. *Journal of Multicultural Counseling and Development, 35,* 219 - 229. doi: http://dx.doi.org/10.1002/j.2161-1912.2007.tb00062.x

Nilsson, J.E. & Anderson, M.Z. (2004). Supervising International Students: The Role of Acculturation, Role Ambiguity, and Multicultural Discussions. *Professional Psychology: Research and Practice, 35,* 306 - 312. doi: http://dx.doi.org/10.1037/0735-7028.35.3.306

Nilsson, J.E., & Dodds, A.K. (2006). A pilot phase in the development of the international student supervision scale. *Journal of Multicultural Counseling and Development, 34,* 50 - 62. doi: http://dx.doi.org/10.1002/j.2161-1912.2006.tb00026.x

Nyman, S.J., Nafziger, M.A., & Smith, T.B. (2010).Client outcomes across counselor training level within a multitiered supervision model. *Journal of Counseling & Development, 88,* 204 -209. doi: http://dx.doi.org/10.1002/j.1556-6678.2010.tb00010.x

Olk, M.E., & Friedlander, M.L. (1992). Trainees' experiences of role conflict and role ambiguity in supervisory relationships. *Journal of Counseling Psychology, 39,* 389 - 397. doi: http://dx.doi.org/10.1037/0022-0167.39.3.389

Ögren, M., Jonsson, C., & Sundin, E.C. (2005).Group supervision in psychotherapy: The relationship between focus, group climate, and perceived attained skill. *Journal of Clinical Psychology, 61,* 373 - 388. doi: http://dx.doi.org/10.1002/jclp.20056

Page, B.J., Pietrzak, D.R., & Sutton, J.M., Jr. (2001).National survey of school counselor supervision. *Counselor Education and Supervision, 41*, 142 - 150. doi: http://dx.doi.org/10.1002/j.1556-6978.2001.tb01278.x

Peace, S.D. & Sprinthall, N.A. (1998). Training school counselors to supervise beginning counselors: Theory, research, and practice. *Professional School Counseling, 1,* 2 - 8.

Peleg-Oren, N., Macgowan, M. J., & Even-Zahav, R. (2007). Field instructors' commitment to Student Supervision: Testing the investment model. *Social Work Education, 26*, 684 - 696. doi: http://dx.doi.org/10.1080/02615470601129875

Ponterotto , J.G. Gretchen, D. Utsey, S.O. Rieger, B.P. & Austin, R. (2002). A revision of the Multicultural Counseling Awareness Scale (MCAS). *Journal of Multicultural Counseling and Development, 30*, 153-180. doi: http://dx.doi.org/10.1002/j.2161-1912.2002.tb00489.x

Pope-Davis, D.B., Reynolds, A.L., Dings, J.G., & Otavi, T.M. (1994).Multicultural competencies of doctoral interns at university counseling centers: An exploratory investigation. *Professional Psychology: Research and Practice, 25,* 466 - 470. doi: http://dx.doi.org/10.1037/0735-7028.25.4.466

Prieto, L. R. (1996). Group supervision: Still widely practiced but poorly understood. *Counselor Education and Supervision, 35*, 295 -3 07.

Rabinowitz, F. E., Heppner, P. P., & Roehlke, H. J. (1986). Descriptive study of process and outcome variables of supervision over time. *Journal of Counseling Psychology, 33,* 292 - 300. doi: http://dx.doi.org/10.1037/0022-0167.33.3.292

Raichelson, S.H., Herron, W.G., Primavea, L.H., & Ramirez, S.M. (1997). Incidence and effects of parallel process in psychotherapy supervision. *The Clinical Supervisor, 15,* 37 – 48. doi: http://dx.doi.org/10.1300/J001v15n02_03

Ramos-Sánchez, L., Esnil, E., Goodwin, A., Riggs, S., Touster, L.O., Wright, L.K., & Ratanasiripon, E.R. (2002). Negative supervisory events: Effects on supervision and supervisory alliance. *Professional Psychology: Research and Practice, 33,* 197 - 202. doi: http://dx.doi.org/10.1037/0735-7028.33.2.197

Ramirez, N. (2003).Views towards organizational arrangements for ethnic-sensitive supervision in clinical settings serving Latino persons. *Journal of Ethnic & Cultural Diversity in Social Work: Innovation in Theory, Research & Practice, 12,* 1 - 18.

Ray, D., & Altekruse, M. (2000). Effectiveness of group supervision versus combined group and individual supervision. *Counselor Education and Supervision, 40,* 19 - 30. doi: http://dx.doi.org/10.1002/j.1556-6978.2000.tb01796.x

Ramos-Sanchez, L., Esnil, E., Goodwin, A., Riggs, S., Touster, L.O., Wright, L.K., Ratanasiripong, P., & Rodolfa, E. (2002). Negative supervisory events: Effects on supervision and supervisory alliance. *Professional Psychology: Research and Practice, 33,* 197 - 202. doi: http://dx.doi.org/10.1037/0735-7028.33.2.197

Reese, R.J., Usher, E.L., Bowman, D.C., Norsworthy, L.A., Halstead, J.L., Rowlands, S.R., & Chisholm, R.R. (2009).Using client feedback in psychotherapy training:

An analysis of its influence on supervision and counselor self-efficacy. *Training and Education in Professional Psychology, 3,* 157 - 168. doi: http://dx.doi.org/10.1037/a0015673

Riggs, S.A. & Bretz, K.M. (2006). Attachment processes in the supervisory relationship: An exploratory investigation. *Professional Psychology: Research and Practice, 37,* 558 - 566. doi: http://dx.doi.org/10.1037/0735-7028.37.5.558

Riva, M.T., Cornish, J.A., & Erickson, J.A. (1995). Group supervision practices at psychology predoctoral internship programs: A national survey. *Professional Psychology: Research and Practice, 26,* 523 - 525. doi: http://dx.doi.org/10.1037/0735-7028.26.5.523

Rodway, M., & Rogers, G. (1993).A comparison of the academic and articulated approaches to graduate field education. *The Clinical Supervisor, 11,* 37 - 54. http://dx.doi.org/10.1300/J001v11n02_04

Romans, J.S. C., Boswell, D.L., Carlozzi, A.F., & Ferguson, D.B. (1995).Training and supervision practices in clinical, counseling, and school psychology programs. *Professional Psychology: Research and Practice, 26,* 407 – 412. doi: http://dx.doi.org/10.1037/0735-7028.26.4.407

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 86,* 638 - 641. doi: http://dx.doi.org/10.1037/0033-2909.86.3.638

Rosenthal, R., & Rubin, D. (1982). Comparing effect sizes of independent studies. Psychological Bulletin, 92, 500 – 504. doi: http://dx.doi.org/10.1037/0033-2909.92.2.500

Rosnow, R.L., & Rosenthal, R. (1996). Computing contrasts, effect sizes, and counternulls on other people's published data: General procedures for research consumers. *Psychological Methods, 4*, 331 – 340. doi: http://dx.doi.org/10.1037/1082-989X.1.4.331

Rossi, J.R. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology, 58,* 646 – 656. doi: http://dx.doi.org/10.1037/0022-006X.58.5.646

Russell, R.K., Crimmings, A.M., & Lent, R.W. (1984). Counselor training and supervision: Theory and research. In S. D. Brown & R. W. Lent (Eds.), *Handbook of counseling psychology* (pp. 625-681). New York: Wiley.

Russell, C.S., Dupree, W.J., Beggs, M.A., Peterson C.M., & Anderson, M.P. (2007).Responding to remediation and gatekeeping challenges in supervision. *Journal of Marital and Family Therapy, 33*, 227 - 244. doi: http://dx.doi.org/10.1111/j.1752-0606.2007.00018.x

Scott, K.J., Ingram, K.M., Vitanza, S.A., & Smith, N.G. (2000). Training in supervision: A survey of current practices. *The Counseling Psychologist*, 28, 403 - 422. doi: http://dx.doi.org/10.1177/0011000000283007

Schact, A.J., Howe, H.E., & Berman, J.J. (1988).A short form of the Barrett-Lennard Relationship *Inventory for supervisory relationships. Psychological Reports, 63,* 699 – 706. doi: http://dx.doi.org/10.2466/pr0.1988.63.3.699

Schectman, Z. & Wirzberger, A. (1999). Needs and preferred style of supervision among Israeli school counselors at different stages of professional development. Journal

of Counseling & Development, 77, 456 – 464. doi: http://dx.doi.org/10.1002/j.1556-6676.1999.tb02473.x

Schoenwald, S.K., Sheidow, A.J., & Chapman, J. E. (2009). Clinical supervision in treatment transport: Effects on adherence and outcomes. *Journal of Consulting and Clinical Psychology, 77*, 410 - 421. doi: http://dx.doi.org/10.1037/a0013788

Schroeder, M., Andrews, J. W. & Hindes, Y.L. (2009). Cross-racial supervision: Critical issues in the supervisory relationship. *Canadian Journal of Counselling, 43*, 295 - 310.

Schultz, J.C., Ososkie, J.N., Fried, J.H., Nelson, R.E., & Bardos, A.N. (2002). Clinical supervision in public rehabilitation counseling settings. *Rehabilitation Counseling Bulletin, 45*, 213 - 222. doi: http://dx.doi.org/10.1177/00343552020450040401

Sells, J. N., Goodyear, R. K., Lichtenberg, J. W., & Polkinghorne, D. E. (1997). Relationship of supervisor and trainee gender to in-session verbal behavior and ratings of trainee skills. *Journal of Counseling Psychology*, 44, 406 - 412. doi: http://dx.doi.org/10.1037/0022-0167.44.4.406

Shadish, W.R., Cook, T.D., and Campbell D.T. (2002).*Experimental and quasi-experimental designs for generalized causal inference.* Boston, MA: Houghton, Mifflin and Company.

Shafranske, E. & Falender, C.A. (2008). Supervision addressing personal factors and countertransference. In Falender, C.A. & Shafranske, E.P. (Eds.) Casebook for clinical supervision: A competency-based approach, (pp. 97-120). Washington, DC: APA. doi: http://dx.doi.org/10.1037/11792-005

Shanfield, S.B. (1993). What do excellent psychotherapy supervisors do? *The American Journal of Psychiatry, 150,* 1081 - 1084.

Shechtman, Z. & Wirzberger, A. (1999). Needs and preferred style of supervision among Israeli school counselors at different stages of professional development. *Journal of Counseling & Development, 77,* 456 - 464. doi: http://dx.doi.org/10.1002/j.1556-6676.1999.tb02473.x

Shulman, L. (2005). The clinical supervisor-practitioner working alliance: A parallel process. *Clinical Supervisor*, *24*, 23 - 47. doi: http://dx.doi.org/10.1300/J001v24n01_03

Smaby, M.H., Smith, M.R., & Maddux, C.D. (2002). Counselor Education and Supervision: Quality of editorial board members' evaluations of manuscripts. *Counselor Education and Supervision*, *41*, 259 - 267. doi: http://dx.doi.org/10.1002/j.1556-6978.2002.tb01289.x

Smith, M.L. & Glass, G.V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist*, 752 – 760. doi: http://dx.doi.org/10.1037/0003-066X.32.9.752

Sodowsky, G.R., Taffe, R.C., Gutkin, T.B., & Wise, S.L. (1994). Development of the multicultural counseling inventory: A self-report measure of multicultural competencies. *Journal of Counseling Psychology, 41,* 137 - 148. doi: http://dx.doi.org/10.1037/0022-0167.41.2.137

Sterner, W.R. (2009). Influence of the supervisory working alliance on supervisee work satisfaction and work-related stress. *Journal of Mental Health Counseling, 31,* 249 - 263.

Stemler, Steven E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, 9. Retrieved from http://PAREonline.net/getvn.

Stevens, D. T., Goodyear, R. K., & Robertson, P. (1998). Supervisor development: An exploratory study in changes in stance and emphasis. *Clinical Supervisor, 16*, 73 – 88. doi: http://dx.doi.org/10.1300/J001v16n02_05

Stoltenberg, C.D., & Delworth, U. (1987). *Supervising counselors and therapists: A developmental approach*. San Francisco: Jossey-Bass.

Strong, S.R. (1968). Counseling: An interpersonal influence process. *Journal of Counseling Psychology, 15*, 215 - 224.

