2013

The Effect Of Instrument-Specific Rater Training On Interrater Reliability And Counseling Skills Performance Differentiation

Paul Douglas Meacham
University of Mississippi

THE EFFECT OF INSTRUMENT-SPECIFIC RATER TRAINING ON INTERRATER RELIABILITY AND COUNSELING SKILLS PERFORMANCE DIFFERENTIATION

A Dissertation

presented as partial fulfillment of the requirements

for the degree of Doctor of Philosophy

in the Department of Leadership and Counselor Education

The University of Mississippi

by

Paul Douglas Meacham, Jr.

May 2013

ABSTRACT

The purpose of this study was to explore the effect of instrument-specific rater training on interrater reliability (IRR) and counseling skills performance differentiation. Strong IRR is of primary concern to effective program evaluation (McCullough, Kuhn, Andrews, Valen, Hatch, & Osimo, 2003; Schanche, Nielsen, McCullough, Valen, & Mykletun, 2010) and counselor education (Baker, Daniels, & Greeley, 1990; Jennings, Goh, Skovholt, & Banerje-Steevens, 2003; Lepkowski, Packman, Smaby, & Maddux, 2009). The ability to differentiate between low and high performances of counseling skills is central to informing the classroom instruction of counseling students and the supervision of early clinical experiences (Byrne & Hartley, 2010; Fitch, Gillam, & Baltimore, 2004; Paladino, Barrio-Minton, & Kern, 2011). Participants were randomly assigned to one of four groups defined by whether they received instrument-specific training and the performance level of the counseling skills they assessed. Data was collected using the Universal Counseling Skills Assessment (UCSA) administered traditionally and through the Dynamic Scoring Interface (DSI). The researcher used a 2 X 2 factorial ANOVA, independent samples *t*-tests, intraclass correlation coefficients, and Fisher's *r* to *z* transformations to analyze the data's validity across the groups and reliability within the groups. Results that brief instrument-specific training and a structure scoring procedure can significantly strengthen IRR. The results of the analyses are discussed within the context of their implications for counselor education and future research possibilities.

DEDICATION

To April, my wife, who believed this was possible long before there was any evidence to warrant such a faith.

and

To Paul and Nydia Meacham, my parents, who have been living examples of what it means to be humble before God and generous to man.

ACKNOWLEDGEMENTS

The list of those who have contributed to this effort is extensive, but some have given of themselves so unselfishly that they must be recognized formally. Dr. Kevin Stoltz began as my professor, became my image of a professional counselor, generously gave of himself as my mentor, chaired my committee with exemplary professionalism, and in the process became my dear friend. His impact on me as a counselor and counselor educator is incalculable, but he has also made a profound impression on me as man. Considering all he has done for me, to say "thank you" seems woefully inadequate. If I can have the kind of impact on my students that Dr. Stoltz has had on me, I will know that I have been a success.

I offer my sincerest gratitude to my committee, Drs. Suzanne Degges, Mathew Reysen, and Marc Showalter. I was advised that the choice of dissertation committee members would make the difference between having an enduring, rewarding experience and having to endure a special kind of hell on earth. Each of my committee members was a paragon of professionalism. They were prompt in their communications, candid and constructive in their feedback, cooperative with each other and with me, and generous with their time and encouragement.

Dr. Degges possesses the seemingly incongruent ability to be wonderfully encouraging while simultaneously dissecting your writing in a way that reveals every weakness. I am grateful for her keen eye, her kind heart, and her dedication to make better whatever comes across her desk.

Dr. Matthew Reysen is one of those rare educators who genuinely enjoys teaching and cares about his students. His advice and counsel were invaluable in helping me make the career

decision that I expect will shape the rest of my life. His service on this committee and his contributions to my research mean more to me than I can express.

I must make special mention of Dr. Marc Showalter, under whose authentic and insightful supervision I worked as a counselor for the past five years. Dr. Showalter has created a great incubator for developing counselors. Under his direction, the University Counseling Center does a wonderful job meeting the counseling needs of the university community, while simultaneously supporting and nurturing beginning counselors. If I ever direct a clinic, it will be his example to which I aspire.

I am also grateful to Dr. Lori Wolff. Not only did I benefit greatly from being a student in her classes, but she accepted me as a teaching assistant and dedicated herself to ensuring that I learned as much about being a teacher as her students learned about statistics. I will always think of her as the template when it comes to an educator's preparation, organization, and efficient use of time.

I am thankful for all the encouragement I have received from my classmates and colleagues. Of special note are the contributions of Dr. Keysha Thomas, Dr. Joshua Magruder, and my brother, Cason Pearce. I will always remember that they believed enough in me and my research ideas that they were willing to invest themselves actively without ever asking, "What's in it for me." I could not have completed this study without their assistance.

My years in graduate studies were by three wonderful professionals in the forms of Shelia Goolsby, the senior executive assistant at the University Counseling Center, and Kim Chrestman and Michelle Wallace, the executive assistants for the Department of Leadership and Counselor Development. These wonderful ladies are the heart and soul of their respective departments and the kind of friends who enrich your life. I have turned to them in many times of crisis and

momentary despair, and they were always ready to listen and, when possible, to help. Not once, even when it would have been right to do so, did any of them ever say, "That's not my job."

One of the keys to accomplishing anything significant is to surround yourself with great, accomplished people. The computer genius behind the *Dynamic Scoring Interface* was supplied by Joey Davis. In addition to being the best and truest of friends, he possesses a mystical ability to speak to computers in languages they understand and to make them do his bidding. I am grateful to Joey for many things far more significant than this dissertation, but without his contribution to this study it would still be just an idea in the mind of a frantic doctoral student.

Finally, I offer my thanks to Dr. Jerry Martin. Dr. Martin was the first to show me the value of a counselor to a soul seeking help. It was he who first caused me to consider how I might be able to help others as a counselor. His compassion for people and his desire to be an influence for eternal good are constant encouragers and motivators to me.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I: INTRODUCTION

To sit in judgment over any part of another person's life is, and should be, a sobering experience. Yet, professional responsibilities impose upon counselor educators and clinical supervisors of counselors just such a role. Both the American Counseling Association (ACA) Code of Ethics (2005) and the Council for the Accreditation of Counseling and Related Educational Programs (CACREP, 2009) standards require counselor educators to provide students with systematic formative and summative feedback, implying the need for ongoing evaluations of students' academic and clinical performances.

Practitioners are also expected to make judgments about their supervisees' performances. This is evidenced by the fact that most if not all supervision models identify *evaluation* as one of a supervisor's primary functions. Ranging from the plainly stated position of Cognitive-Behavioral supervision—"The purpose of supervision is to teach appropriate therapist behaviors and extinguish inappropriate behavior" (Boyd, 1978, p. 89)—to the gentler more collaborative Constructivist-Narrative approach, which identifies the supervisor as the "editor or catalyst" (Bernard & Goodyear, 2009, p. 87), all of the Psychotherapy-based models note that the supervisor functions as an evaluator of the supervisee. Even the developmental models, which are conceptualizations of supervisee progress more than frameworks for doing supervision (Falenders et al., 2004), speak of "supervisors assessing their supervisees' level of functioning" (Bernard & Goodyear, 2009, p. 90) and provided instrumentation for doing so (McNeill,

Stoltenberg, & Romans, 1992). The Social Roles Models of supervision, however, most plainly speak of the supervisor as an evaluator (Carroll, 1996; Hess, 1980; Holloway, 1995; Williams, 1995). The most popular of the Social Roles Models, the Discrimination Model (Bernard, 1979), does not explicitly list *evaluator* as one of a supervisor's roles, but this model does base all three of its roles, *teacher*, *counselor*, and *consultant*, on the foundation of having first "made a judgment about their supervisee's abilities" (Bernard & Goodyear, 2009, p. 102). There is, therefore, a clear expectation of and responsibility on counselor educators and clinical supervisors to evaluate their students and supervisees.

Counselor educators and clinical supervisors meet their evaluative responsibilities through a number of techniques. Students' counseling knowledge is assessed through testing, writing assignments, and oral presentations. Counseling skills, however, are most commonly evaluated by reviewing video recordings of mock or real counseling sessions and scoring the demonstrated performances with a formal instrument (Freeman & McHenry, 1996; CACREP, 2009). Studies (Daniels & Larson, 1992; Steward, Breland, & Neil, 2001) have shown that feedback based on video review is a powerful tool having a profound impact on counseling students' levels of self-esteem and feelings of professional self-efficacy. Social cognitive theory explains this relationship between performance feedback and feelings of self-efficacy by identifying feedback as the social metric against which beginning counselors measure themselves (Bandura, 1991, 1997). Counselor educators and supervisors need to be aware of the potential power of their feedback given that the combination of an external locus of evaluation and negative or disappointing feedback can raise students' anxiety to levels that have been shown to be deleterious to counseling performance (Hiebert, Uhlemann, Marshall, & Lee, 1998; Ronnestad & Skovholt, 1993).

Considering the potential influence of evaluative feedback over beginning counselors, it behooves counselor educators and supervisors to ensure the feedback they provide is accurate and reliable. Unfortunately, this standard is not always met. Fitch, Gillam, and Baltimore (2004) found that even expert raters, defined as counselor educators or Master's level practitioners certified to supervise, can vary widely in the scoring of the most basic counseling skills. The study's data demonstrated that within the group of participant raters, some assigned failing scores while others assigned nearly perfect performance scores to the same video. A challenge exists, therefore, for supervisors to perform accurate and reliable assessments of beginning counselors' skills and to provide reliable feedback that can shape the continuing development of effective counseling skills. The purpose of this study is to explore the effect of brief instrument-specific rater training on the reliability and accuracy of counseling skills assessment when said assessment is conducted by video review.