Studer, J.R., Oberman, A. (2006). The use of the ASCA National Model® in supervision. *Professional School Counseling, 10*, 82 - 87.

Sue, D.W., Arrendondo, P. & McDavis, R.J. (1992). Multicultural counseling competencies and standards: A call to the profession. *Journal of Counseling & Development, 70*, 477 – 486. doi: http://dx.doi.org/10.1002/j.1556-6676.1992.tb01642.x

Sue, D.W., & Sue, D. (1999). *Counseling the culturally different: Theory and practice (3rd ed.).* Hoboken, NJ: John Wiley & Sons.

Sumerel, M.B., & Borders, L.D. (1996). Addressing personal issues in supervision: Impact of counselors' experience level on various aspects of the supervisory relationship. *Counselor Education and Supervision, 35*, 268 - 286.

Sundin, E.C., Ögren, M.L., & Boëthius, S.B. (2008). Supervisor trainees' and their supervisors' perceptions of attainment of knowledge and skills: An empirical

evaluation of a psychotherapy supervisor training programme. *British Journal of Clinical Psychology, 47,* 381 – 396. doi: http://dx.doi.org/10.1348/014466508X304414

Szymanski, D.M. (2005). Feminist identity and theories as correlates of feminist supervision practices. *The Counseling Psychologist, 33,* 729 - 747. doi: http://dx.doi.org/10.1177/0011000005278408

Thielsen, V.A. & Leahy, M.J. (2001). Essential knowledge and skills for effective clinical supervision in rehabilitation counseling. *Rehabilitation Counseling Bulletin, 44,* 196 – 208. doi: http://dx.doi.org/10.1177/003435520104400402

Tinsley, H.E.A., & Tinsley, D. J. (1987). Uses of factor analysis in counseling psychology research. *Journal of Counseling Psychology, 34,* 414 - 424. doi: http://dx.doi.org/10.1037/0022-0167.34.4.414

Tobin, D.J. & McCurdy, K.G. (2006). Adlerian-focused supervision for countertransference work with counselors-in-training. *Journal of Individual Psychology, 62,* 154 - 167.

Toporek, R.L., Ortega-Villalobos, L., & Pope-Davis, D. B. (2004). Critical incidents in multicultural supervision: Exploring supervisees' and supervisors experiences. *Journal of Multicultural Counseling and Development, 32,* 66 - 83. doi: http://dx.doi.org/10.1002/j.2161-1912.2004.tb00362.x

Tromski-Klingshirn, D.M. & Davis, T.E. (2007). Supervisees' perceptions of their clinical supervision: A study of the dual role of clinical and administrative supervisor. *Counselor Education and Supervision, 46,* 294 - 304. doi: http://dx.doi.org/10.1002/j.1556-6978.2007.tb00033.x

Tyler, J.D., Sloan, L.L., & King, A.R. (2000). Psychotherapy supervision practices of academic faculty: A national survey. *Psychotherapy: Theory, Research, Practice, Training, 37*, 98 - 101. doi: http://dx.doi.org/10.1037/h0087750

Tryon, G.S. (1996). Supervisee development during the practicum year. *Counselor Education and Supervision*, 35, 287 - 294. doi: http://dx.doi.org/10.1002/j.1556-6978.1996.tb01929.x

Utsey, S.O. & Gernat, C.A. (2002).White racial identity attitudes and the ego defense mechanisms used by White counselor trainees in racially provocative counseling situations. *Journal of Counseling & Development, 80*, 475 - 483. doi: http://dx.doi.org/10.1177/0011000004269058

Utsey, S.O, Gernat, C.A., & Hammar, L. (2005). Examining white counselor trainees' reactions to racial issues in counseling and supervision dyads. *The Counseling Psychologist, 33,* 449 - 478. doi: http://dx.doi.org/10.1177/0011000004269058

Vacarro, N., & Lambie, G.W. (2007).Computer-based counselor-in-training supervision: Ethical and practical implications for counselor educators and supervisor. *Counselor Education and Supervision, 47*, 46 - 57. doi: http://dx.doi.org/10.1002/j.1556-6978.2007.tb00037.x

Vespia, K.M., Heckman-Stone, C., & Delworth, U. (2002). Describing and facilitating effective supervision behavior in counseling trainees. *Psychotherapy: Theory, Research, Practice, Training, 39*, 56 - 65. doi: http://dx.doi.org/10.1037/0033-3204.39.1.56

Wampold, B.E., Davis, B., & Good, R.H. (1990). Hypothesis validity of clinical research. *Journal of Consulting and Clinical Psychology, 58,* 360 – 367. doi: http://dx.doi.org/10.1037/0022-006X.58.3.360

Walker, J. A., Ladany, N., & Pate-Carolan, L., M. (2007). Gender-related events in psychotherapy supervision: Female trainee perspectives. *Counseling & Psychotherapy Research, 7,* 12 - 18. doi: http://dx.doi.org/10.1080/14733140601140881

Watkins, C.E. (1998). Psychotherapy supervision in the 21st century: Some pressing needs and impressing possibilities. *Journal of Psychotherapy Practice and Research, 7,* 93 - 101.

Wester, S.R., Vogel, D.L., & Archer, J., Jr. (2004). Male Restricted Emotionality and Counseling Supervision. *Journal of Counseling & Development, 82*, 91 - 98. doi: http://dx.doi.org/10.1002/j.1556-6678.2004.tb00289.x

Westefeld, J.S. (2009). Supervision of psychotherapy: Models, issues, and recommendations. *The Counseling Psychologist, 37*, 296 - 316. doi: http://dx.doi.org/10.1177/0011000008316657

White, J.H. & Rudolph, B.A. (2000). A pilot investigation of the reliability and validity of the Group Supervisory Behavior Scale (GSBS). *The Clinical Supervisor, 19,* 161 171. doi: http://dx.doi.org/10.1300/J001v19n02_09

White, M.B. & Russell, C.S. (1995).The essential elements of supervisory systems: A modified Delphi study. *Journal of Marital and Family Therapy, 21,* 33 - 53. doi: http://dx.doi.org/10.1111/j.1752-0606.1995.tb00137.x

White, M.B. & Russell, C.S. (1997). Examining the multifaceted notion of isomorphism in marriage and family therapy supervision: A quest for conceptual clarity. *Journal of Marital and Family Therapy, 23*, 315 - 333. doi: http://dx.doi.org/10.1111/j.1752-0606.1997.tb01040.x

Wilbur, M.P., Roberts-Wilbur, J., Hart, G.M., & Morris, J.R. (1994). Structured Group Supervision (SGS): A pilot study. *Counselor Education and Supervision, 33*, 262 - 279. doi: http://dx.doi.org/10.1002/j.1556-6978.1994.tb00293.x

Winter, M., & Holloway, E.L. (1991). Relation of trainee experience, conceptual level, and supervisor approach to selection of audiotaped counseling passages. *The Clinical Supervisor, 9*, 87 - 103. doi: http://dx.doi.org/10.1300/J001v09n02_09

Worthen & McNeill (1996). A phenomenological investigation of "good" supervision events. *Journal of Counseling Psychology, 43*, 25 - 34. doi: http://dx.doi.org/10.1037/0022-0167.43.1.25

Yourman, D.B. & Farber, B.A. (1996). Nondisclosure and distortion in psychotherapy supervision. *Psychotherapy: Theory, Research, Practice, Training, 33*, 567 - 575. doi: http://dx.doi.org/10.1037/0033-3204.33.4.567

Zabock, G., Drews, M., Bodansky, A., Dahme, B. (2009). The evaluation of supervision: Construction of brief questionnaires for the supervisor and the supervisee. *Psychotherapy Research, 19*, 194 - 204. doi: http://dx.doi.org/10.1080/10503300802688478

Zaslavsky, J., Nunes, N.L.T., Eizirik, C., & Nurse, G. (2005). Approaching countertransference in psychoanalytic supervision: A qualitative investigation.

The International Journal of Psychoanalysis, *86*, 1099 - 1131. doi:

http://dx.doi.org/10.1516/4GM8-5NUA-YKDG-9CWV

Table 1

Journals reviewed in Ellis et al. (1996) and in the current study

Ellis et al. (1996)	*American Journal of Psychiatry, Clinical Supervisor, Counselor Education and Supervision, Journal of Counseling Psychology, Professional Psychology: Research and Practice, Psychological Reports*
Current Study	*American Journal of Family Therapy, American Psychologist, Bulletin of Educational Psychology, Canadian Journal of Counselling, Clinical Psychology & Psychotherapy, Counselor Education and Supervision, Group Dynamics: Theory, Research, and Practice, Journal of Addictions & Offender Counseling, Journal of Clinical Child and Adolescent Psychology, Journal of Consulting and Clinical Psychology, Journal of Counseling Psychology, Journal of Counseling Development, Journal of Ethnic & Cultural Diversity in Social Work: Innovation in theory, Research, and Practice, Journal of Individual Psychotherapy, Journal of Marital and Family Therapy, Journal of Mental Health Counseling, Journal of Multicultural Counseling and Development, Journal of Social Behavior & Personality, Journal of Social Work Education, Measurement and Evaluation in Counseling and Development, Professional School Counseling, Professional Psychology Research and Practice,*

Psychotherapy: Theory, Research, Practice, Training, Rehabilitation Counseling Bulletin, Social Work Education, The Clinical Supervisor, The Counseling Psychologist, Training and Education in Professional Psychology

Table 2

Extension Articles: Authors, Publication Dates, Journal Name, and Number of Participants

Author(s)	Publication year	Journal	# participants (study 1, study 2)
1. Anderson et al.	2000	*Journal of Marital and Family Therapy*	160
2. Bhat & Davis	2007	*Journal of Multicultural Counseling and Development*	119
3. Birk & Mahalik	1996	*The Clinical Supervisor*	29
4. Borders et al.	1995	*Counselor Education and Supervision*	190
5. Butler & Constantine	2006	*Professional School Counseling*	48
6. Callahan et al.	2009	*Training and Education in Professional Psychology*	11
7. Carlozzi et al.	1997	*The Clinical Supervisor*	48
8. Cashwell & Dooley	2001	*The Clinical Supervisor*	33
9. Chagnon & Russell	1995	*Journal of Counseling & Development*	48
10. Clingerman & Bernard	2004	*Counselor Education and Supervision*	19
11. Coll	1995	*Counselor Education and Supervision*	60
12. Coleman et al.	2009	*Group Dynamics: Theory, Research, and Practice*	214 , 49
13. Constantine et al.	2005	*Journal of Counseling Psychology*	108
14. Culbreth	1999	*Journal of Addictions & Offender Counseling*	547
15. Dow et al.	2009	*The Clinical Supervisor*	247
16. Ellis et al.	2002	*Journal of Counseling Psychology*	71, 80
17. Fernando & Hulse-Killacky	2005	*Counselor Education and Supervision*	82
18. Gatmon et al.	2001	*Journal of Multicultural Counseling and Development*	289
19. Geller et al.	2010	*Psychotherapy: Theory, Research, Practice Training*	115
20. Gloria et al.	2008	*Training and Education in Professional Psychology*	211
21. Hart & Nance	2003	*Counselor Education and Supervision*	258
22. Henry et al.	2004	*The Clinical Supervisor*	190
23. Inman	2006	*Journal of Marital and Family Therapy*	147
24. Johnson & Stewart	2008	*Training and Education in Professional Psychology*	155
25. Kanno & Koeske	2010	*Journal of Social Work Education*	144
26. Kahn	1999	*Professional School Counseling*	119

27. Ladany et al.	2001	*Counselor Education and Supervision*	137
28. Ladany et al.	1999	*Journal of Counseling & Development*	107
29. Ladany & Lehrman-Waterman	1999	*Counselor Education and Supervision*	210
30. Ladany et al.	1999	*The Counseling Psychologist*	151
31. Ladany et al.	1997	*Journal of Counseling Psychology*	116
32. Ladany et al.	1997	*Counselor Education and Supervision*	105
33. Ladany et al.	1996	*Journal of Counseling Psychology*	108
34. Ladany, N., & Friedlander, Myrna L.	1995	*Counselor Education and Supervision*	123
35. Ligiéro, D.P. & Gelso, C.J.	2002	*Psychotherapy: Theory, Research, Practice, Training*	100
36. Locke, L.D. & McCollum, E.E.	2001	*Journal of Marital and Family Therapy*	108
37. McCarthy et al.	1994	*Professional Psychology: Research and Practice*	232
38. McCurdy & Owen	2008	*The Journal of Individual Psychology*	31
39. Miller & Larrabee	1995	*Counselor Education and Supervision*	315
40. Mori et al.	2009	*Training and Education in Professional Psychology*	104
41. Navin et al.	1995	*Journal of Mental Health Counseling*	321
42. Nilsson & Duan	2007	*Journal of Multicultural Counseling and Development*	69
43. Nilsson & Anderson	2004	*Professional Psychology: Research and Practice*	42
44. Page et al.	2001	*Counselor Education and Supervision*	267
45. Raichelson et al.	1997	*The Clinical Supervisor*	300
46. Ramos-Sanchez et al.	2002	*Professional Psychology: Research and Practice*	126
47. Reese et al.	2009	*Training and Education in Professional Psychology*	28
48. Riggs & Bretz	2006	*Professional Psychology: Research and Practice*	87
49. Riva et al.	1995	*Professional Psychology: Research and Practice*	243
50. Romans et al.	1995	*Professional Psychology: Research and Practice*	46
51. Schultz et al.	2002	*Rehabilitation Counseling Bulletin*	111
52. Scott et al.	2000	*The Counseling Psychologist*	688
53. Sterner	2009	*Journal of Mental Health Counseling*	71
54. Stevens et al.	1997	*The Clinical Supervisor*	60
55. Studer & Oberman	2006	*Professional School Counseling*	73
56. Szymanski	2005	*The Counseling Psychologist*	135
57. Tromski-Klingshirn & Davis	2007	*Counselor Education and Supervision*	158
58. Tyler et al.	2000	*Psychotherapy: Theory, Research, Practice, Training*	300
59. Utsey & Gernat	2002	*Journal of Counseling & Development*	145

60. Walker et al.	2007	*Counseling & Psychotherapy Research*	111
61. Wester et al.	2004	*Journal of Counseling & Development*	103
62. Wilbur et al.	1994	*Counselor Education and Supervision*	194

Table 3

Measures Utilized in the Research Articles Included in the Current Study

Measure	Used by	Created Validity/ for study	Modified for study	Reliability reported ?
American-International Relations Scale (*AIRS*; Sodowsky & Plake, 1991, 1992) Nilsson & Anderson (2004)	Mori et al. (2009)	N	N	Y
Beck Depression Inventory-II (Beck, Steer, & Brown, 1996)	Callahan et al. (2009)	N	N	Y
	Culbreth (1999)	N	N	Y
Case Conceptualization Exercise (Butler & Constantine, 2006)	Butler & Constantine (2006)	Y	-	N
Child Behavior Checklist (*CBCL*; Achenbach, 1991)	Schoenwald et al. (2009)	N	N	Y
Client Satisfaction Questionnaire-8 (*CSQ-8;* Attkisson et al., 1989)	Locke & McCollum (2001)	N	N	Y
Clinical Supervision Questionnaire (1) (McCarthy et al., 1994)	McCarthy et al. (1994)	Y	-	Y
Clinical Supervision Questionnaire (2) (Tromski-Klingsirn & Davis, 2007)	Tromski-Klingsirn & Davis (2007)	Y	-	N

Collective Self-Esteem Scale (*CSES*; Luhtanen & Crocker, 1992)	Butler & Constantine (2006)	N	N	Y
Counseling Self-Estimate Inventory (*COSE*; Larson et al., 1992)	Cashwell & Dooley (2001)	N	N	Y
	Fernando & Hulse-Killacky (2005)	N	N	Y
	Nilsson & Anderson (2004)	N	N	Y
	Nilsson & Duan (2001)	N	N	Y
	Reese et al. (2009)	N	N	Y
	Wester et al. (2004)	N	N	Y
Counselor Skill and Personnel Development Rating Form (*CSPD-RF*; Wilbur & Roberts-Wilbur, 1994)	Wilbur & Roberts-Wilbur (1994)	Y	-	N
Counselor Rating Form (*CRF-S*; Corrigan & Schmidt, 1983)	Anderson (2000)	N	N	Y
	Callahan et al. (2009)	N	N	Y
Countertransference Index (*CT*; Hayes, Riker, & Ingram, 1997)	Ligiéro & Gelso (2002)	N	N	Y
Cross-Cultural Counseling Inventory—Revised (*CCCI-R*; LaFromboise et al., 1991)	Constantine et al. (2005)	N	N	Y
	Ladany et al. (1997)	N	N	Y
	Gloria et al. (2008)	N	Y	Y
Cultural Identity Attitude Scale (*CIAS*; Helms & Carter, 1990)	Ladany et al. (1997)	N	N	Y
Deferred Imitation Scale (*DIS*; Geller et al., 2010)	Geller et al. (2010)	Y	-	N

Discussion of cultural variables questions

(Gatmon et al., 2001)	Gatmon et al. (2001)	Y	-	N
Feminist Perspectives Scale (*FPS*; Henley et al., 1998)	Szymanski (2005)	N	N	Y
Existing and Preferred Supervision Practices (Borders and Usher, 1992)	Coll (1995)	N	N	Y
Feminist Supervision Scale (*FSS*; Szymanski, 2003)	Szymanski (2005)	N	N	Y
Gender Related Events Survey (Walker et al., 2007)	Walker et al. (2007)	Y	-	N
Gender Role Conflict Scale (*GCRS*; O'Neil et al., 1986)	Wester et al. (2004)	N	N	Y
International Student Supervision Scale (*ISSS*; Nilsson & Dodds, 2004)	Mori et al. (2009)	N	N	Y
	Nilsson & Anderson (2004)	N	N	Y
Majority –Minority Relations Survey (*MMRS*: Sodowsky, Lai, & Plake, 1991)	Nilsson & Duan (2001)	N	N	Y
Minnesota Satisfaction Questionnaire—Short Form (*MSQ*; Weiss et al., 1967)	Sterner (2009)	N	N	Y
MST Therapist Adherence Measure—Revised (*TAM-R*; Henggeler et al., 2006)	Schoenwald et al. (2009)	N	N	Y
Multicultural Case Conceptualization Ability (Inman, 2006)	Inman (2006)	Y	-	N

Multicultural case conceptualization ability exercise (Constantine et al., 2005)	Constantine et al. (2005)	Y	-	N
Occupational Stress Inventory—Revised (OSI-R) (Osipow, 1998)	Sterner (2009)	N	Y	N
ORS (Miller & Duncan, 2000)	Reese et al. (2009)	N	N	Y
Parallel Process Survey (Raichelson et al., 1997)	Raichelson et al. (1997)	Y	-	N
People of Color Racial Identity Attitude Scale (*PRIAS*; Helms, 1995)	Bhat & Davis (2007)	N	N	Y
Perceptions of Supervisor Racial Identity (PSRI; Ladany et al., 1997)	Ladany et al. (1997)	Y	-	N
Perceptions of Supervisee Racial Identity for POC (*PSeRIP*; Modification of PSRI; Ladany, 1997)	Bhat & Davis (2007)	N	Y	N
Perceptions of Supervisee Racial Identity for Whites (*PSeRIW;* Modification of *PSRI*; Ladany, 1997)	Bhat & Davis (2007)	N	Y	N
Priorities and practices in field supervision of school counseling students (Kahn, 1999)	Kahn (1999)	Y	-	N
Psychotherapy supervision practices of academic faculty (Tyler et al., 2000)	Tyler et al. (2000)	Y	-	N
Purdue Live Observation Satisfaction Scale				

(*PLOSS*; Sprenkle et al., 1982)	Locke & McCollum (2001)	N	N	Y
Questions regarding multicultural supervision (Gloria et al., 2008)	Gloria et al. (2008)	Y	-	N
Rahim Leader Power Inventory (*RLPI*; Rahim, 1988)	Schultz et al. (2002	N	N	Y
Referent and Manner scales of the Therapist Experiencing Scale (Klein & Keisler, 1986).	Ellis et al. (2002)	N	N	Y
Relationship Questionnaire (Bartholomew & Horowitz, 1991)	Ligiéro & Gelso (2002)	N	N	Y
	Ramos-Sanchez et al. (2002) N	N	Y	Y
Revised Feminist Identity Development Scale (*FIDS*; Bargard & Hyde, 1991)	Szymanski (2005)	N	N	Y
Role Conflict and Role Ambiguity Inventory (*RCRAI*; Olk & Friedlander, 1992)	Nilsson & Anderson (2004)	N	N	Y
	Nilsson & Duan (2001)	N	N	Y
	Ladany & Friedlander (1995)	N	N	Y
Schedule of Race-Related Ego Defenses-Counselor Form (SHRED-C; Utsey & Gernat, 2002)	Utsey & Gernat (2002)	Y	-	Y
School Counselor Supervision Questionnaire (Studer & Oberman, 2006)	Studer & Oberman (2006)	Y	-	N
SCS (Prentice-Dunn & Rogers, 1982)	Ellis et al. (2002)	N	N	Y
Self-Efficacy Inventory (*SEI*; Friedlander & Snyder, 1983)	Ladany et al. (1999b)	N	N	Y

Self-efficacy measure (Stevens et al., 1998)	Stevens et al. (1998)	Y	-	N
SRS (Miller et al., 2000)	Reese et al. (2009)	N	N	Y
State form of the State–Trait Anxiety Inventory (*STAI*; Spielberger et al., 1970)	Birk et al. (1994)	N	N	Y
State form of the State–Trait Anxiety Inventory—Form Y (*SAI*; Spielberger, 1977)	Ellis et al. (2002)	N	N	Y
Supervisee Description Questionnaire (Ossana, 1991)	Birk et al. (1994)	N	N	Y
Supervisee Levels Questionnaire—Revised (*SLQ-R*; McNeill, Stoltenberg, & Romans, 1992)	Ramos-Sanchez et al. (2002)	N	N	Y
Supervision Level Scale (*SLS*; Wiley & Ray, 1986)	Birk et al. (1994)	N	Y	N
	Chagnon & Russell (1995)	N	N	Y
Supervision Outcomes Survey (*SOS*; Worthen & Isakson, 2003)	Reese et al. (2009)	N	N	Y
Supervision Survey (Johnson & Stewart, 2008)	Johnson & Stewart (2008)	Y	-	N
Supervision Questionnaire—Revised (Worthington & Roehlke, 1979)	Gatmon et al. (2001)	N	N	Y
Supervisor Ethical Behavior Scale (*SEBS*; Ladany et al., 1999)	Ladany et al. (1999)	Y	-	Y

Supervisor Ethical Practices Questionnaire (*SEPQ*; Ladany et al., 1999)	Ladany et al. (1999)	Y	-	Y
Supervisor Self-Disclosure Questionnaire (SSDQ; Ladany & Lehrman-Waterman, 1999)	Ladany & Lehrman-Waterman (1999)	Y	-	Y
Supervisor Questionnaire (Ladany et al., 1996)	Tromski-Klingshirn & Davis (2007)	N	Y	Y
Supervisor Adherence Measure (*SAM*; Schoenwald et al., 1998)	Schoenwald et al. (2009)	N	N	Y
Supervisory Embodiment Scale (*SES*; Geller & Schaffer, 1992)	Geller et al. (2010)	N	N	Y
Supervisory Emphasis Report Form-Revised (SERF-R; Lannine & Freeman, 1993)	Stevens et al. (1998)	N	N	Y
Supervisory Interactional Dynamics (Anderson et al., 2000)	Anderson et al. (2000)	Y	-	N
Supervisory Functions Scale (*SFS*; Geller & Schaffer, 1992)	Geller et al. (2010)	N	N	Y
Supervisor Multicultural Competency Inventory (Inman, 2005)	Inman (2006)	N	N	Y
	Mori et al. (2009)	N	N	Y
Supervisory Occasion Scale (*SOS*; Geller & Schaffer, 1992)	Geller et al. (2010)	N	N	Y
Supervisor Self-Disclosure Inventory				