Technology, Instrument Development, and Skills Assessment

Two important factors have served to facilitate the proliferation of assessing counseling skills through the scoring of video recordings. First, technological advancements have made the video recording of sessions a simple and common procedure. What once required the use of elaborate commercial equipment and professionally trained technicians and operators (Kagan, Krathwohl, & Miller, 1963) can now be accomplished by counselors with no technical training and a pocket-sized digital recorder (Byrne & Hartley, 2010). Today's counseling students' ubiquitous familiarity with modern technology has changed the shape of the profession. Every counselor educator and clinical supervisor can reasonably expect trainees to be able to provide

high quality video recordings for assessment and supervision purposes. Additionally, CACREP (2009) requires that students' practicum experience includes either live supervision or supervision using session recording (Section III.F.4).

Just as technology has placed the power of video recording into everyone's hand, technological advancements in desktop publishing and inexpensive high quality printers have made it possible for anyone to produce their own professional looking instruments. These advances coupled with the current emphasis on measuring learning and counseling outcomes (Schanche, Nielsen, McCullough, Valen, and Mykletun, 2010) prompted a late 20th century/early 21st century wave of new assessment instruments.

These new instruments are by no means the first attempts to measure counseling skills. In fact, the development of instruments to measure counseling skills coincides with the earliest efforts to teach discrete therapeutic skills. Though Rogers (1957) did not emphasize counseling skills, his concentration on the *facilitative conditions* that bring about a *necessary and sufficient therapeutic relationship* began the genre of model-specific companion instruments. Among these were the Relationship Inventory (Barrett-Lennard, 1962), the Working Alliance Inventory (Horvath, 1981), and the Therapeutic Bond Scale (Saunders, Howard, & Orlinsky, 1989). However, a distinction should be made between the assessments associated with Rogers' person-centered model and those that followed. Because Rogers did not deconstruct the facilitative conditions and define their elements, instruments associated with Rogers' model were more concerned with the degree to which a therapist possessed a particular facilitative characteristic (e.g. congruence, positive regard, or empathy) rather than with observable skills that demonstrated those constructs. This method of assessment is rendered unsatisfactory by today's

demand for outcome based evaluation of behaviors, but it did pave the way for the observable skills assessments that would follow.

Carkhuff, through his Human Resource Development model (Carkhuff, 1969a, 1969b, 2000, 2009; Truax, Carkhuff, & Dounds, 1964), expanded upon Rogers work by operationalizing the empathy construct. The concept that a construct like empathy could be observed and measured through the behaviors that conveyed it promptly brought about the Carkhuff Empathy Scale (1969a, 1969b) so that empathy might be measured. Ivey's introduction of the Microskills model (Allen, 1967; Forsyth & Ivey, 1980; Ivey & Authier, 1971) was predicated on the view of skills being discrete behaviors (observable and measurable) to be learned individually and then integrated into a whole. Microskills' wide acceptance gave birth to a number of instruments including the Counselor Effectiveness Scale (Ivey, 1968) and the Counselling Interview Rating Form (Russell-Chapin & Sherman, 2000). While many of the studies using instruments associated with Carkhuff's and Ivey's models demonstrated the reliability and validity of these instruments, such support was always limited to the context of scoring the performances of students trained according to their relative methods. No attempts were made to establish cross modality validity.

Modern models of skills instruction have continued the trend of producing accompanying instrumentation. Young (1998, 2009) borrowed Ivey's (1968) microskills and suggested that they be presented within the context of Frank and Frank's (1991) six common therapeutic factors. Young referred to Ivey's (1971) microskills as *building blocks* and to the common therapeutic factors as *mega-skills* and so was born the Mega-skills model. Accompanying the Mega-skills approach is a collection of instrumentation designed to measure each component of the model.

In a similar spirit of model integration, the Skilled Counselor Training Model (Smaby, Maddux, Torres-Rivera, & Zimmick, 1999; Urbani et al., 2002) is a combination of Carkhuff's skills, which Smaby and Maddux (2011) refer to as basic skills, and Ivey's microskills, which they refer to as advanced skills. To assess the progress of students being trained according to their model, Smaby and Maddux included no less than five well-researched instruments including the Skilled Counselor Scale (SCS) to assess students' skills. Yet, no effort has been made to demonstrate cross modality validity of any skills instrument.

Other new instruments fall into the category of situation-specific measures designed to provide workshop and training course directors with a measure of participant progress (Rubel, Sherpell, Sobell, & Miller, 2000; Saitz, Sullivan, & Samet, 2000; Walters, Matson, Baer, & Ziedonis, 2005). Many of these instruments, however, rely solely on the participants' self-reported sense of self-efficacy. This and the contextually specific nature of the instruments tend to render them invalid for assessing anything beyond participants' sense of progress as they exit the course for which the instruments were designed.

Program Evaluation and Skills Assessment

In addition to the requirement that counselor educators perform continuous systematic evaluations of students' progress, CACREP (2009) also requires each counseling program to institute a formal program evaluation plan that describes the performance of the program and informs the faculty's future decisions concerning curricular and program development. The multifaceted nature of counseling skills assessment can make such formative evaluations challenging. Counseling programs, even with relatively small faculties, often have more than one

professor teaching counseling skills courses. The complication arises because professors often differ in the preferred method of teaching counseling skills. One professor may teach skills according to the Human Resource Training model and use an instrument designed and validated for Carkhuff's model. Another professor might teach skills according to the Mega-skills model and use an instrument validated for Young's model. This arrangement raises a number of questions. Are both instruments equally reliable and valid? Are both instruments equally sensitive at detecting high and low performing students? Are both instruments equally rigorous in the evaluation of beginning counselors' skills? If the various instruments are not equal, is it fair that students in the program are evaluated by unequal measures? The difficulties are compounded in programs where the evaluation plans call for all students to be assessed according to the same metric. To avoid encroaching on the academic freedom of the professors teaching the courses, a single instrument that can assess students' counseling skills across teaching modalities would be a valuable contribution to the profession.

A return to the instruments of Roger's day that measure various constructs theoretically associated with effective counseling could be acceptable cross-modality assessments, but they do not satisfy the latest CACREP standards (2009) which specify at least 88 different times that mastery of knowledge or possession of skills must be "demonstrated." For example under Professional Practice (Section III), introductory paragraphs of each of the areas outlining the requirements for each accreditation track end with the same requirement for counseling programs to "provide evidence that student learning has occurred in the following domains." In every case, those domains specifically include counseling skills. For example, students preparing to be Clinical Mental Health Counselors must demonstrate "appropriate use of culturally responsive individual, couple, family, group, and systems modalities for initiating, maintaining, and

terminating counseling" (Section III,CMHC.D.5). Training-model-specific instruments do not seem to be appropriate because of their narrow focus and the lack of research support for their use beyond their limited context. Self-report instruments can span the divide across teaching models, but they are not appropriate as they do not measure learning outcomes, and their use would abdicate the counselor educators' or supervisors' responsibilities to evaluate. The unavoidable conclusion is that a need remains for an instrument that trained expert raters can use to measure demonstrated counseling skills accurately, regardless of the skills training method used to teach the students. Based on this literature review, counselor educators need a universal skills assessment instrument with demonstrated cross modality validity to demonstrate learning outcomes and insure counselors-in-training are acquiring the appropriate skills for professional practice.

Rater Training and Skills Assessment

In addition to the need for a universal counseling skills measure, the need also exists to refine the training of raters. Following established research practices, researchers report the methods used to train raters and describe the scoring procedures used in their studies. Some studies show a significant investment in rater training with impressive interrater reliability (Schaefle, Smaby, Packman, & Maddux, 2007). Others studies reflect little effort put forth to training raters, and the interrater reliability results are correspondingly weak (Fitch, Gillam, and Baltimore, 2004).

In a study of rater training effectiveness, as measured by interrater reliability (IRR), Schanche, Nielsen, McCullough, Valen, and Mykletun (2010) were able to achieve poor to fair

interrater reliability (ICC = .28 - .55) after raters received eight hours of instrument specific training. When the training was increased to 15 hours, raters' data produced IRR of ICC = .42 - .71. An additional 20 hours of practice and reducing the raters' focus from six subscales to two subscales raised the participants' IRR to ICC = .76 -.95. These results led the authors to speculate that a substantial investment in rater training could produce IRR coefficients strong enough to warrant single rater assessments. At first glance, 15 hours of training and another 20 hours of practice may seem excessive; however, some researchers recommend even more training.

In a study designed to explore the possibility of eliminating rater bias, Wang (2010) concluded that a three-tiered training approach is best. Raters in this study assessed international student's language skills using English as a second language. While the application of the raters was not associated with scoring counseling skills, Wang's discoveries about rater training can help inform counselor educators' thinking on the matter. The raters in Wang's study began by receiving 20 hours of pre-service training. Because Wang found that skills learned during rater trainer are not retained unless they are constantly refreshed by use, he added a second tier of on-service training. This consisted of a refresher training session immediately before the assessment sessions were scheduled to begin. The third tier of training was in the form of pilot-on-task training. This included having the raters score an old example of language skills demonstration and compare each rater's score to the mean produced by previous raters. Wang's striking conclusion was that "Rater training is not a once-for-all matter, it is on-going business" (p. 110). While it may not yet be clear how much rater training is enough training, what is clear is that through instrument-specific training researchers, program evaluators, supervisors, and counselor educators can enhance greatly the IRR of their assessments of beginning counselor's basic skills.

Scoring Procedures and Skills Assessment

An additional finding of the Schanche et al. (2010) study was that narrowing raters' focus can improve IRR. To raise the IRR coefficients from the *good* range to the *strong* range, raters were not only given 20 hours of practice, but instead of scoring performances on all six of the instrument's subscales, they were responsible for scoring on only two subscales at a time. How much of the improvement in IRR was attributable to practice and how much was attributable to the narrowing of raters' field of focus was not examined. However, the authors noted that the difficulty in achieving strong IRR coefficients "seems to be when the students have to focus on several complex processes at a time. They are significantly more reliable when they are allowed to focus on only two instead of six clinical phenomena" (p. 14).

Although manipulating scoring procedures to improve IRR is not a new practice, narrowing the raters' focus is not a common approach. Schanche et al. (2010) narrowed raters' focus by limiting the number of instrument subscales for which each rater was responsible. The proposed study extends that concept through the use of an original scoring concept, the Dynamic Scoring Interface (DSI). The DSI narrows a rater's focus not by eliminating parts of the assessment instrument but by transforming the rating process from series of global evaluations performed item by item into a myriad of singular session events culminating in an overall rating. For example, the traditional method of scoring a student's mastery of reflecting skills would be for the raters to watch the session video, or assigned portions of the video, and then assign an overall score to the student's ability to make effective reflections of content, reflections of feeling, and reflections of meaning. The DSI narrows raters' focus by asking them to rate every

response the counselor makes at the moment it is made. That raters consider the student's overall reflecting skills is neither necessary nor desirable. Every time the counselor responds to the client (or fails to respond when he or she should have responded), a rating moment is created. In that moment, the rater's focus is limited to only that one response. This substantially narrows the raters focus incorporating the results of Schanche et al. into this investigation.

A significant part of counselor educators' and clinical supervisors' professional responsibilities is associated with evaluating their students and supervisees. A substantial part of that evaluation is the assessment of trainees' basic counseling skills. The review and scoring of video recorded sessions is the method most used to evaluate trainees' basic skills. Yet, a challenge remains to provide an efficient framework for evaluative feedback that is valid across various skills teaching modalities and reliability across raters. That challenge prompts the question, can a single instrument designed to assess basic counseling skills be implemented in a structured scoring environment in such a way that provides a valid cross-teaching-modality evaluation with acceptable IRR?

Purpose Statement

The purpose of this study is to explore the effect of brief instrument specific rater training on interrater reliability (IRR) and counseling performance differentiation using an original instrument, the Universal Counseling Skills Assessment (UCSA) and an original scoring method, the Dynamic Scoring Interface (DSI). The training given to raters will be based on the Human Resource Development model (Carkhuff, 2000), but the language of Microcounseling (Ivey & Authier,1978), Megaskills (Young, 2013), and the Skilled Counselor Training Model (Smaby &

Maddux, 2011) will be incorporated to facilitate cross modality understanding. Study participants will be asked to identify by which skills training model they were taught and which skills training model they use when they teach a counseling skills course. Ideally, the proposed study sample will consist of sufficiently varied and equal sized groups across the various training modalities to allow for statistical exploration of the UCSA-DSI's cross modality validity. However, because the use of a volunteer sample precludes prior knowledge of the sample's constituency, the exploration of the effect of demographic characteristics on counseling skills scoring has been relegated to a potential *post hoc* analysis of the data.

This study seeks to address the need for a cross-methodology skills scoring instrument by presenting the Universal Counseling Skills Assessment (UCSA), a 12-item assessment of basic counseling skills. Each item is scored on a 5-point Likert-like scale. The instrument produces three subscales. The first five items comprise the *Attending Scale*, the second five items make up the *Basic Listening Scale*, and the final two items constitute the *Deepening Scale*.

In this study, I will create on-line video based instruction to train raters in the use of the UCSA and the DSI. The study's data provide validity and reliability information regarding the UCSA in both static (traditional instrument scoring method) and dynamic scoring environments. The data will be analyzed to explore whether there is a difference in mean score and IRR between untrained raters (counselor educators, doctoral students in counselor education who have already completed their supervision internship, and master's level counselors who are certified as clinical supervisors) and trained raters (expert raters who have already completed the on-line video based training concerning the use of the UCSA and DSI). Also, the data will be analyzed to explore what affect the quality of the counseling performance has on the rater

performance by comparing raters who score an example of a low counseling skills performance with raters who score an example of a high counseling skills performance.

Hypotheses

This study will formally test the following hypotheses:

H_{O_1}: There is no significant difference in the mean UCSA scores (S_{low}, S_{high}, S_{all}) by rater type (untrained, trained)

H_{O_2}: There is no significant difference in IRR coefficient by rater type (untrained, trained)

H_{O_3}: There is no significant difference in $IRR_{untrained}$ coefficient by counselor performance level (low, high).

H_{O_4}: There is no significant difference in $IRR_{trained}$ coefficient by counselor performance level (low, high).

Depending on the demographic characteristics of the sample, a fifth hypothesis may be tested. Because of the uncertainty associated with the demographic characteristics of a volunteer sample, the decision to test a fifth hypothesis and the final composition of that hypothesis will be determined as a post hoc analysis. Hypothesis five is presented here as an example of what the post hoc hypothesis will be if all seven demographic characteristics are present in sufficiently equal quantities.

H_{O_5}: There is no significant difference in mean UCSA scores (S_{Low}, S_{High}, S_{All}) by demographic characteristic (age, years of experience, degree held, race, gender, method by which taught counseling skills, method used to teach counseling skills).

Conclusion

Counselor educators and clinical supervisors are compelled by ethical standards, accreditation bodies, and the expectations of the discipline to provide accurate, consistent, outcome based evaluation of counseling students' clinical skills. Technology has paved the way for the production of high quality video recordings of sessions to be within the ability of every counseling student. Varied teaching methods have created a need for a universal (across teaching modality) assessment instrument and scoring procedure. This study will present such an instrument (the UCSA) and scoring procedure (the DSI) and examine the effect of instrument specific training on rater performance.

CHAPTER II: REVIEW OF THE LITERATURE

The purpose of this study is to explore the effect of instrument specific rater training on interrater reliability (IRR) and counseling skills performance differentiation using the Universal Counseling Skills Assessment (UCSA) in a dynamic scoring environment. This chapter is comprised of a review of the literature regarding the counseling skills teaching models and the literature regarding the assessment of counseling skills. The review of teaching models includes Carkhuff's Human Resource Development model(1969), Ivey's Microcounseling model (1971), and two modern iterations of them, the Skilled Counselor Training Model (Smaby & Maddux, 2011) and the Megaskills (Young, 2009) model. The review of literature regarding the assessment of counseling skills will include an overview of various assessment instruments, the role of raters and rater training in assessment, the introduction of an original instrument, the Universal Counseling Skills Assessment, and an original scoring procedure, the Dynamic Scoring Interface.

Teaching Counseling Skills

As was noted in chapter one, an historical link exists between counseling skills teaching models and counseling skills assessment instrumentation. Therefore, before entering a discussion of how skills are assessed, literature regarding the teaching of skills is presented. Any

discussion of the teaching and assessment of counseling skills must begin with the work of Carl Rogers (1957). Baker, Daniels, and Greely (1990) credit the introduction of person-centered counseling (Rogers) with opening the therapeutic profession to supervisors providing either live supervision or supervision based on the review of recorded sessions provided by beginning therapists. Previously, therapist-patient interactions were considered sacrosanct (Conver, 1944; Matarazzo, Phillips, Wiens, & Saslow, 1965; Matarazzo, Wiens, & Saslow, 1966), relegating the assessment of students' skills to classroom simulations. "In essence, the trainee's supervised experiences were of a therapeutic nature as supervisors aimed to help work through the student's difficulties in describing the sessions" (Baker et al., p. 356-357). Rogers' emphasis on the *facilitative conditions* that brought about the *necessary and sufficient therapeutic relationship* began to focus attention on the assessment of therapist behaviors. However, Rogers stopped short of identifying a set of discrete counseling skills, and therefore, the assessment instruments associated with client-centered therapy do not measure counseling skills. Rather they seek to measure the characteristics and strength of the counselor-client relationship (Barrett-Lennard, 1962; Horvath, 1981; Saunders, Howard, & Orlinsky, 1989). Not until the work of Robert Carkhuff and the Human Resource Development (HRD; 1969) model did educators begin teaching counseling skills as measurable learning outcomes.

Carkhuff and the Human Resource Development Model

A strong congruence exists between Carkhuff's model of teaching counseling skills and his personal philosophy of life. In the prologue of his iconic book, *The Art of Helping*, Carkhuff presented his concepts as those that "move the humankind to change" (2000, p. xi). He reflected

happily on the fact that because of his work the words "*interpersonal* and *skills* are linked together in a growthful embrace" (p. xii). He concluded his opening remarks with the hope that humans might learn to relate to each other better, and that by doing so, we might become "growthful people" (p. xii). His belief in human growth is further emphasized in the opening words to his *Credo of a Militant Humanist*. "My fundamental assumption in life is this: The only reason to live is to grow and therefore growth is worth any price" (1972, p. 237). This personal passion that humans might grow, learn to relate better, and realize their full potential is at the heart of Carkhuff's model of counseling skills instruction.