(*SSDI*; Ladany & Lehrman-Waterman, 1999)	Ladany et al. (2001)	N	N	Y
Supervisory Satisfaction Questionnaire				
(*SSQ*; Larsen et al., 1979)	Fernando & Hulse-Killacky (2005)	N	N	Y
	Ladany et al. (1996)	N	N	Y
	Ladany et al. (1999a)	N	N	Y
	Mori et al. (2009)	N	N	Y
Supervisory Styles Inventory				
(Friedlander & Ward, 1984)	Fernando & Hulse-Killacky (2005)	N	N	Y
	Ladany et al. (1996)	N	N	Y
	Ladany & Lehrman-Waterman (1999)	N	N	Y
	Ladany et al. (2001)	N	N	Y
Supervisory Styles Inventory				
(Hart & Nance, 2003)	Hart & Nance (2009)	N	N	Y
Supervisory Working Alliance Inventory—Trainee Form				
(*SWAI—Trainee*; Efstation et al., 1990)	Ladany et al. (1999b)	N	N	Y
	McCurdy & Owen (2008)	N	N	Y
	Nilsson & Anderson (2004)	N	N	Y
	Reese et al. (2009)	N	N	Y
	Schultz et al. (2002)	N	N	Y
	Sterner (2009)	N	N	Y
	Wester et al. (2004)	N	N	Y
Survey for Counselors of Licensure Applicants				
(Borders et al., 1995)	Borders et al. (1995)	Y	-	N

Survey of MSW Students' Perceptions of their Field Placement (Kannon & Koeske, 2010)	Kanno & Koeske (2010)	Y	-	N
Survey for training director (Romans et al., 1995)	Romans et al. (1995)	Y	-	N
Survey of supervisor ethical behavior (Navin et al., 1995)	Navin et al. (1995)	Y	-	N
Survey of supervision training practices (Scott et al., 2000)	Scott et al. (2000)	Y	-	N
Survey of supervision training in predoctoral internship sites (Scott et al., 2000)	Scott et al. (2000)	Y	-	N
Symptom Checklist-9 (Derogatis, 1992)	Callahan et al. (2009)	N	N	Y
Topics of Supervision Report Dow et al. (2009)	Dow et al. (2009)	Y	-	N
Trainee Disclosure Scale (*TDS*; Walker et al., 2007)	Walker et al. (2007)	Y	-	Y
Trainee Satisfaction with Supervision Scale (Holloway & Wampold, 1984)	Ladany et al. (1999b)	N	N	Y
Unethical intimacy survey (Glaser and Thorpe, 1986)	Miller & Larrabee (1995)	N	Y	N
Vanderbilt Functioning Inventory (*VFI*; Bickman et al., 1998)	Schoenwald et al. (2009)	N	N	Y

Working Alliance Inventory (Horvath & Greenberg, 1989)	Gatmon et al (2001)	N	N	Y
Working Alliance Inventory—Revised (Baker, 1990)	Ramos-Sanchez et al. (2002)	N	N	Y
Working Alliance Inventory—Supervisor Version (*WAI-S*; Bahrick, 1989)	Ladany et al. (2001)	N	N	Y
	Bhat & Davis (2007)	N	N	Y
Working Alliance—Trainee Version (*WAI—T*; Bahrick, 1989)	Inman (2006)	N	N	Y
	Ladany & Friedlander (1995)	N	N	Y
	Ladany et al. (1997)	N	N	Y
	Ladany et al. (1999)	N	N	Y
	Ladany & Lehrman-Waterman (1999)	N	N	Y
	Walker et al. (2007)	N	N	Y
Working Alliance Inventory for Therapists—Short Version (*WAI—Therapist*; Tracey & Kokotovic, 1989)	Ligiéro & Gelso (2002)	N	N	Y
White Racial Identity Attitude Scale (*WRIAS*; Helms & Carter, 1990)	Bhat & Davis (2007)	N	N	Y
	Constantine et al. (2005)	N	N	Y
	Ladany et al. (1997)	N	N	Y
	Utsey & Gernat (2000)	N	N	Y

Note: where measures are reported as modified, the column "validity/reliability reported" refers to the modification, not the original measure. Also, in cases where authors created measures with the same name as pre-existing measures, the measures are listed in ascending order of original publication date and numbered 1, 2, etc.

Table 4

Means, Standard Deviations, Standard Errors, and 95% Confidence Intervals for Statistical Variables

Variable	*M*	*SD*	*Mdn*	*SE*	*95% CI (Mdn)*
		Across 1,202 statistical tests (sufficient information presented)			
N per analysis[1]	121.457	118.507	107	3.444	100.251-113.749
Sample effect size[2] (partial η^2)	0.1212	0.144	0.065	0.004	0.057-0.074
Minimum detectable effect size[3] (partial $\eta^2_{min(N)}$)	0.072	0.051	0.051	0.002	0.048-0.054
Post-hoc power - Ellis et. al. definition[4] ($P_{(\eta 2)}$)	0.676	0.378	0.899	0.013	0.874-0.924
Post-hoc power assuming a small effect[5] ($P_{(Small)}$)	0.196	0.121	0.181	0.004	0.174-0.188
Post-hoc power assuming a medium effect[6] ($P_{(Med)}$)	0.795	0.217	0.897	0.007	0.882-0.912
Post-hoc power assuming a large effect[7] ($P_{(Large)}$)	0.967	0.102	1	0.004	0.991-1.009
Type II error[8] ($\beta_{(\eta 2)}$)	0.333	0.383	0.11	0.013	0.084-0.135
		Across 51 studies (sufficient information presented)			
N per study[9]	154.984	192.24	111	27.628	56.848 -165.152
Sample effect size[10] (partial η^2)	0.139	0.1	0.112	0.015	0.084 - 0.141
Minimum detectable effect size[11] (partial $\eta^2_{min(N)}$)	0.074	0.049	0.058	0.007	0.043 - 0.072
Post-hoc power - Ellis et. al. definition[12] ($P_{(\eta 2)}$)	0.721	0.235	0.771	0.034	0.705 – 0.837
Post-hoc power assuming a small effect[13] ($P_{(Small)}$)	0.214	0.152	0.181	0.022	0.138--0.224
Post-hoc power assuming a medium effect[14] ($P_{(Med)}$)	0.773	0.21	0.831	0.031	0.771--0.891
Post-hoc power assuming a large effect[15] ($P_{(Large)}$)	0.947	0.108	0.997	0.017	0.964--1.03
Experiment-wise Type I error[16] (α_{EW})	0.347	0.268	0.226	0.041	0.145--0.307
Experiment-wise Type II error[17] ($\beta_{EW(\eta 2)}$)	0.447	0.301	0.431	0.042	0.348--0.513

Number of statistical tests per study[18]	23.333	30.634	13	4.532	4.118 – 21.882
Number of tests significant[19]	16.804	24.496	11	3.534	4.074--17.926

Note. CI = confidence interval

[a] n = 56. [b] n =#. [c] ###studies included sufficient information to compute α_{EW}

[1]range: 28 – 751; [2] range: 0.0247 – 0.34357; [3] range: 0.1846 – 0.1610; [4] range: 0.2323 – 1; [5] range: 0.0956 – 0.6293; [6] range: 0.27 - ; [7] range: 0.7697 – 1; [8] range: 0 – 0.9777; [9] range: 29 – 751; [10] range: 0.0021 – 0.3302; [11] range: 0.0215 – 0.9667; [12] range: 0.38 – 1; [13] range: 0.221 – 0.531; [14] range: 0.27 - 1; [15] range: 0.8254 – 1;[16] range: 2 – 38; [17] range: 3 – 23.

Table 5

Prevalence of Validity Threats in Ellis et al. (1996) and the Current Study

	Ellis et al. (1996)	Current Study	Kappas for Current Study
Statistical conclusion validity			
Low statistical power	76.67	77.00	98%
Violation of assumption of statistics	24.67	**5.00**	90%
Inflated error rate	84.67	**6.00**	95%
Unreliability of dependent/ indep measures	80.67	*58.00*	87%
Unreliability of treatment implementation	15.33	*42.00*	85%
Irrelevance in experimental setting	27.33	**94.00**	98%
Heterogeneity of participants	55.33	*81.00*	100%
Internal validity			
History	19.33	18.00	98%
Maturation	20.00	36.00	100%
Testing	20.67	7.00	100%
Instrumentation	3.33	7.00	88%
Statistical regression	10.67	13.00	100%
Differential attrition	20.00	13.00	88%
Interaction with selection	35.33	28.00	90%
Ambiguity of causal direction	68.67	*83.00*	100%
Diffusion of treatment	0.67	0.00	100%
Compensatory equalization of treatments	0.00	0.00	100%
Resentful demoralization	0.00	0.00	100%
Construct validity			
Inadequate preoperationalization explication	68.67	**13.00**	90%
Mono-operation bias	24.00	*44.00*	100%
Monomethod bias	78.67	72.00	100%
Hypothesis guessing within treatments	13.34	9.00	95%
Evaluation apprehension	16.00	**94.00**	98%
Experimenter expectancies	18.00	5.00	100%
Confounding of construct with levels of construct	68.67	**30.00**	88%
Interaction of treatments	6.00	7.00	85%
Interaction of testing and treatments	22.67	*00.00*	88%
Restricted generalizability across constructs	55.33	*21.00*	90%
External validity			
Interaction of selection and treatment	94.67	**35.00**	88%
Interaction of setting and treatment	88.67	**57.00**	85%
Interaction of history and treatment	82.67	**35.00**	88%

Russell et al.'s (1984) threats			
Lack of adequate control group	30.67	**92.00**	100%
No pretreatment assessment	43.33	*22.00*	100%
Inadequate sample size	78.00	**00.00**	100%
Variations/confounds in length of training	7.33	4.00	100%
Nonrandom assignment to conditions	58.67	*35.00*	100%
Widely discrepant cell sizes	20.67	*30.00*	95%
Restricted range of dependent variables	22.00	33.00	98%
Nonrepresentative supervisee/supervisor pop	7.33	**61.00**	95%
Lack of follow-up assessment	62.00	**22.00**	100%
Use of roleplay or audiotaped client statement	7.33	*20.00*	100%
Exclusive reliance on self-report data	66.00	*86.00*	100%
Overly brief training period	2.67	0.00	100%

Notes: Numbers reported in column of *threat present in Ellis et al. (1996)* were copied from Table 1 in Ellis et al. (1996), p 40.

Bold indicates a large difference (> 50%) between Ellis et al. (1996) and current study

Italics indicate a medium difference (between 25% and 50% difference) between Ellis et al.(1996) and current study

Underline indicates a small difference (less than 25%) between Ellis et al. (1996) and current study

Table 6

Top Most Salient Methodological Threats

Type of Threat	Percent
Evaluation apprehension	94.00
Irrelevance in experimental setting	94.00
Lack of adequate control group	92.00
Exclusive reliance on self-report data	86.00
Ambiguity of causal direction	83.00
Instrumentation	82.00
Heterogeneity of participants	81.00
Monomethod bias	72.00
Nonrepresentative supervisee/supervisor pop	61.00
Unreliability of dep/ indep measures	58.00
Interaction of setting and treatment	57.00
Mono-operation bias	44.00
Unreliability of treatment implementation	42.00
Maturation	36.00
Interaction of history and treatment	35.00
Interaction of selection and treatment	35.00
Nonrandom assignment to conditions	35.00
Restricted range of dependent variables	33.00
Widely discrepant cell sizes	30.00
Confounding of construct with levels of construct	30.00

Table 7

Inference categories and associated subcategories

First Inference: Inferences about the Supervisory Relationship	*Fifth Inference: Inferences about the Use of Technology in Supervision*
Supervisory Working Alliance Model	E-mail
Role Conflict and Ambiguity	Online Discussion
Structure of the Supervisory Relationship	
Supervisor Style	
Ethics in the Supervisory Relationship	*Sixth Inference: Inferences about Supervisor Training*
Parallel Process	
	No subcategories
Second Inference: Inferences Regarding the Supervisee	
	Seventh Inference: General Inferences about the Practice of Supervision
Supervisee Nondisclosures	
Self-efficacy	Supervision Practices of Academic Faculty
Developmental models	Supervision of Community College Counselors
Third Inference: Inferences about Client Outcome	Supervision of Marriage and Family Counselors
	Supervision of School Counselors
No subcategories	Supervision of Substance Abuse Counselors
	General Perceptions of Clinical Supervision
Fourth Inference: Inferences about Culture and Multicultural Competence	
Multicultural Competence	
International Trainees/Students	
Matching in Supervision	
Gender	

Appendix A

108 Articles reviewed in the current study

---

*Anderson, S.A., Schlossberg, M., & Rigazio-DiGilio, S. (2000). Family therapy trainees' evaluations of their best and worst supervision experiences. *Journal of Marital and Family Therapy, 26*, 79-91.