Theoretical Foundation

Carkhuff's (1969) philosophy of producing skilled helpers is comprehensive in scope. The nature of the HRD model can best be described as integrative, behavioral, and relational.

Integrative.

Within Carkhuff's model, counselor educators are called upon to model the skills they are teaching and assessing. This reliance on parallel process is seen as being fundamental to positive outcomes. Carkhuff's early studies demonstrated that training outcomes were determined not only by the quality of didactic presentations, but also by educators' level of functioning within the skills.

Educators are expected to make effective classroom presentations of material, but also necessary is that they be able to coach beginners through the early stages of practicing new

skills. Carkhuff's model also relies upon students being assessed and students assessing themselves as they grow into this new skill set. Educators must, therefore, be able to assess student performance accurately and to give effective feedback. To summarize, Carkhuff's model of teaching counseling skills requires an integration of didactic presentation, coached practice, repeated assessment, and instructor modeling.

Behavioral.

The skills of the HRD model can be divided into two categories. The first set of skills is associated with the student's ability to discriminate between more effective and less effective counselor responses. The second set of skills revolves around the student's ability to form and communicate effective responses. In the early stages of training, responses are provided from books, video recordings, or in-class demonstrations and students have the time to consider the responses and rate them according to their respective merits. However, students must quickly develop the ability to formulate and communicate their own effective responses, while under the pressure of interacting with clients.

Communication encompasses everything a counselor does or says that in any way speaks to the client, including, but not limited to, body language, tone and volume of voice, attire, and the content of the spoken word. The behavioral emphasis of the HRD model is demonstrated by Carkhuff's comparison of the relative merits of discrimination skills and communication skills. Although students should master both, communication skills are considered to be of greater importance to the beginning counselor. Carkhuff's research forced him to conclude, "While there is evidence to relate helper level of communication to indices of constructive helpee change,

there is no evidence to relate discrimination to client change" (Carkhuff, Collingwood, & Renz, 1969, p. 461). In other words, a counselor might excel in discriminating between high and low level responses, but without the necessary communication skills, the counselor will not be able to help clients effect change in their lives.

Furthermore, the HRD model is classified as being a behavioral approach because the communication component can be taught only through the process of doing. Carkhuff, Collingwood, and Renz (1969) studied participants who had received 16 hours of training exclusively on differentiation skills. They found that reading or hearing well-considered responses and learning to differentiate between high level responses and low level responses did not significantly improve students' abilities to formulate responses. "The direct implication is that to effect differences in communication, the training must emphasize a behavioristic approach providing practice in communication" (p. 461).

Relationship Based.

Carkhuff's training model is recognizable for the emphasis it places on counselor-client interaction. Client-centered, existential, and constructivist models emphasize the clients' contribution to and guiding hand in the therapeutic relationship. Almost all other models focus intently on the activity of the counselor. The HRD model is predicated upon the balanced interaction of counselor and client. The counselors' activities are described as attending, responding, personalizing, and initiating. The clients' activities are defined as involving, exploring, understanding, and acting. While these activities are not unique to the HRD model, the scaffolding of the reciprocal nature of these activities are.

Counselors attend to clients so that clients might become involved in the process. Counselors respond to involved clients to facilitate clients' exploration. Counselors personalize clients' explorations to bring about understanding. Counselors initiate client's actions to bring about change. Once action has begun on the part of the client, counselor and client together reenter the exploring phase and recycle through the acting stage. Counselors are never asked to be experts on their clients' lives. However, counselors are taught to be experts on this process of helping. This balanced, flowing interaction is the foundation of Carkhuff's theory of helping.

Pedagogical Techniques

Pedagogical techniques associated with the HRD model can be divided into three groups. The desired skills are taught and modeled by the instructor, practiced by the student, and assessed by both the instructor and student.

Modeled.

As was noted earlier, instructors must model the skills they expect students to learn. Carkhuff explained "the trainer is the key ingredient insofar as he offers a model of a person who is living effectively" (1969, p. 201). He even went as far as to say that without an instructor who can model congruently the expected skills, the rest of the training process is meaningless.

The modeling of HRD tenets is most apparent in the communication between the instructor and student. Students are met with warmth and sensitivity but not with unconditional acceptance of their performance. Students are directed toward more effective practices and, when

necessary, corrected. That direction and correction, however, take place in an atmosphere of trust and honesty.

Practiced.

Instructors bridge the gap from their modeling of counseling skills to students' performance of those skills by leading them through experiential exercises. These exercises place students in counseling-like situations with instructors to lead them through a session, coaching them on a response-by-response basis when necessary. As students struggle to master new skills, they experience a useful analog to the stresses and uncertainties they will experience with clients. For many, this is an unsettling experience. However, this anxiety laden struggle for mastery is necessary for most students to be able to make the transition from talking about what they might say during counselor-client interactions to actually responding during counseling-like situations.

Some instructors use these early learning experiences as class-wide instruction and development sessions within their counseling skills courses. Paladino, Barrio-Minton, and Kern (2011) have married Carkhuff's model with Andersen's reflecting team concept (1991), which assigns the observing classmates a reflective feedback group role, and staged these experiences in front of the class. Live observation provides input from other class members but at the cost of additional pressure on the student struggling with unfamiliar skills. Live observation also takes advantage of the vicarious learning principle allowing classmates to learn from their peers' errors and triumphs. Paladino et al. have demonstrated the efficacy of staging the mock counseling sessions in front of the class, videotaping them, and then reviewing the recordings with the class similar the way a football coach might review the previous week's game tapes. This arrangement

capitalizes on the benefits of the class as a reflecting team and provides the opportunity for students to review and assess formally their performances. This self-assessment is a key element in developing self-reflection practices and learning to become a self-monitoring professional (Schon, 1983).

*Assessed***.**

Accurate assessment of demonstrated counseling skills is vital to the HRD model of skills instruction. Students and instructors regularly score videotapes of students' mock counseling sessions and compare the ratings. Throughout this process, instructor assessment serves as the standard against which students refine their self-assessments skills. Instructor feedback facilitates students' efforts to transfer counseling skills from the textbook into practice.

Formative feedback from those recognized as experts has been demonstrated to have a powerful influence on counseling students (Daniels & Larson, 1992). Daniels and Larson explored the effect of performance feedback on counseling students' levels of self-efficacy and anxiety. The sample of 45 counseling trainees (39 females, six males; 37 Caucasian, six African-American, two Internationals) were randomly given either positive or negative feedback they believed was based on their counseling performance in a mock session. Prefeedback-postfeedback comparisons of self-efficacy and state anxiety measures indicated that positive feedback correlated with significant increases in self-efficacy and significant decreases in state anxiety. The data also demonstrated a correlation of negative feedback with significant decreases in self-efficacy and significant increases in state anxiety.

Empirical Support

Carkhuff's integrative behavioral relational approach has not only stood the test of time but has been formally tested hundreds of times. Some research focused on testing various aspects of the helping process (Carkhuff, 1969c; Collingwood, Renz, & Carkhuff, 1969), some on aspects of training and supervising counselors (Pierce, 1968; Carkhuff, Friel, & Kratochvil, 1970), and others on the unique research issues faced when exploring aspects of the HRD model (Cannon & Carkhuff, 1969; Carkhuff & Burstein, 1970). Over the first twenty years of his work, Carkhuff (1983) reported 164 studies (n_{total} = 158,940) related to the HRD model and indicating positive outcomes for more than 90% of the participants. Since then, his method has continued to be used, tested, and modified (Smaby & Maddux, 2011).

Allen Ivey and the Microcounseling Model

Perhaps the greatest testament to Carkhuff's work is that it has spawned other significant paradigms. From a historical vantage, Microcounseling (MC; Ivey, 1971) was born when "Allen Ivey responded to Truax and Carkhuff's call for training reform by expanding their idea to use skills-based instruction to bridge the gap between theory and practice" (Ridley, Kelly, & Mollen, 2011). The last 40 years have seen MC become the "ubiquitous" (Ridley et al.) choice of counselor training programs for skills instruction. Because one was the progenitor of the other, many similarities exist between the HRD model and the microcounseling (MC) model. Both are grounded in social learning theory and emphasize counselor behaviors (skills) that bring therapeutic benefit to the counselor-client relationship. Also, both models include "an

instructional component, supervised practice, and immediate and concrete feedback" (Baker, Daniels, & Greeley, 1990, p. 358). Both models also make use of recordings for feedback purposes.

Differences, however, do exist, and MC's unique contributions to skills training are accentuated by those differences. While Carkhuff made use of recordings, primarily for evaluation and feedback purposes, Ivey greatly expanded their use. Ivey made significant use of the then new technology of video recordings for both didactic and evaluative purposes (Reivich & Geertsma, 1969). Ivey also expanded the scope of counseling skills from basic beginning skills to more advanced interviewing techniques. Microcounseling's greatest contribution, however, was "its ability to isolate and specify discrete objective interviewer behaviors that the counseling trainee can be taught" (Baker et al., 1990, p. 364).

Theoretical Foundation

Ivey never intended MC to be associated with any one theoretical orientation, but that does not mean it has no theoretical moorings. Microcounseling was founded on Ivey's epistemological belief that knowledge is most understandable when rooted in experience (1973, 1978), which shaped the behavioral nature of his theory. He was also strongly influenced by social learning theory (Weinrach, 1987) which helps to account for its incremental aspects. Ivey (1973) referred to the teaching of MC as competency-based microtraining and founded it upon five principles. Ivey believed: (a) that counseling is complex but can be learned by focusing on a single skill at a time, (b) effective training includes time for students self-reflection, (c) that interviewing skills can be learned in part by observation of those skills being properly

demonstrated, (d) that microskills are atheoretical and can, therefore, be effectively employed across a diverse palette of theoretical orientations, and (e) "microtraining sessions are real interviewing" (Ivey & Authier, 1978, p. 13). The behavioral nature of MC is further seen in Ivey's intended outcome. "The outcome variable of the effective training is an effective trainee – a helper who can demonstrate specific competencies of helping with equally demonstrable affects on the helpee" (Ivey & Authier, 1978, p. 563).

Microcounseling can, therefore, be summarized as a model based primarily on students' learning and demonstrating specific behaviors with an eye toward bringing about behavioral changes in their clients. Microtraining seeks to bring about student learning incrementally. Notice again the first of the five principles upon which microcounseling is built: The complexity of counseling is to be learned a single skill at a time. "Microtraining procedures are not concerned with producing effective counselors in one session. The primary concern is helping trainees to grow with time" (Ivey & Authier, 1978, p. 302).

Pedagogical Techniques

In a sense, Ivey saw the MC model as a continuation of the movement toward behaviorism started by Carkhuff's early work, but MC was more concretely and operationally defined counseling. Many of the training techniques are similar to those of Carkhuff's method. Students learn through a combination of didactic presentation, experiential practice, reflective self-assessment, and instructor provided formative feedback. Ivey did place much more importance on the formal use of materials during training than did Carkhuff. The appendix of his book (Ivey & Authier, 1978) contains nine different forms that can be used to score various

aspects of student performance as they progress through training exercises and early counseling sessions.

Use of Assessment

Ivey's use of assessment is very similar to Carkhuff's but much more formalized. The Taxonomy Scoring Form (TSF, Ivey & Authier, 1978, p. 470-471) is a good example of that formalizing of assessment instrumentation. The TSF is a detailed and complicated scoring sheet designed to catalog every counselor response across 24 categories. The purpose for this detailed scoring is the same as assessment in the HRD model. Student learning of microskills and self-assessment skills is enhanced by assessing their own performances and comparing them to expert ratings. In a sense, the TSF is the early progenitor of the Dynamic Scoring Interface (DSI) which is introduced in this study. More will be said about this connection between the TSF and the DSI in the section on Counseling Sills Assessment.

Empirical Support

Like Carkhuff's HRD model, Ivey's MC model has stood the test of time. Countless researchers have used, tested, and modified the MC model. Daniels' (2003) meta-analysis of MC studies reviewed more than 450 studies conducted between 1967 and 2002, and although they have taken issue with some of the extant research, Ridley et al. (2011) acknowledged that "For more than four decades, the microskills approach has been the dominant paradigm for training entry-level counseling students" (p. 800). In their latest textbook, Ivey and Ivey (2007) claim,

"Over 450 data-based studies affirm the validity of the microskills approach; thus, the approach is the most researched model of interviewing and counselor training available" (p. 7), a claim which even MC's detractor do not question.

One of the ramifications of being the most used and tested counseling skills instructional models is that others have sought to modify the MC or integrate it with other paradigms. Some of the modern attempts at modified usage of MC training include a self-instructional version (Schonrock-Adema, Van der Molen, & van der Zee, 2009) and a version integrating MC and the reflecting team (Anderson, 1991) model of collaborative skills development (Hawley, 2006).

Schonrock-Adema et al. found that microskills could be self-taught as long as an experienced trainer was available for the practice, assessment, and reflection portions of the training. Their study with undergraduate psychology students (n = 193, 97 self-taught, 96 trainer-taught) reduced the amount of trainer-student interaction time by 50% while producing significant learning effects with both groups (pretest-posttest means comparison: TT = 2.24; z = 13.17, $p < .001$ and ST = 2.32; z = 19.07, $p < .001$). While there was no significant difference found between the trainer-taught and the self-taught posttest means, the effects size for both groups were large (TT: d = 2.11; ST: d = 2.29). This study potentially opens the door to more efficient methods of skills instruction, possibly indicating that MC may yet have more to offer.

Hawley's (2006) grounded theory study (n = 15, 13 females, two males; one African-American, 14 European-Americans) of MC in a reflecting team (RT) environment yielded three themes related to the use of MC. Student in the RTs felt they benefited from seeing other beginners bring the microskills to life in a mock-session environment. Students in the role-play positions believed the RT helped them understand what skills they were using in a way that appeared helpful to the client and helped them understand when they had not yet used a skill

well. For example, one student observed "I will start with one questions and I will [change the subject to] something else…instead of letting someone answer that question" (p. 205). Hawley's overall sense of the integration of the MC model with the RT model was that student anxiety seemed less while there seemed to be an increase in the "volume of internal thought processes compared with other skill training methods (e.g., in-class role plays, fishbowls, supervisory experiences)" (p. 206). Hawley posited that this study might point a way toward merging the "behavioristic and deductive nature of the training method [MC, added by PDM]" with "the art of the therapeutic relationship" (p. 206). Hawley's conclusion may indicate that in some quarters there is a desire to bring MC on a full circle return to Roger's therapeutic relationship conditions. An alternative explanation may be that because of its moorings through Carkhuff's work, MC skills instruction, in the hands of qualified instructors, never left "the art of the therapeutic relationship."

Modern Extensions of Carkhuff's and Ivey's Models

One of the impressive aspects of both Carkhuff's and Ivey's work is that they have endured. These models have endured because they have demonstrated themselves to be effective. They have also proven to be adaptable. Two modern efforts to modify the HRD and MC models have progressed beyond the single study phase into burgeoning skills training models.

Recently, the HRD and the MC models of skills instruction have been synthesized to form the Skilled Counselor Training Model (SCTM; Smaby, Maddux, Tores-Rivera, & Zimmick, 1999; Urbani et al., 2002). Smaby and Maddux (2011) have written the textbook outlining the SCTM, and in time, their names will probably become most closely associated with

the model. They have openly associated the SCTM with the previous work of Carkhuff and Ivey and have retained much of the original nomenclature. The merger has been described as the taking of Carkhuff's model for teaching basic skills and Ivey's model for teaching advanced skills and combining them into a structured system of instruction rife with experiential learning and formal assessment (Crews et al., 2005). "The SCTM emphasizes modeling, mastery, persuasion, arousal, and supervisory feedback during counseling training as key elements to promote skills acquisition, self-appraisal of counseling skills, self-monitoring behavior, and cognitive complexity" (Little, Packman, Smaby, & Maddux, 2005, p. 190). In the SCTM, counseling is presented as a three stage process (exploring, understanding, and acting), with each stage being uniquely associated with: "(a) a purpose, (b) two counseling processes, and (c) six counseling skills" (Urbani et al., 2002, p. 93).

The second recent skills teaching model to come out of Carkhuff's and Ivey's work is the Megaskills model produced by Mark Young (2009). All of the models of skills instruction discussed thus far can be referred to as *bottom-up* theories. The idea is that building a counselor is analogous to building a house. A foundation of fundamental skills is first laid and more complex tasks (i.e. case conceptualization, formalizing treatment plans, etc.) are then built on that foundation.

Young (2009), however, has espoused a top-down approach. Rather than building a house, the house is believed to already exist in the form of the trainees' personality, life experiences, and innate abilities. The house is merely in need of certain renovations to adapt the person for counseling tasks. Young's message is somewhat confusing as he denounces the skills-up concept of learning counseling and suggests that microskills be presented only in the context of how they "evoke common curative factors, or 'megaskills'" (p. 37).

Though presented as being a departure from skills-based training and named Megaskills, Young's model is more like Ivey's Microskills model than different from it. Discrete skills almost identical to microskills (Ivey, 1979) are presented as *building blocks* (eye contact, body position, attentive silence, voice tone, facial expressions and gestures, physical distance, touch, encouragers, questioning, reflecting content, reflecting feelings, reflecting meaning, paraphrasing, non-judgmental listening, confrontation, and goal setting). These skills do not seem to represent a departure from those taught in Carkhuff's and Ivey's models. The mega-skills to which he attaches the building blocks are the common therapeutic factors identified by Frank and Frank (1991), which do provide an excellent context in which students can conceptualize the microskills of counseling.

Counseling Skills Assessment

Assessing counseling skills is a complex and sometimes controversial undertaking. Some question whether counseling skills are worthy of assessment separate from a global sense of the therapeutic relationship they engender, preferring rather to concentrate on counselors' personality traits as the central constructs of counseling (Grencavage & Norcross, 1990; Lambert, 1989; Stein & Lambert, 1995; Stevens, Dinoff, & Donnenworth, 1998). Others have concluded that the therapeutic arena is too ambiguously defined for anyone to be able to differentiate who the expert counselors are (Dawes, 1994; Lichtenberg, 1997; Smith & Glass, 1977). Jennings, Goh, Skovholt, Hanson, and Banerjee-Stevens (2003) concluded that in counseling literature "there is no *gold-standard* or agreed upon definition of *expert* or *master* counselor or therapist" (p. 60) and that "expertise in counseling and therapy is a dynamic and

complex phenomenon that requires multiple explanations" (p. 62). The profile of an expert counselor Jennings et al. extracted from the literature included the following factors: (a) how much experience a counselor had, (b) the personal characteristics the counselor brought to the task, (c) how multiculturally competent the counselor was (although this was defined without consideration of any of the demonstrable skills associated with multicultural competencies), and (d) how comfortable they were with the ambiguity inherent in the counseling process. Mastery of counseling skills was mentioned only in the context of being an anxiety producing concern of beginning counselors. Those holding to this profile of an expert counselor have little use for the assessment of counseling skills, relying more on therapists' surveyed opinions than empirical measures of effectiveness.