Barnett-Queen, T. & Larrabee, M.J. (2000). Sexually oriented relationships between educators and students in mental-health-education programs. *Journal of Mental Health Counseling, 22*, 68-84.

*Bhat, C.S. & Davis, T.E. (2007). Counseling supervisors' assessment of race, racial identity, and working alliance in supervisory dyads. *Journal of Multicultural Counseling and Development, 35*, 80-91.

*Birk, J.M. & Mahalik, J.R. (1996). The influence of trainee conceptual level, trainee anxiety and supervision evaluation on counselor developmental level. *The Clinical Supervisor, 14*, 123-137.

Borders, L.D., Cashwell, C.S., & Rotter, J.C. (1995). Supervision of counselor licensure applicants: A comparative study. *Counselor Education and Supervision, 35*, 54-69.

Butler, S.K. & Constantine, M.G. (2006). Web-based peer supervision, collective self-esteem, and case conceptualization ability in school counselor trainees. *Professional School Counseling*, *10*, 146-152.

*Callahan, J.L., Almstrom, C.M., Swift, J.K., Borja, S.E., & Heath, C.J. (2009). Exploring the contribution of supervisors to intervention outcomes. *Training and Education in Professional Psychology*, *3*, 72-77.

Carter, J.W., Enyedy, K.C., Goodyear, R.K., Arcinue, F., & Puri, N. N. (2009). Concept mapping of the events supervisees find helpful in group supervision. *Training and Education in Professional Psychology*, *3*, 1-9.

*Carlozzi, A.F., Romans, J.S.C., Boswell, D.L., Ferguson, D.B., & Whisenhunt, B.J. (2001). Training and supervison practices in counseling and marriage and family therapy programs. *The Clinical Supervisor*, 15, 51-60. Cashwell, T.H., & Dooley, K. (2001). The impact of supervision on counselor self-efficacy. *The Clinical Supervisor, 20*, 39-47.

*Chagnon, J. & Russell, R.K. (1995). Assessment of supervisee developmental level and supervision environment across supervisor experience. *Journal of Counseling & Development, 73*, 553-558.

Chui, E.W. T. (2010). Desirability and feasibility in evaluating fieldwork performance: Tensions between supervisors and students. *Social Work Education, 29*, 171-187.

*Clingerman, T.L. & Bernard, J.M. (2004). An investigation of the use of e-mail as a supplemental modality for clinical supervision. *Counselor Education and Supervision, 44*, 82-95.

*Coleman, M.N., Kivlighan, D.M., Jr., & Roehlke, H.J. (2009). A taxonomy of the feedback given in the group supervision of group counselor trainees. *Group Dynamics: Theory, Research, and Practice, 13*, 300-315.

*Coll, K.M. (1995). Clinical supervision of community college counselors: Current and preferred practices. *Counselor Education and Supervision, 35*, 111-117.
Constantine, M.G., Warren, A.K., & Miville, M. L. (2005).White Racial Identity dyadic interactions in supervision: Implications for supervisees' multicultural counseling competence. *Journal of Counseling Psychology, 52*, 490-496.
Culbreth, J.R. (1999). Clinical supervision of substance abuse counselors: *Current and preferred practices. Journal of Addictions & Offender Counseling, 20*, 15-25.
Dennin, M.K., & Ellis, M.V. (2003). Effects of a method of self-supervision for counselor trainees. *Journal of Counseling Psychology, 50*, 69-83.
deMayo, R.A. (2000). Patients' sexual behavior and sexual harassment: A survey of clinical supervisors. *Professional Psychology: Research and Practice, 31*, 706-709.
*Dow, D.M., Hart, G.M., & Nance, D.W. (2009). Supervision styles and topics discussed in supervision. *The Clinical Supervisor, 28*, 36-46.
Dressel, J.L., Consoli, A.J., Kim, B.S., & Atkinson, D.R. (2007). Successful and unsuccessful multicultural supervisory behaviors: A Delphi poll. *Journal of Multicultural Counseling and Development, 35*, 51-64.
*Ellis, M.V., Krengel, M., & Beck, M. (2002). Testing self-focused attention theory in clinical supervision: Effects of supervisee anxiety and performance. *Journal of Counseling Psychology, 40*, 101-116.
Enyedy, K.C., Arcinue, F., Puri, N.N., Carter, J.W., Goodyear, R.K., & Getzelman, M.A. (2003).Hindering phenomena in group supervision: Implications for practice. *Professional Psychology: Research and Practice, 34*, 312-317.
*Fernando, D.M. & Hulse-Killacky, D. (2005). The relationship of supervisory styles to satisfaction with supervision and the perceived self-efficacy of master's-level counseling students. *Counselor Education and Supervision, 44*, 293-304.
Fortune, A., & Abramson, J. (1993). Predictors of satisfaction with field practicum among social work students. *The Clinical Supervisor, 11*, 95-110.
Gabbay, M. B., Kiemle, G., Maguire, C. (1999). Clinical supervision for clinical psychologists: Existing provision and unmet needs. *Clinical Psychology & Psychotherapy, 6*, 404-412.
*Gainor, K.A., & Constantine, M.G. (2002). Multicultural group supervision: A comparison of in-person versus web-based formats. *Professional School Counseling, 6*, 104-111.
*Gatmon, D., Jackson, D., Koshkarian, L., Martos-Perry, N., Molina, A., Patel, N., & Rodolfa, E. (2001). Exploring ethnic, gender, and sexual orientation variables in supervision: Do they really matter? *Journal of Multicultural Counseling and Development, 29*, 102-113.
*Geller, J.D., Farber, B.A., & Schaffer, C.E. (2010). Representations of the supervisory dialogue and the development of psychotherapists. *Psychotherapy: Theory, Research, Practice, Training, 47*, 211-220.
Gloria, A.M., Hird, J. S., & Tao, K. W. (2008). Self-reported multicultural supervision competence of White predoctoral intern supervisors. *Training and Education in Professional Psychology, 2*, 129-136.
*Hart, G., & Nance, D. (2003). Styles of counselor supervision as perceived by supervisors and supervisees. *Counselor Education and Supervision, 43*, 146-159.

Haverkamp, B. (1994). Using assessment in counseling supervision: Individual differences in self-monitoring. *Measurement and Evaluation in Counseling and Development, 27,* 316-324

Henggeler, S.W., Schoenwald, S.K., Liao, J.G., Letourneau, E. J., & Edwards, D. L. (2002). Transporting efficacious treatments to field settings: The link between supervisory practices and therapist fidelity in MST programs. *Journal of Clinical Child and Adolescent Psychology, 31*, 155-167.

*Henry, P.J., Hart, G.M., & Nance, D.W. (2004). Supervision topics as perceived by supervisors and supervisees. *The Clinical Supervisor, 23,* 139-152.

Herbert, J.T., Ward, T.J., & Hemlick, L.M. (1995). Confirmatory factor analysis of the Supervisory Style Inventory and the Revised Supervision Questionnaire. *Rehabilitation Counseling Bulletin, 38,* 334-349.

Hilton, D.B., Russell, R.K., & Salmi, S.W. (1995). The effects of supervisor's race and level of support on perceptions of supervision. *Journal of Counseling and Development, 73,* 57-563.

*Inman, A.G. (2006). Supervisor multicultural competence and its relation to supervisory process and outcome. *Journal of Marital and Family Therapy, 32,* 73-85.

*Johnson, E. A., Stewart, D.W. (2008). Perceived competence in supervisory roles: A social cognitive analysis. *Training and Education in Professional Psychology, 2,* 229-236.

*Kahn, B.B. (1999). Priorities and practices in field supervision of school counseling students. *Professional School Counseling, 3,* 128-136.

*Kanno, H., Koeske, G.F. (2010). MSW students' satisfaction with their field placement: The role of preparedness and supervision quality. *Journal of Social Work Education, 46,* 23-38.

Knight, C. (2001). The process of field instruction: BSW and MSW students' views of effective field supervision. *Journal of Social Work Education, 37,* 357-379.

*Ladany, N., & Friedlander, M.L. (1995).The relationship between the supervisory working alliance and trainees' experience of role conflict and role ambiguity. *Counselor Education and Supervision, 34,* 220-231.

*Ladany, N., Hill, C.E., Corbett, M.M., & Nutt, E.A. (1996).Nature, extent, and importance of what psychotherapy trainees do not disclose to their supervisors. *Journal of Counseling Psychology, 43,* 10-24.

*Ladany, N., Brittan-Powell, C.S., & Pannu, R.K. (1997).The influence of supervisory racial identity interaction and racial matching on the supervisory working alliance and supervisee multicultural competence. *Counselor Education and Supervision, 36,* 284-304.

*Ladany, N., Inman, A.G., Constantine, M.G., & Hofheinz, E.W. (1997).Supervisee multicultural case conceptualization ability and self-reported multicultural competence as functions of supervisee racial identity and supervisor focus. *Journal of Counseling Psychology, 44,* 284-293.

*Ladany, N., Ellis, M.V., & Friedlander, M. L. (1999).The supervisory working alliance, trainee self-efficacy, and satisfaction. *Journal of Counseling & Development, 77,* 447-455.

*Ladany, N. & Lehrman-Waterman, D. E. (1999).The content and frequency of supervisor self-disclosures and their relationship to supervisor style and the supervisory working alliance. *Counselor Education and Supervision, 38,* 143-160.

*Ladany, N., Lehrman-Waterman, D., Molinaro, M., & Wolgast, B. (1999).Psychotherapy supervisor ethical practices: Adherence to guidelines, the supervisory working alliance, and supervisee satisfaction. *The Counseling Psychologist, 27,* 443-475.

*Ladany, N., Walker, J. A., & Melincoff, D. S. (2001). Supervisory style: Its relation to the supervisory working alliance and supervisor self-disclosure. *Counselor Education and Supervision, 40,* 263-275.

Ladany, N., Marotta, S., & Muse-Burke, J. L. (2001). Counselor experience related to complexity of case conceptualization and supervision preference. *Counselor Education and Supervision, 40,* 203-219

Lehrman-Waterman, D., & Ladany, N. (2001).Development and validation of the Evaluation Process Within Supervision Inventory. *Journal of Counseling Psychology, 48,* 168-177.

Lent, R.W., Cinamon, R.G., Bryan, N.A., Jezzi, M.M., Martin, H.M., & Lim, R. (2009). Perceived sources of change in trainees' self-efficacy beliefs. *Psychotherapy: Theory, Research, Practice, Training, 46,* 317-327.

*Ligiéro, D.P. & Gelso, C.J. (2002). Countertransference, attachment, and the working alliance: The therapist's contribution. *Psychotherapy: Theory, Research, Practice, Training, 39,* 3-11.

*Lochner, B.T. & Melchert, T.P. (1997). Relationship of cognitive style and theoretical orientation to psychology interns' preferences for supervision. *Journal of Counseling Psychology, 44,* 256-260.

*Locke, L.D. & McCollum, E.E. (2001). Clients' views of live supervision and satisfaction with therapy. *Journal of Marital and Family Therapy, 27,* 129-133.

Lovell, C. (1999). Supervisee cognitive complexity and the integrated developmental model. *The Clinical Supervisor, 18,* 191-201.

*McCarthy, P., Kulakowski, D., Kenfield, J.A. (1994).Clinical supervision practices of licensed psychologists. *Professional Psychology: Research and Practice, 25,* 177-181.

*McCurdy, K. G. & Owen, J. J. (2008). Using sandtray in Adlerian-based clinical supervision:An initial empirical analysis. *The Journal of Individual Psychology, 64,* 96-112.

McMahon, M. & Simons, R. (2004).Supervision training for professional counselors: An exploratory study. *Counselor Education and Supervision, 43,* 301-309.

*Miller, G.M. & Larrabee, M.J. (1995). Sexual intimacy in counselor education and supervision: A national survey. *Counselor Education and Supervision, 34,* 332-343.

Miller, M., M., Korinek, A.W. & Ivey, D. C. (2006). Integrating spirituality into training: The Spiritual Issues in Supervision Scale. *American Journal of Family Therapy, 34,* 355-372.