Ironically, there are others who rely on surveyed opinions (self-report) as though they were empirical measures. Assessing skills improvement by self-report is tempting to many as is demonstrated by the use of self-report surveys as outcome measures. Walters, Matson, Baer, and Ziedonis' (2005) study of workshop effectiveness began with a review of 353 workshops purporting to increase substance abuse and addictions counselors' abilities, knowledge, or skills. Their inclusion criteria required that the workshop "measure some outcome related to workshop participation (e.g., knowledge, attitudes, demonstration of skills)" (p. 284). Many of the workshops reviewed rely on the participants' self-report of improvement or learning to assess their effectiveness, even though the validity of self-report has been called into question (Miller & Mount, 2001; Miller, Yahne, Moyers, Martinez, & Pirritano, 2004).

In the case of assessing counseling skills, the accuracy of self-report has been severely discredited. Urbani et al. (2002) sought to study self-efficacy among counseling students by comparing their self-ratings to the scores of expert raters and monitoring the effect of the rating

comparisons on self-efficacy. They found that before skills training counselors-in-training (CIT) tended to rate their abilities significantly higher than did expert raters (means comparison of students to expert ratings: $t = 6.25$, $p < .001$, $d = .87$). Immediately after skills training, counselors tended to rate their skills significantly lower than did expert raters (means comparison of students to expert ratings: $t = 2.84$, $p < .01$, $d = .39$). These results fail to support the validity of student self-ratings of counseling skill both before and immediately after completing a counseling skill course. Counselor educators are left to seek some more accurate measure of student's skills. That search often leads back to the originators of the counseling skills teaching model being used.

Two factors combine to explain the proliferation of orientation specific or model specific counseling skills measures. First, educators and practitioners intuitively rely on the originators of a particular teaching model as the experts who can best measure performance of those trained in that model. Second, those who conceive of define, and present new pedagogies naturally want to measure and demonstrate their effectiveness. These factors have lead to many model specific instruments.

Assessment Reliability and Rater Training

Those who have been charged with providing accurate and reliable evaluations of counseling trainees' skills inevitably face the challenge of producing trained raters who can assess reliably students' performances. "Raters are not born raters," one researcher observed (Wang, 2010). However, studies (Fitch, Gillam, & Baltimore, 2004; Lepkowski, Packman, Smaby, & Maddux, 2009; Little, Abney, Packman, Smaby, & Maddux, 2005; Urbani et al.,

2002) do seem to support the idea that assessment, both the assessment of other's skills and the self-assessment of skills, is a learned skill and can be improved through training.

Fitch et al. (2004) explored the reliability of counselor educators' and supervisors' ratings when scoring counseling students' video recordings of mock sessions. The sample of 21 educators and supervisors (12 males, nine females; 20 Caucasians, one Native American; 18 holding Ph.D., two holding Master's degrees, one did not specify) found that supervisors can differ widely in their ratings of trainees' skills. Some raters assigned almost perfect performance scores and others assigned failing scores to the same video. Other studies (Lepkowski et al., 2009; Little et al., 2005; Urbani et al., 2002) have demonstrated that counseling students' untrained self-assessments are equally unreliable. However, those studies also demonstrate that after students receive skills training and consolidate their learning with a period of clinical practice, their assessments do not differ significantly from expert ratings'. Control groups (receiving counseling instruction but no skills training) continued to rate themselves significantly higher than did expert raters.

The Lepkowski et al. (2009) study further explored whether this trend in students' self-ratings differed according to gender. The sample (n = 69, 52 females, 17 males) produced data indicated that males' self-ratings were significantly higher than females' self-ratings on the pre-test, and that both groups' pre-training self-ratings were significantly higher than expert ratings. There was, however, no significant difference between males', females', and experts' ratings after skills training. This would seem to indicate that self-assessment is a learned skill, and that teaching counseling skills helps bring students' self-ratings in line with expert ratings. However, because student's skills need to be evaluated during and immediately after the skills training processes self-assessment, even self-assessment that through training and practice eventually

comes into line with expert assessment, does not address counselor educators' needs. The challenge remains to produce expert trained raters who can make reliable assessments of counseling students' skills.

In each of the above studies (Lepkowski et al., 2009; Little et al., 2005; Urbani et al., 2002), the expert raters were experienced users of the chosen skills assessment instrument and received specific rater training. Recent studies have demonstrated that rater training occupies a position of growing importance. Binhong Wang's (2010) meta-analysis of rater training studies examined the effect of rater training in the context of raters who assessed the language skills of international students using English as a second language. Two studies stood out among the others. Haizhen Wang's (2008) study (24 raters, 20 females, 4 males; all Chinese) indicated two difficulties in using raters who had not been specifically trained for the task. First, raters did not limit themselves to the evaluative criteria of the rating instrument but applied a variety of other criteria as well. Second, raters disagreed in their interpretation of the constituent items of the rating instrument. The second study was similar except that Wen, Liu, and Jin (2005) introduced the element of cultural diversity among the raters. Even though the sample of raters (n = 11; five native English speakers, six nonnative English speakers) used the same scoring instrument, the data indicated significant differences between native and nonnative raters in the areas of applied criteria and the focus of the raters' interests. These two studies show that expertise in the skill being rated does not assure that raters will: (a) interpret the scoring instrument the same way, (b) limit their assessment to the elements of the scoring instrument, or (c) agree on what elements of the demonstrated skills deserved their attention. Binhong Wang's (2010) analysis led him to conclude that raters need to receive three tiers of training: (a) pre-service training consisting of 20 hours of instruction, (b) on-service training consisting of a refresher training session just prior

to the beginning of the assessment, and (c) pilot-on-task training consisting of a *dry run* practice rating that is openly reviewed and discussed prior to beginning the actual evaluation.

While not a perfect parallel to the assessment of counseling skills, the assessment of spoken language skill does possess similarities. In both cases, raters are being asked to evaluate a very complex set of skills. Also, in both cases the raters are scoring, according to an instrument, skills that are performed primarily through the spoken word. These commonalities make for theoretical parallels between the effects of rater training in the two disciplines. This assumption is supported by Schanche, Nielsen, McCullough, Valen, and Mykletun's (2010) study of the effects of training on psychology graduate student raters. The participants in this study (n = 32) were measured for interrater reliability on their use of the Achievement of Therapeutic Objectives Scale (ATOS, McCullough et al., 2003). Their performance as raters was measured after having eight hours of rater instruction (ICC's ranging from .28 - .55 on the instrument's six subscales), 15 hours of rater instruction (ICC's ranging from .43 - .71 on six subscales), and again after an additional 20 hours of practice (ICC's ranging from .76 - .95 with the proviso that raters were responsible for only two of the scoring instruments six subscales at a time).

The findings of the Schanche et al. (2010) study clearly support that additional rater specific training improves IRR, as measured by intraclass correlation coefficients. What remains unclear is how much of the final improvement in IRR was a factor of the additional 20 hours of practice and how much was attributable to the narrowing of the raters' focus by making them responsible for only two of the instrument's six subscales.

Skills Assessment Through a Structured Scoring Procedure

An alternative method for focusing raters' attention is to divide an assessment session into multiple single-response scoring sessions. The Dynamic Scoring Interface (DSI) is an original scoring procedure that, like the Taxonomy Scoring Form (TSF, Ivey & Authier, 1978, p. 470-471), creates a response by response scoring environment, which serves to focus and limit raters' attention to the last response made by the counselor. Ivey's TSF could, with practice, be used effectively to track, categorize, and score a counselor's performance on a response by response basis. However, the instrument did have drawbacks. As responses were entered on the form, the TSF became an unwieldy combination of assessment instrument and scoring database. Considering that every counselor response required the rater to consider 24 different columns into which scores could be entered, using the TSF could become overwhelming for the rater. The TSF also called for the rater to categorize both the counselor's and the client's response, doubling the number of scoring entries per session.

The fundamental difference in the TSF and the DSI is that the DSI is not an assessment instrument. The DSI is a scoring procedure – an environment in which an instrument is employed. Using an instrument in the DSI facilitates a structured approach to instrument specific training of raters and creates a highly structured scoring environment that is hypothesized to improve IRR. Because the DSI is computer based, the assessment instrument remains separate from the database of scores giving the rater the same efficient interface for scoring no matter how many responses have previously been entered. Additionally, the DSI breaks down the sessions into component responses which mirror the way skills are taught. The scoring of the component responses creates feedback that intuitively matches the skill-by-skill feedback students receive during training. The highly structured nature of the DSI and the focus it brings to raters' efforts may also reduce the amount of time needed to train raters and improve IRR. In

this study, the instrument employed through the DSI will be the Universal Counseling Skills Assessment (UCSA).

The Universal Counseling Skills Assessment (UCSA)

The counseling skills literature chronicles an impressive variety of skills teaching methods, replete with instrumentation for measuring counseling skills. However, the extant measures do not include an instrument that has demonstrated validity across the teaching models and acceptable reliability. The Universal Counseling Skills Assessment (UCSA) is presented to meet this need.

The UCSA is a highly modified revision of Eriksen and McAuliffe's (2003) Counseling Skills Scale (CSS). The CSS was developed with the intention of creating an instrument that was: (a) valid and reliable, (b) relied on observations of actual in-session performance, (c) had face validity and ease of use, (d) used expert judges' ratings, and (e) required qualitative judgments as to the contextual appropriateness of the skills demonstrated (p. 123). Their final product was a 22-item measure, subdivided into six subscales by which demonstrated skills are rated on a 5-point Likert-type scale by expert raters. The 5-point scale is defined as (+2) *highly developed*, (+1) *well developed*, (0) *developing skills*, (-1) *continue practice*, or (-2) *major adjustment needed*. A sixth non-scoring category allowed raters to indicate that any individual skill was not demonstrated, but that it also was not needed in the observed session.

Using two raters and a sample of Master's level counseling students (N = 29), Eriksen and McAulifee (2003) measured internal consistency using Cronbach's alpha (.91). "Construct validity was assessed by examining pre- to post course changes in student performance and

conducting an item analysis" (p. 130). Comparison of pre-course/post-course scores indicated significant changes in the overall scores and in five of the six subscales ($t = 4.51$, $p < .000$, $ES = .80$). Item analysis produced Cronbach's alpha ranging from .18 to .71 with all items being positively correlated with the overall score. The authors noted as limitations the small sample size, the use of only two raters, and the absence of demonstrated interrater reliability.

The Scoring System

The revision of the CSS that produced the UCSA centered around two major changes—a more intuitive Likert-like scoring system and the reduction of the number of items from 22 to 12. The first revision clarified the scoring system by changing the -2 to +2 scale to a 5-point scale that is scored from 1 to 5 removing the unnecessary complication of a scale encompassing both negative and positive numbers. The new scale was also redefined as follows: (1) *Not observed when required by the situation,* (2) *Demonstrated insufficiently to proceed to practicum/internship,* (3) *Demonstrated sufficiently to proceed to practicum/internship but inconsistently,* (4) *Demonstrated sufficiently and consistently enough to proceed to practicum/internship, and* (5) *Demonstrated at a level normally indicative of one who has at least completed internship.* A sixth non-scoring category continues to identify skills that are (N/O*) not observed but not required for the situation.*

These scoring descriptors tailored the UCSA to our needs as a gatekeeper assessment into the counseling practicum and honored Eriksen and McAuliffe's (2003) rejection of scoring scales that rely solely on the number of times a skill is demonstrated as being "unrefined" (p. 122). Eriksen and McAuliffe concluded that such scales do not adequately consider the quality or

contextual appropriateness of the skill. However, complete removal of the quantitative element seems to introduce an inherent weakness into the scoring procedure and was, therefore, avoided. Scales dependent exclusively on the qualitative judgments of raters have been demonstrated to produce widely varying IRR coefficients that are positively correlated with the targets' levels of performance (Carkhuff, Kratochvil, & Friel, 1968). To ignore completely the number of times a skill is demonstrated will, therefore, call into question the reliability of the data. Through consultation with faculty members and doctoral students experienced as expert skills raters, the current scoring descriptors were chosen to anchor the instrument in its intended use. After refining the scoring descriptors to facilitate the intended usage, consideration was given as to how well each of the individual items furthered the instrument's intended use.

Item Determination

The second major revision to the CSS was to limit the scope of the assessment. This was desirable for two reasons. First, the UCSA was to serve as a rubric by which required video recordings would be assessed in a beginners' skills class. The individual skills measured needed, therefore, to be congruent with and limited to the scope of the skills expected of students during the chronological strictures of the course. Some of the items of the CSS assessed skills that were beyond the scope of that course material. Second, as the skills course is taught by various faculty members, different models of skills training are employed. Some items on the CSS assessed skills that are usually associated only with a specific skills-based model. After consulting with faculty and doctoral students who had been instrumental in teaching the skills course, the UCSA was limited to 12 items divided into three subscales.

The first subscale, *attending*, is comprised of five items: (1) body *language & appearance*, (2) *eye contact*, (3) *minimal encouragers*, (4) *vocal tone*, and (5) *verbal tracking*. Descriptions of each of these items appear on the form to aid raters in understanding the specific elements of each item.

The second subscale, basic listening, is also comprised of five items: (1) *selective attending*, (2) *directs and encourages client to talk*, (3) *paraphrasing (reflections of content)*, (4) *reflections of feeling/meaning*, and (5) *summarizing*. Again, each item is accompanied by an explanation. For example item 10, *summarizing*, is explained as "makes statements at key moments in the sessions that capture the overall sense of what the client has been expressing."

The third subscale, *deepening the session*, is comprised of only two items: (1) *using immediacy* and (2) *challenging/pointing out discrepancies*. Descriptions of these items include sample statements that students might use. For example, item 12, challenging/pointing out discrepancies is explained by the example, "You expect yourself to do ___ when facing the problem of ___ but you do ___ instead. When this happens, you feel ___ about yourself."

The resultant instrument (the UCSA) is undeniably the product of Eriksen and McAuliffe's (2003) Counseling Skills Scale. The number and significance of modifications, however, demand that it stand or fall on its own reliability and validity studies.

Reliability and Validity of the UCSA

An initial examination of the reliability and validity of the UCSA was undertaken in a pilot study performed by the dissertation author and his committee chair. In this study, the professor teaching a Counseling Skills course and two doctoral student co-teachers employed the

UCSA independently to review and score the video recordings students submitted as part of their final exam. The recordings were of 15-minute-long mock counseling sessions in which the submitting student assumed the role of the counselor and another student assumed the role of the client. Students were instructed to demonstrate within the 15 minutes all of the basic skills they had been taught during the course. Scores for individual items were recorded, and Overall Scores and the Attending and the Basic Listening subscales results were calculated. Individual raters' scores from this administration were used to calculate IRR coefficients using intraclass correlation (ICC; Shrout & Fleiss, 1979). These results also served as a pre-test as the students entered practicum and were compared by dependent samples *t* test with a second administration of the instrument 10 weeks later at the completion of practicum. The practicum faculty supervisor, a doctoral student supervisor, and the student's on-site supervisor served as raters during the post-practicum assessment by reviewing a taped session with a client who had consented to being taped for this purpose.

For the pre-test administration of the UCSA, two sections of the Counseling Skills course (N = 20) participated in the research. All participants were female. Sixteen were White, three were African-American, and one was Asian. Eleven were in the clinical mental health track of instruction, and the remaining nine were in the school counseling track. Two students did not complete the course successfully. One of these students failed to complete the course because of low academic performance. Her performance ratings (0.4 sd above the class mean) would have been sufficient to advance her to practicum. The other student was not advanced to practicum because her performance ratings (1.45 sd below the class mean) were below the advancement threshold. However, because these students attended the full term of the course and completed all the assignments, their scores remained part of the data set.

For the post-practicum administration, the participants were limited to those students in the clinical mental health track (N = 11, 11 females, nine Caucasians, one African-American, one Asian). The nine students in the school counseling track were not included in the pre-post comparison because the practicum faculty supervisor chose not to administer the UCSA to those students.

Results

Using the data from the first administration of the instrument, an intraclass correlation coefficient of .80 was calculated. Shrout and Fliess (1979) categorize this as evidence of strong interrater reliability. Internal consistency was also supported as the data produced a Cronbach's alpha = 0.82.

Face validity was supported through consultation with experienced faculty and supervisors during the revision process. A pre-practicum to post-practicum dependent samples *t*-test gave evidence of construct validity. A comparison of Overall Scores indicated significant improvement in demonstrated skills ($t = 10.8$, $p < .0001$) with each of the subscales indicating similar significant gains. The mean pre-practicum Overall Scores averaged 3.0 compared to mean post-practicum Overall Score averages of 4.3. The effect size of this 1.3 point difference was demonstrated to be large by both Cohen's $d = 3.25$ and $r^2 = .92$. These initial results supporting the reliability and validity of the UCSA are sufficient to conclude that the instrument is ready for further testing and for use in this study.

Conclusion

The literature chronicles many different methods for teaching counseling skills. Carkhuff (1969) and Ivey's (1973) work to operationalize and define the discrete skills counselors need has not only stood the test of time but has experienced a renaissance in Smaby and Maddux's Skilled Counselor Training Model (2011) and Young's Megaskills (2009) model. Although each of these pedagogical models has accompanying assessment instruments, to date little attention has been paid to developing a basic skills assessment that can demonstrate cross modality validity.

Regardless of the skills teaching model employed, scoring a video recording of beginning counselors' sessions has become the most common method of skills assessment. Because this feedback is powerfully influential and because the outcome of the assessments often bears directly on the students' success in their academic programs, care must be exercised to ensure accurate and reliable scoring. A common safeguard is to use multiple raters to mitigate the effects of any one rater's biases. This procedure raises the need to ensure that there is sufficient agreement between the raters to consider their scoring statistically reliable. The literature has identified two factors that can have a significant impact on interrater reliability. First, task specific training of raters can improve the reliability of their scoring. Second, narrowing their focus by making them responsible for scoring across fewer domains can improve their interrater reliability. The question that remains is how can the interaction of these two factors be used to ensure acceptable interrater reliability with minimal training time? In other words, can a scoring procedure that limits the raters' focus be employed in conjunction with brief instrument-specific rater training to provide a more efficient path to robust interrater reliability? This study is designed to explore that question.