Milne, D. (2010).Can we enhance the training of clinical supervisors? A national pilot study of an evidence-based approach. *Clinical Psychology & Psychotherapy, 17,* 321-328.

*Mori, Y., Inman, A.G., & Caskie, G. I. (2009). Supervising international students: Relationship between acculturation, supervisor multicultural competence, cultural discussions, and supervision satisfaction. *Training and Education in Professional Psychology, 3*, 10-18

*Navin, S., Beamish, P., & Johanson, G. (1995). Ethical practices of field-based mental health counselor supervisors. *Journal of Mental Health Counseling, 17,* 243-253.

Nelson, M.L., & Friedlander, M.L. (2001).A close look at conflictual supervisory relationships: The trainee's perspective. *Journal of Counseling Psychology, 48,* 384-395.

*Nilsson, Johanna E., Duan, Changming (2007).Experiences of prejudice, role difficulties, and counseling self-efficacy among U.S. racial and ethnic minority supervisees working with White supervisors. *Journal of Multicultural Counseling and Development, 35*, 219-229.

Nilsson, J.E., & Dodds, A.K. (2006). A pilot phase in the development of the international student supervision scale. *Journal of Multicultural Counseling and Development, 34*, 50-62.

*Nilsson, J.E. & Anderson, M.Z. (2004). Supervising International Students: The Role of Acculturation, Role Ambiguity, and Multicultural Discussions. *Professional Psychology: Research and Practice, 35*, 306-312.

Nyman, S.J., Nafziger, M.A., & Smith, T.B. (2010).Client outcomes across counselor training level within a multitiered supervision model. *Journal of Counseling & Development, 88*, 204-209.

*Page, B. J., Pietrzak, D.R., & Sutton, J.M., Jr. (2001). National survey of school counselor supervision. *Counselor Education and Supervision, 41*, 142-150.

Peace, S.D. & Sprinthall, N.A. (1998). Training school counselors to supervise beginning counselors: Theory, research, and practice. *Professional School Counseling, 1*, 2-8.

Peleg-Oren, N., Macgowan, M. J., & Even-Zahav, R. (2007). Field instructors' commitment to Student Supervision: Testing the investment model. *Social Work Education, 26*, 684-696.

Prieto, L. R. (1996). Group supervision: Still widely practiced but poorly understood. *Counselor Education and Supervision, 35*, 295-307.

*Raichelson, S.H., Herron, W.G., Primavea, L.H., & Ramirez, S.M. (1997). Incidence and effects of parallel process in psychotherapy supervision. *The Clinical Supervisor, 15*, 37 – 48.

Ramirez, N. (2003).Views towards organizational arrangements for ethnic-sensitive supervision in clinical settings serving Latino persons. *Journal of Ethnic & Cultural Diversity in Social Work: Innovation in Theory, Research & Practice, 12*, 1-18.

*Ramos-Sanchez, L., Esnil, E., Goodwin, A., Riggs, S., Touster, L.O., Wright, L.K., Ratanasiripong, P., & Rodolfa, E. (2002). Negative supervisory events: Effects on supervision and supervisory alliance. *Professional Psychology: Research and Practice, 33*, 197-202.

*Reese, R.J., Usher, E.L., Bowman, D.C., Norsworthy, L.A., Halstead, J.L., Rowlands, S.R., & Chisholm, R. R. (2009).Using client feedback in psychotherapy training:

An analysis of its influence on supervision and counselor self-efficacy. *Training and Education in Professional Psychology, 3,* 157-168.

*Riggs, S.A. & Bretz, K.M. (2006). Attachment processes in the supervisory relationship: An exploratory investigation. *Professional Psychology: Research and Practice, 37,* 558-566.

*Riva, M.T., Cornish, J.A., & Erickson (1995). Group supervision practices at psychology predoctoral internship programs: A national survey. *Professional Psychology: Research and Practice, 26,* 523-525.

*Romans, J.S. C., Boswell, D.L., Carlozzi, A.F., & Ferguson, D.B. (1995).Training and supervision practices in clinical, counseling, and school psychology programs. *Professional Psychology: Research and Practice, 26,* 407-412

Schroeder, M., Andrews, J. W. & Hindes, Y.L. (2009). Cross-racial supervision: Critical issues in the supervisory relationship. *Canadian Journal of Counselling, 43,* 295-310.

*Scott, K. J., Ingram, K. M., Vitanza, S. A., & Smith, N. G. (2000). Training in supervision: A survey of current practices. *The Counseling Psychologist,* 28, 403-422.

Schoenwald, S.K., Sheidow, A.J., & Chapman, Jason E. (2009). Clinical supervision in treatment transport: Effects on adherence and outcomes. *Journal of Consulting and Clinical Psychology, 77,* 410-421.

*Schultz, J.C., Ososkie, J.N., Fried, J.H., Nelson, R.E., & Bardos, A.N. (2002). Clinical supervision in public rehabilitation counseling settings. *Rehabilitation Counseling Bulletin, 45,* 213-222.

Sells, J. N., Goodyear, R. K., Lichtenberg, J. W., & Polkinghorne, D. E. (1997). Relationship of supervisor and trainee gender to in-session verbal behavior and ratings of trainee skills. *Journal of Counseling Psychology,* 44, 406-412.

Shechtman, Z. & Wirzberger, A. (1999). Needs and preferred style of supervision among Israeli school counselors at different stages of professional development. *Journal of Counseling & Development, 77,* 456-464.

Smaby, M.H., Smith, M.R., & Maddux, C.D. (2002). Counselor Education and Supervision: Quality of editorial board members' evaluations of manuscripts. *Counselor Education and Supervision, 41,* 259-267.

*Sterner, W.R. (2009). Influence of the supervisory working alliance on supervisee work satisfaction and work-related stress. *Journal of Mental Health Counseling, 31,* 249-263.

*Steven, D. T., Goodyear, R. K., & Robertson, P. (1998). Supervisor development: An exploratory study in changes in stance and emphasis. *Clinical Supervisor, 16,* 73-88

*Studer, J.R., Oberman, A. (2006). The Use of the ASCA National Model® in Supervision. *Professional School Counseling, 10,* 82-87.

Sumerel, M.B., & Borders, L.D. (1996). Addressing personal issues in supervision: Impact of counselors' experience level on various aspects of the supervisory relationship. *Counselor Education and Supervision, 35,* 268-286.

*Szymanski, D.M. (2005). Feminist identity and theories as correlates of feminist supervision practices. *The Counseling Psychologist, 33,* 729-747.

*Thielsen, V.A. & Leahy, M.J. (2001). Essential knowledge and skills for effective clinical supervision in rehabilitation counseling. *Rehabilitation Counseling Bulletin, 44,* 196-208

*Tromski-Klingshirn, D.M. & Davis, T. E. (2007). Supervisees' perceptions of their clinical supervision: A study of the dual role of clinical and administrative supervisor. *Counselor Education and Supervision, 46,* 294-304.

*Tyler, J.D., Sloan, L.L., & King, A.R. (2000). Psychotherapy supervision practices of academic faculty: A national survey. *Psychotherapy: Theory, Research, Practice, Training, 37,* 98-101.

*Utsey, S.O. & Gernat, C.A. (2002).White racial identity attitudes and the ego defense mechanisms used by White counselor trainees in racially provocative counseling situations. *Journal of Counseling & Development, 80,* 475-483.

Utsey, S. Gernat, C.A., Hammar, L. (2005). Examining white counselor trainees' reactions to racial issues in counseling and supervision dyads. *The Counseling Psychologist, 33,* 449-478.

Vespia, K.M., Heckman-Stone, C., & Delworth, U. (2002). Describing and facilitating effective supervision behavior in counseling trainees. *Psychotherapy: Theory, Research, Practice, Training, 39,* 56-65.

*Walker, J. A., Ladany, N., & Pate-Carolan, L. M. (2007). Gender-related events in psychotherapy supervision: Female trainee perspectives. *Counseling & Psychotherapy Research, 7,* 12-18.

*Wester, S.R., Vogel, D.L., & Archer, J., Jr. (2004). Male Restricted Emotionality and Counseling Supervision. *Journal of Counseling & Development, 82,* 91-98.

White, M.B. & Russell, C.S. (1995).The essential elements of supervisory systems: A modified Delphi study. *Journal of Marital and Family Therapy, 21,* 33-53.

White, M.B. & Russell, C.S. (1997). Examining the multifaceted notion of isomorphism in marriage and family therapy supervision: A quest for conceptual clarity. *Journal of Marital and Family Therapy, 23,* 315-333.

*Wilbur, M.P., Roberts-Wilbur, J., Hart, G.M., & Morris, J.R. (1994). Structured Group Supervision (SGS): A pilot study. *Counselor Education and Supervision, 33,* 262-279.

Yourman, D.B. & Farber, B.A. (1996). Nondisclosure and distortion in psychotherapy supervision. *Psychotherapy: Theory, Research, Practice, Training, 33,* 567-575.

Zabock, G., Drews, M., Bodansky, A., Dahme, B. (2009). The evaluation of supervision: Construction of brief questionnaires for the supervisor and the supervisee. *Psychotherapy Research, 19,* 194.

---

*article included in final sample of 62

Appendix B

Detailed Calculation Procedures for converting to eta squared (η^2)

- *Procedures for Computing η^2 Given a Statistic and Degrees of Freedom*
    - For the F statistic:

$$\eta^2(F, df1, df2) = \frac{F * df1}{F * df1 + df2}$$

    - For the t statistic: $\eta^2(t, df) = \frac{t^2}{t^2 + df}$
    - For correlation coefficients, r: $\eta^2(r) = \left(\frac{r}{1.253}\right)^2$ *
    - The chi-square test statistic must be converted to Cohen's effect size measure, w, before it can be converted to η^2.
        - To convert to w: $w(\chi^2, N) = \sqrt{\frac{\chi^2}{N}}$
        - Then to convert to η^2: $\eta^2(w) = \left(\frac{w}{1.253}\right)^2$ **
- *Procedures for Computing η^2 Given Cohen's other Effect Measures and Vice Versa*
    - To convert to Cohen's f: $f(\eta^2) = \sqrt{\frac{\eta^2}{1-\eta^2}}$
    - To convert to Cohen's f^2: $f^2(\eta^2) = \frac{(1.253*\eta)^2}{1-(1.253*\eta)^2}$
    - To convert from Cohen's f to Cohen's d: $d(f) = 2 * f$

Note: *Where 1.253 is a conversion factor to convert the bi-serial r measure to the point bi-serial since η^2 is a point bi-serial measure (η^2 is traditionally used for ANOVA tests which are equivalent to a point bi-serial linear regression). When a point bi-serial test was performed, the r was just squared and set equal to η^2.
**Where χ^2 is the value of the chi-square statistic and N is the sample size.

Appendix C

Test-Specific Procedures for Computing the non-centrality parameter (λ), η^2, and the MEs used by G*Power given test statistics and test-specific parameters

---

- *t-test of means: difference test between two dependent means (matched pairs)*

$$\eta^2 = \eta^2(t, df)$$

$$d(\eta^2) = d(f(\eta^2))$$

$$\lambda = d\sqrt{\frac{n1 * n2}{(n1 + n2)}}$$

- *Procedure for the t-test of means: difference between two independent means (two groups)*

$$\eta^2 = \eta^2(t, df)$$

$$d(\eta^2) = d(f(\eta^2))$$

$$\lambda = d\sqrt{\frac{n1 * n2}{(n1 + n2)}}$$

- *Procedure for the t-test of means: difference from constant (one sample)*

$$\eta^2 = \eta^2(t, df)$$

$$d(\eta^2) = d(f(\eta^2))$$

$$\lambda = d\sqrt{N}$$

- *Procedure for the Wilcoxon signed-rank test (one sample case)*

$$\eta^2 = \eta^2(t, df)$$

$$d(\eta^2) = d(f(\eta^2))$$

$$\lambda = \sqrt{0.955} * d\sqrt{N}$$

- *Procedure for the Wilcoxon signed-rank test (matched pairs)*

$$\eta^2 = \eta^2(t, df)$$

$$d(\eta^2) = d(f(\eta^2))$$

$$\lambda = \sqrt{0.955} * d \sqrt{\frac{n1 * n2}{(n1 + n2)}}$$

- *Procedure for the Wilcoxon-Mann-Whitney test (two groups)*

$$\eta^2 = \eta^2(t, df)$$

$$d(\eta^2) = d(f(\eta^2))$$

$$\lambda = \sqrt{0.955} * d \sqrt{\frac{n1 * n2}{(n1 + n2)}}$$

- *Procedure for the Correlation: simple bivariate r*

$$\eta^2 = \eta^2(r)$$

Cohen's r = r

$$\lambda = \sqrt{\frac{Nr^2}{1 - r^2}}$$

- *Procedure for the Correlation: point biserial*

$$\eta^2 = r^2$$

Cohen's r = r

$$\lambda = \sqrt{\frac{Nr^2}{1 - r^2}}$$

- *Procedure for the Rank correlations*

$$\eta^2 = \eta^2(r)$$

Cohen's r = r

$$\lambda = \sqrt{\frac{Nr^2}{1 - r^2}}$$

- *Procedure for the Linear multiple regression: Fixed model, R^2 deviation from zero*

$$\eta^2 = \eta^2(F, df1, df2)$$

$$f^2 = f^2(\eta^2)$$

$$\lambda = N\ f^2$$

- *Procedure for the Linear multiple regression: Fixed model, R^2 increase*

$$\eta^2 = \eta^2(F, df1, df2)$$

$$f^2 = f^2(\eta^2)$$

$$\lambda = N\ f^2$$

- *Procedure for the Multivariate multiple regression*

$$\eta^2 = \eta^2(F, df1, df2)$$

$$f^2 = f^2(\eta^2)$$

$$\lambda = sN\ f^2 \quad *$$

- *Procedure for the Canonical correlation – Using approx F statistic*

$$\eta^2 = \eta^2(F, df1, df2)$$

$$f^2 = f^2(\eta^2)$$

$$\lambda = sN\ f^2$$

- *Procedure for the ANCOVA: main effects and interactions*

$$\eta^2 = \eta^2(F, df1, df2)$$

$$f = f(\eta^2)$$

$$\lambda = N\ f^2$$

- *Procedure for the 1-way ANOVA*

$$\eta^2 = \eta^2(F, df1, df2)$$

$$f = f(\eta^2)$$

$$\lambda = N f^2$$

- *Procedure for the n-way ANOVA: main effects and interactions*

$$\eta^2 = \eta^2(F, df1, df2)$$

$$f = f(\eta^2)$$

$$\lambda = N f^2$$

- *Procedure for the Kruskal-Wallis Test (non-parametric ANOVA, uses chi-square test of proportions)*

$$\eta^2 = \eta^2(w)$$

$$w(\chi^2, N) = \sqrt{\frac{\chi^2}{N}}$$

$$\lambda = Nw^2$$

- *Procedures for the MANCOVA and MANOVA: Global effects, Special effects, and Interactions*

$$\eta^2 = \eta^2(F, df1, df2)$$

$$f^2 = f^2(\eta^2)$$

$$\lambda = sN f^2 \ **$$

- *Procedure for the Repeated measures: between interaction - univariate approach and MANOVA approaches*

$$\eta^2 = \eta^2(F, df1, df2)$$

$$f = f(\eta^2)$$

$$\lambda = \sqrt{\frac{Nmf^2}{1+(m-1)*\rho}} \qquad ***$$

- *Procedure for the Repeated measures: within interactions and within-between interactions, univariate and MANOVA approaches*

$$\eta^2 = \eta^2(F, df1, df2)$$

$$f = f(\eta^2)$$

$$\lambda = \sqrt{\frac{Nmf^2}{1-\rho}} \quad ****$$

- *Procedures for the Chi-square test of independence equality of proportions, Chi-square goodness of fit test, and the Canonical correlation (using approximate Chi-Square Statistic)*

$$\eta^2 = \eta^2(w)$$

$$w(\chi^2, N) = \sqrt{\frac{\chi^2}{N}}$$

$$\lambda = Nw^2$$

Note: *s = min{number of dependent variables, df1 + 1}
**s = min{number of dependent variables, df1}
***Where m is the number of repeated measures and ρ is the bivariate correlation between the measures.
****Where m is the number of repeated measures and ρ is the bivariate correlation between the measures.

Appendix D

Coding Chart

CODER		
AUTHOR/TITLE		
ABSTRACT/SUMMARY		
HYPOTHESIS(ES)		
STATED PURPOSE		
STATED HYPOTHESES		
RESEARCH QUESTIONS		
Sample description		
Type of Design		
Limitations acknowledged?		
Threat? Y/N	**Hypothesis Validity threats**	
	1. Does hypothesis ask a critical question? (Inconsequential hypotheses)	
	2. Are there multiple hypoth to reduce risk of inconsequential hypotheses?	
	3. Is the hypothesis clear? (Ambiguous hypotheses)	
	4. Does the statistical hypoth (null and alternate hypoth) correspond to the research hypoth? (Noncongruence of research and statistical hypothesis)	
	5. Are multiple tests used to test the hypothesis? (Diffuse statistical hypotheses and tests)	
Threat? Y/N	**Russell et al. (1984)'s Methodological Threats**	
	1. Is there an adequate comparison group?	
	2. Is there a pretreatment assessment?	
	3. Is there an adequate sample size?	
	4. Variations or confounds in length of training across conditions	
	5. random assignment of participants to conditions	

	6. Widely discrepant cell sizes	
	7. Restricted range of dependent variables	
	8. Representative supervisee or supervisor sample?	
	9. Is there a follow-up assessment?	
	10. Use of role play or audiotaped client statements to assess supervised change?	
	11. Exclusive reliance on self-report data?	
	12. Overly brief training period?	
Threat? Y/N	**Threats to Internal Validity**	
	1. History-anything occur btwn pre & posttest?	
	2. Maturation—or gain experience btwn testing?	
	3. Testing—did they become familiar with tests given multiple times?	
	4. Instrumentation ceiling/floor effects?	
	5. Statistical regression—poss regression to the mean?	
	6. Mortality—any drop outs?	
	7. Interactions with selection of sample & other threats?	
	8. Ambiguity about the direction of causal influence btwn dep & indep variables?	
	9. Diffusion of treatments –did control grp learn about experimental group?	
	10. compensatory rivalry by respondents receiving less desirable treatments—did supervisors attempt to equalize tx?	
	11. Resentful demoralization of respondents receiving less desirable treatment?	
Threat? Y/N	**Threats to External Validity**	
	1. Interaction of selection and treatment?	
	2. Interaction of setting and treatment?	
	3. Interaction of history and treatment?	
Threat? Y/N	**Threats to Construct Validity**	
	1. Inadequate preoperational explication of constructs	

	2. Mono-operation bias—only 1 operation?	
	3. Monomethod bias—only one method?	
	4. Hypothesis guessing within experimental conditions	
	5. Evaluation apprehension/ socially desirable responding?	
	6. Experimenter expectancies—were raters biased by own expectations?	
	7. Confounding of constructs and levels of constructs	
	8. Interaction of different treatments	
	9. Interaction of testing and treatment	
	10. Restricted generalizability across constructs? I.e. not enough constructs were affected by tx	
Threat? Y/N	**Statistical Conclusion Validity (Methodological threat)**	
	1. low statistical power	
	2. violation of assumptions of statistical tests—do the tests assume that the sample is normally distributed, or that they can freely say what they want, etc.?	
	3. Type I error—do the authors say there is no phenomena when there actually may be?	
	4. unreliability of measures	
	5. unreliable treatment implementation	
	6. random irrelevancies in the experimental setting	
	7. random heterogeneity of respondents—some of the variety in the sample may be related to the phenomena under investigation, but at least part is likely to just to constitute individual differences that are irrelevant to the relationship being observed.	

Appendix E

Vita

EDUCATION

Ph.D., Counseling Psychology, May 2012
Lehigh University, Bethlehem, PA
American Psychological Association-approved program

Dissertation topic: Meta-analysis, power study, and qualitative examination of published research of psychotherapy supervision

Qualifying project: Development and validation of a multicultural counseling competency checklist for counselor training programs

M.Ed., Counseling & Human Services, May, 2001
Lehigh University, Bethlehem, PA

BA, Psychology, May, 1997
Temple University, Philadelphia, PA

PRE-DOCTORAL INTERNSHIP

Allegheny General Hospital Pittsburgh, PA July 2005 to July 2006
Completion with honors
The AGH internship consisted of three major four-month long rotations in Adult, Child/Adolescent, and Neuropsychology, and one intern-specific year-long minor rotation. The Adult rotation included a) facilitation of process-oriented therapy groups in an intensive outpatient/partial hospital setting and b) carrying a caseload of individual outpatient clients for both short and long-term therapy c) coordination of services and case management as needed/appropriate. The Child/Adolescent rotation included a) outpatient individual therapy with children and adolescents b) coordination of auxiliary and support services c) case management as needed/appropriate. The Neuropsychology rotation included a) administration and interpretation of neuropsychological assessment batteries to inpatient children/adolescent, adult, and geriatric population post closed head injury b) administration and interpretation of full outpatient neuropsychological assessment batteries to children/adolescent, adult, and geriatric populations representing a range of referral questions including learning disorders, attention, cognitive abilities, memory, and executive functioning.

The minor rotation was divided in two six month sections and interns could select these. My first minor was spent with The Center for Traumatic Stress at AGH. Responsibilities included initial intake, evaluation, and outpatient therapy for children and adolescent victims of trauma. Treatment included family work, contact with support and auxiliary organizations (such as the Department of Children, Youth, and Families and Family-

based programs), and utilization of a trauma intervention treatment program. My second minor was an extension of the Neuropsychology rotation required of all interns. I chose to receive additional training in neuropsychology assessment, which was in the form of outpatient testing, assessment, report writing, and follow up with patients.

Additional responsibilities of the internship included: providing lectures to medical students, participating in journal club, attending and presenting at Grand Rounds, attending 105 hours of didactic training, receiving 6 hours of supervision per week, and engaging in case presentations.

PRACTICUM EXPERIENCE

Friends Hospital August 2004 to June 2005
Philadelphia, PA
Conducted short and long-term individual counseling, group therapy, crisis assessment and intervention, and psychological testing with adult and older adult clients with chronic mental illness in a residential placement. Receive one hour of supervision and two hours of training per week.

Veterans Administration August 2003 to May 2004
Allentown, PA
Provided mental health treatment for veterans of Vietnam, Korea, and Gulf Wars, including: provided short and long-term individual counseling, conducted intake interviews, provided crisis intervention, co-facilitated process-oriented group therapy, and co-facilitated smoking cessation group. Received one hour of individual supervision per week on-site.

Lehigh University Counseling Center August 2002 to May 2003
Bethlehem, PA
Provided counseling services to young adults in a college population. Conducted short and long-term individual counseling, couples counseling, intake interviews, psychological testing; co-facilitated process-oriented group therapy for undergraduate women and a support group for those with loved ones in the war. Provided outreach to university students regarding eating disorders and sexual assault. Supervision included one hour of individual, two hours of group, and one hour of group therapy facilitation supervision per week; Conducted experiential Diversity Training for Residential Advisors.

Kutztown University Counseling Center August 1999 to May 2000
Kutztown, PA
Master's practicum: Provided counseling services to young adults in a college population. Conducted short and long-term individual counseling, intake interviews, and crisis mediation/intervention. Received one hour on-site supervision and one hour off-site supervision weekly.

CLINICAL EXPERIENCE

Sayegh Pediatric Therapy Services. P.C. 2010 to present
Whitehall, PA
Position: Psychological Consultant
Responsibilities: Provide consultation and education to treatment group consisting of occupational, physical, and speech therapists and special educators treating children age birth to 3 enrolled in Early Intervention and children ages 6 to 21 enrolled in cyber schools. Consultation includes case review with therapists and recommendations of strategies for behavioral management and support of functional emotional development, as well as education regarding psychological and behavioral strategies for addressing specific behavioral or emotional presentations. Also provide recommendations for further testing, evaluation, and treatment, including consideration of higher levels of care such as wrap around, respite, or even alternative placement. Provide education and guidance about confidentiality, ethics, and accessing county services.

Family Psychological Associates 2005 to 2008
Kittanning, PA
Position: Postdoctoral Clinician, Supervisor
Responsibilities: Provided outpatient individual and family therapy to children, adolescents, and adults of managed care population. Conducted psychological evaluations to ascertain appropriate recommendations for behavioral rehabilitation services. Provided supervision to master's counseling students and masters-level therapists.

Variety Club Camp and Developmental Center Summer 2004
Worcester, PA
Position: Behavioral Consultant/Program Supervisor
Responsibilities: Designed and conducted emotional wellness program in summer camp for emotionally, physically, and developmentally challenged children and adolescents ages 6 through 21. Program centered on self-esteem, coping skills, anger management, communication, self-efficacy, and stress management. Supervised and facilitated delivery of program, provided counseling, crisis intervention, consultation, and behavioral analysis/intervention.

Allentown School District 2001 to 2004
Allentown, PA
Position: Clinical Supervisor/Counselor
Responsibilities: Provided clinical supervision to practicum students in a school counseling masters' degree program. Supervision included the following: provision of one weekly individual supervision and weekly group supervision, review of audiotapes of supervisee's counseling sessions, provision of written and verbal feedback, and on-site support and supervision. Met regularly with professors, on-site school counselor supervisors, and Allentown School District administrators regarding supervisees' progress in the practicum. Conducted individual and group counseling, crisis intervention, and assessment with the children in the school district (ages 6 through 14).

KidsPeace Center for Kids in Crisis 2001 to 2003
Orefield, PA
Position: Clinical Therapist
Responsibilities: Provided treatment to culturally-diverse populations of adolescents living in a long-term acute residential setting; Conducted individual, group, and family therapy; Created treatment plans with input from psychiatrists and staff, maintained regular communication with county offices and insurance representatives.

KidsPeace Center for Kids in Crisis 2000 to 2001
Orefield, PA
Position: Assistant Treatment Team Supervisor
Responsibilities: Provided supervision to residential staff and treatment to a culturally diverse, inner-city group of adolescent girls in an acute residential setting; Supervised a staff of 12 childcare counselors, created and maintained treatment plans for all clients, facilitated hospitalization as necessary, conducted individual and group counseling, provided crisis intervention, met with social workers and psychiatrists to review treatment, and communicated with county offices and insurance representatives; hired staff; administered medication to clients.

The Westmeade Center 1998 to 2000
Hartsville, PA
Position: Milieu Counselor
Responsibilities: Created and facilitated process and psychoeducational groups for adolescents, including anger management, family issues, and a sexual abuse survivors group for girls. Provided individual and crisis intervention and counseling.

Milestones Community Healthcare 1998 to 1999
Glenside, PA
Position: Assistant Program Director
Responsibilities: Provided leadership support for three residential homes for CHIPPS consumers from Norristown and Allentown State Hospitals; Created and supervised delivery of program schedule; Provided clinical leadership to support staff; hired and scheduled staff; Worked with Bucks County MH/MR in transitioning clients from the hospital.

Penn Foundation 1997 to 1998
Sellersville, PA
Position: Case Manager
Responsibilities: Coordinated mental health services for mental health consumers; assisted consumers in accessing benefits, including SSI, SSD, medical/cash assistance, food stamps, low income housing, and indigency pharmaceuticals; acted as liaison between organizations for psycho-social rehabilitation and vocational training services; provided counseling at medication clinic; monitored compliance of outpatient involuntary commitments; acted as community hospital liaison and acted as gatekeeper to state hospitals by evaluating referrals from community hospital.

St. Luke's Hospital 1997 to 1999
Quakertown, PA
Position: Crisis Worker
Responsibilities: Evaluated Emergency Room patients for suitability of admission to inpatient unit or potential of referral to outpatient treatment; discussed all cases with the on-call psychiatrist and ER physician and completed admission to the inpatient unit if indicated. Attained pre-certification from insurance companies and reviewed cases regularly with insurance providers. Pursued/completed involuntary petitions and commitments.

TEACHING EXPERIENCE

Kenyatta University, Kenya December 2000; course development through 2001
Instructor
Project Supervisor: Dr. Muugi
Counseling and Therapeutic Techniques graduate course
Co-created and co-taught a graduate-level counseling skills course at Kenyatta University in Kenya. Course was a 2-week long part of an emergency certification program for HIV/AIDS counselors. Content of the course centered on introductory counseling, helping skills, and therapeutic techniques for use with HIV/AIDS clients and their families.

Lehigh University, PA January 2002 to May 2002
Teaching Assistant
Professor: Dr. Tina Q. Richardson
Standardized Tests, Measurement, and Appraisal graduate course
Conducted biweekly labs to teach clinical assessment and interviewing skills. Students were required to videotape role-plays clients and counselors, for which I provided verbal and written feedback. I also was responsible for grading tests, papers, and projects.

Lehigh University, PA Summer 2003
Lecturer
Professor: Carl Persing, MS
Industrial/Organizational Psychology undergraduate course
Taught multicultural issues modules; lectures focused on dimensions of culture as it affects interaction in the workplace. Discussion included the following: race and racial identity, sexual orientation, gender, SES, implications of worldview (Euro-American) and value orientations, and dealing with ethnocentrism and factors influencing interactions with those of different cultures.

Lehigh University, PA Summer 2004
Lecturer
Professor: Carl Persing, MS
Industrial/Organizational Psychology undergraduate course
Taught multicultural issues modules; lectures focused on dimensions of culture as it affects interaction in the workplace. Discussion included the following: race and racial

identity, sexual orientation, gender, SES, implications of worldview (Euro-American) and value orientations, and dealing with ethnocentrism and factors influencing interactions with those of different cultures.

Lehigh University, PA Fall 2004
Teaching Assistant
Professor: Dr. Colleen McDonough
Child Development undergraduate course
Met weekly with students individually and in groups to review and teach course materials and teach study skills; Created exam questions, administer and score exams, and conduct review sessions.

Lehigh University, PA Spring 2005
Teaching Assistant
Professor: Linda Dench, MA, ABD
Child Development undergraduate course
Met weekly with students individually and in groups to review and teach course materials and teach study skills; Created exams, administer and score exams, and conduct review sessions.

RESEARCH EXPERIENCE

Principal Researcher 2010 to 2012
Dissertation: Replication and Extension of Ellis et al. (1996); Meta-analysis and power study of supervision research from 1994 through 2010
Advisor: Dr. Arnold Spokane, Lehigh University

Principal Researcher 2007 to 2009
The Impact of Supervision and Training on the Development of a Counselor Trainee: A Case Study
Advisor: Dr. Nicholas Ladany, Lehigh University
The study followed a large-scale research project of four counselor trainees, four supervisors, and sixteen actual counseling clients at a university counseling clinic; utilized qualitative methodology to examine data of one trainee-supervisor dyad.

Principal Researcher 2001 to 2006
Development of a Multicultural Competency Checklist for Counselor Training Programs
Advisor: Dr. Tina Q. Richardson, Lehigh University
Design, creation, and validation of a measure for assessing counselor multicultural competence; construction of web-based assessment; conducted initial validation research with 193 graduate student participants nationwide, using the Miville Universality-Diversity Scale, Counselor Self-Efficacy Scale, and Marlowe-Crowne Social Desirability Scale.

Core Research Team Member 2001to 2002
Positive Attitudes toward Gay Men: A Qualitative Investigation of Heterosexual Allies

Principal Researcher: Kevin Castro-Convers, M.Ed., Lehigh University
Transcribed interviews with heterosexual allies of gay men; Worked with team members to code data and conduct qualitative analysis of participant interviews using CQR method;.

Core Research Team Member 2001-2002
Trainee Learning: An Exploratory Investigation into Experiences of Counselor Trainees through Practice and Supervision.
Principle Researcher: Katja Spradlin, M.Ed., Lehigh University
Conducted and transcribed interviews with counselor trainees; conducted qualitative analysis of participant interviews using Discovery-Oriented method.

Core Research Team Member 2001-2002
Supervisors' and Trainees' Perceptions of Helpful and Hindering Events in Supervision
Principal Researcher: Laurie Gray, M.Ed.
Conducted and transcribed interviews with counselor trainees; conducted qualitative analysis of participant interviews using Discovery-Oriented method and analyzed data.

Core Research Team Member 2000-2001
Parallel Process in Supervision and Psychotherapy
Principle Researcher: Laurie Gray, M.Ed., Lehigh University
Conducted and transcribed interviews with counselor trainees, supervisors, and clients; conducted qualitative analysis of participant interviews using Discovery-Oriented method and analyzed data.

Research Team Member 2000 to 2001
Psychotherapy and Supervision Research Project: *Assessing Counselor Trainee Development*
Primary Researcher: Dr. Nicholas Ladany, Lehigh University
Conducted pre and post-therapy session interviews of clients and therapists; conducted pre and post supervision interviews of trainees and supervisors; administered instruments, including the Trainee Anxiety Scale OQ-45, Supervisor attitudes scale, and supervisor, client, and therapist Working Alliance Inventory; responsible for audio-visual recording.

Research Assistant 1998 to 1999
Study of the use of silence in therapy
Primary Researcher: Dr. Nicholas Ladany, Lehigh University
Transcribed qualitative interviews of practicing psychologists regarding use of silence in therapy.

INVITED PRESENTATIONS

School Counselors Assoc. Leadership Development Academy August, 2003
Villanova University
Multicultural Competency in Counselor-Educator Leadership

Conducted presentation and training on multicultural competent leadership for school counselor and psychologists; focus on multicultural competency in supervision, professional leadership, team building, communication, advocacy, and interaction with peers, students, families, and communities; presentation of worldview orientations, personal dimensions, and racial identity models.

School Counselor Training in Multicultural Competence August 2002; 2003
Allentown School District
Conducted presentation and training on multicultural competence in school counseling; focus on multicultural competency in counseling interventions, communication, supervision, and interaction with peers, students, families, and communities.

PRESENTATIONS

Schutt, M., & Richardson, T.Q. (2004). *Tools for Increasing Awareness and Developing Multicultural Competence in Counselor Training Programs.* Paper presentation at the 21st Annual Teachers College Winter Roundtable on Cultural Psychology and Education: Strategies for Building Cultural Competence in Psychology and Education.

Castro-Convers, K., Metzler, A., Kelly, J., Rothermel, C., Schutt, M., & Walker, J. (2004). *A Qualitative Investigation of Heterosexual Allies of Gay Men.* Paper presentation at the Annual Meeting of the American Psychological Association, Hawaii.

Schutt, M., & Richardson, T.Q. (2002). *Application of a Multicultural Counseling Competency Checklist to High-Stakes Testing*. Paper presentation at the Second Annual Diversity Challenge at Boston College, through the Institute for the Study and Promotion of Race and Culture, Boston, Massachusetts.

Castro-Convers, K., Metzler, A., Kelly, J., Rothermel, C., Schutt, M., & Walker, J. (2002). *Positive Attitudes toward Gay Men: A Qualitative Investigation of Heterosexual Allies*. Poster presentation at the Second Annual Meeting of the Mid-Atlantic Society for Psychotherapy Research, University Park, Pennsylvania.

Gray, L.G, Schutt, M., & Spradlin, K. (2001). *Parallel Process in Supervision and Psychotherapy*. Paper presented at the Annual Meeting of the Society for Psychotherapy Research, Montevideo, Uruguay.

Gray, L.G. & Schutt, M. (2001). *Parallel Process in Supervision and Psychotherapy*. Paper presented at the Annual Meeting of the American Psychological Association, San Francisco, California.

ASSESSMENT TRAINING

Attention Deficit Disorders Evaluation Scale; Behavior Rating Inventory of Executive Function (BRIEF-SR); Verbal Learning (CAVLT-2); Child Behavior Checklist (CBCL); Child Depression Inventory (CDI); Connors' Continuous Performance Test-II; Delis

Kaplan Executive Function System (D-KEFS); BAADS House-Tree-Person (HTP); Minnesota Multiphasic Personality Inventory (MMPI-2); Millon Clinical Multiaxial Inventories (MCMI-3); Millon Index of Personality Styles (MIPS); Rorschach; Thematic Apperception Test (TAT); Strong Interest Inventory (SII); Trail Making Test, Verbal Fluency Test; Revised Children's Manifest Anxiety Scale (MASC); Wechsler Intelligence Scale for Children, Fourth Edition (WISC-IV); Wechsler Adult Intelligence Scale (WAIS-III); Wisconsin Card Sorting Test