CHAPTER III: METHODOLOGY

This study was designed as an investigation of the effects of instrument-specific rater training on interrater reliability and counseling skills performance differentiation using the Universal Counseling Skills Assessment (UCSA). Participants assumed the roles of raters and were randomly assigned to one of four approximately equally sized groups: (a) expert raters without instrument-specific training scoring a demonstration of low counseling skills performance (this group will be designated the *UL Group*), (b) trained expert raters scoring a demonstration of low counseling skills performance (this group will be designated the *TL Group*), (c) expert raters without instrument-specific training scoring a demonstration of high counseling skills performance (this group will be designated the *UH Group*), and (d) trained expert raters scoring a demonstration of high counseling skills performance (this group will be designated the *TH Group*).

The training the TL and TH groups received included an overview of the UCSA and the basic counseling skills it assesses and instruction using the UCSA in a *dynamic scoring* environment. Dynamic scoring is a process of rating students' demonstration of basic counseling skills by categorizing each counselor response and assigning a score reflective of the overall quality and effectiveness of that individual response. Each of these raw scores was logged into a database and composite scores were calculated from this data. The categories to which the responses are assigned corresponded to the individual scoring items on the UCSA. The composite scores produced by the dynamic scoring procedure corresponded to the three subscale

scores and the overall score from the UCSA. As the dynamic scoring interface is an original production, the only exposure any participant had to this method of scoring came from the instruction provided in the study. As the UCSA is a highly modified and unpublished derivation of the Counseling Skills Scale (CSS, Eriksen & McAuliffe, 2003), the only exposure any participant had to the instrument came from its use in this study.

The untrained expert groups watched a video recording demonstrating either a low or high counseling skills performance (depending on the group assignment) and scored the performance by completing the 12-item UCSA. The trained expert groups first watched a training video demonstrating the use of the UCSA in a dynamic scoring environment (using the dynamic scoring interface) to score either the low or the high counseling skills performance video depending on group assignment. The same low performance video was scored by both the untrained expert group and the trained expert group. The same high performance video was scored by both groups assigned to the high counseling skills performance groups. To control for the potentially confounding variables of race, ethnicity, gender, personality, and physical appearance biases, the role of the counselor was played by the same person in both the low and high performing videos, and the role of the client was played by the same person in both videos. The general subject matter of the sessions was also the same.

Sample

In this study, the investigator sampled the population comprised of counselor educators, counselor education doctoral students, and Master's level counseling practitioners who were certified as clinical supervisors. For the purpose of this study, counselor educators were defined

as those holding an earned Ph.D. or Ed. D. in Counselor Education and Supervision or a closely related field (i.e., Counseling Psychology or Educational Psychology). Active employment as a professor was not required for a person to qualify for this study. Master's level supervisors were those who have been certified as clinical supervisors by the state in which they practiced or who held the Approved Clinical Supervisor (ACS) credential from the Center for Credentialing and Education. Qualified doctoral students were those who had completed a doctorial level course in supervision.

Materials

Consent to Participate in Research

The *Consent to Participate in Research* form was administered electronically via the first web page the participants encountere when entering the study's website. A printed copy of the form is included in Appendix A. The consent form outlined the study for potential participants by providing the following: (a) a description of the study, (b) an explanation of the risks and benefits associated with participation, (c) an explanation of the financial costs and payments associated with participation, (d) an explanation of the limitations of confidentiality, (e) an explanation of a participant's right to withdraw, and (f) information regarding The University of Mississippi's Institutional Review Board approval.

The Demographics Form

A demographics form was administered electronically via the second page of the study's web site. A printed copy of the form is included in Appendix B. This form was used to collect the participants' age by direct request for that information and their gender through the following provided choices: (a) male and (b) female. The form captured the participants' race through the following choices: (a) African-American, (b) Asian, (c) Caucasian, (d) Latina/Latino, (e) Middle Eastern, (f) Native American/First People, (g) Pacific Islander, and (h) Other. The form captured the participant's professional standing through the following choices: (a) Master's in counseling (state certified supervisor), (b) Master's in counseling (ACS), (c) Doctoral student having completed a supervision practicum or internship, or (d) Ph.D. or Ed.D. in counselor education or closely related field. The form captured the participants' years of experience by direct request for that information. The form collected the method by which each participant was trained in counseling skills through the following choices: (a) Carkhuff's Human Resource Development Model, (b) Egan's Skilled Helper Model, (c) Roger's Client-centered Model, (d) Ivey's Microcounseling Model, (e) Smaby & Maddux's Skilled Counselor Training Model, (f) Young's Megaskills Model, or (g) Other. Finally, the form collected the method each participant would use to teach counseling skills if he or she were to teach that course next semester. Participants chose from the following: (a) Carkhuff's Human Resource Development Model, (b) Egan's Skilled Helper Model, (c) Roger's Client-centered Model, (d) Ivey's Microcounseling Model, (e) Smaby & Maddux's Skilled Counselor Training Model, (f) Young's Megaskills Model, or (g) Other.

Low and High Performance Videos

The low performance video and the high performance video were alike in many respects. The same person served as the counselor in both videos. The same person served as the client in both videos. The camera angles, lighting, set, backdrop, sound levels, and all other technical aspects of the videos were as identical as is possible to produce. In each case, the client began the session with the same presenting concern. The low performance video was 14:43 long and the high performance video was 11:12 long.

In the low performance video, the counselor demonstrated the proper attending skills of *good appearance and body language, good eye contact*, and the *use of minimal encouragers*. Although the counselor was a personable discussion companion with good general social skills and an apparent genuine interest in the client, he only marginally demonstrated any counseling skills. His responses were limited to direct questions regarding the content of the client's statements, explaining things to the client, and well-meaning expressions of opinions and advice.

In the high performance video, the counselor demonstrated proper attending skills: (1) *body language & appearance*, (2) *eye contact*, (3) *minimal encouragers*, (4) *vocal tone*, and (5) *verbal tracking*; quality basic listening skills: (1) *selective attending*, (2) *directs and encourages client to talk*, (3) *paraphrasing (reflections of content)*, (4) *reflections of feeling/meaning*, and (5) *summarizing*; and appropriate deepening skills of *using immediacy* and *challenging/pointing out discrepancies*.

Training Video and Quiz

The training video was 32:11 long and included a brief introduction to skills assessment, an introduction to the UCSA, and instruction in the use of the dynamic scoring interface. The

instruction included an explanation of each item of the UCSA with corresponding video demonstrations of each assessed skill. The training video was followed by a five-question multiple choice quiz on the material covered. Any question answered incorrectly prompted a brief review of the question's subject and an explanation of why each response is correct or incorrect. Participants were required to answer all questions correctly before being allowed to advance in the study. For study participants who desired to receive continuing education credit, the result of this quiz was used to demonstrate their mastery of the information covered in the training session and justify the awarding of one supervision level CEU. A copy of the quiz with the pop-up messages participants received in response to incorrect answers is provided in Appendix D.

The Universal Counseling Skills Assessment (UCSA)

Instrument Derivation.

The UCSA is a revision of Eriksen and McAuliffe's (2003) Counseling Skills Scale (CSS). The CSS was developed with the intention of creating an instrument that: (a) was valid and reliable, (b) relied on observations of actual in-session performance, (c) had face validity and ease of use, (d) used expert judges' ratings, and (e) required qualitative judgments as to the contextual appropriateness of the skills demonstrated (p. 123). Their final product is a 22-item measure, subdivided into six subscales by which demonstrated skills are rated on a 5-point Likert-type scale by expert raters.

The revisions of the CSS that produced the UCSA centered on two major changes—a plainer, more intuitive scoring system that is linked directly to a student's readiness for advancement from skills training to practicum and the reduction of the number of items from 22 to 12. The first revision clarified the scoring system by changing the -2 to +2 scale to a 5-point scale that is scored from 1 to 5 removing the unnecessary complication of a scale encompassing both negative and positive numbers. The new scale was also redefined as follows: (1) *Not observed when required by the situation,* (2) *Demonstrated insufficiently to proceed to practicum/internship,* (3) *Demonstrated sufficiently to proceed to practicum/internship but inconsistently,* (4) *Demonstrated sufficiently and consistently enough to proceed to practicum/internship,* and (5) *Demonstrated at a level normally indicative of one who has at least completed internship.* A sixth non-scoring category continued to identify skills that are (N/O*) not observed but not required for the situation.*

The second major revision to the CSS was to limit its scope of assessment. This was considered desirable because some of the items of the CSS assessed skills that were beyond the scope of a beginning skills course. Additionally, some items on the CSS assessed skills that are usually associated only with a specific model of skills instruction. After consulting with faculty and doctoral students who had been instrumental in teaching the skills course, the UCSA was limited to 12 items divided into three subscales.

The first subscale, *attending,* is comprised of five items: (1) *body language & appearance,* (2) *eye contact*, (3) *minimal encouragers*, (4) *vocal tone*, and (5) *verbal tracking*. Descriptions of each of these items appear on the form to aid raters in understanding the specific elements of each item.

The second subscale, basic listening, is also comprised of five items: (1) *selective attending*, (2) *directs and encourages client to talk*, (3) *paraphrasing (reflections of content)*, (4) *reflections of feeling/meaning*, and (5) *summarizing*. Again, each item is accompanied by an explanation. For example item 10, *summarizing*, is explained as "makes statements at key moments in the sessions that capture the overall sense of what the client has been expressing."

The third subscale, *deepening the session*, is comprised of only two items: (1) *using immediacy* and (2) *challenging/pointing out discrepancies*. Descriptions of these items include sample statements that students might use. For example, item 12, challenging/pointing out discrepancies is explained by the example, "You expect yourself to do ___ when facing the problem of ___ but you do ___ instead. When this happens, you feel ___ about yourself."

Reliability and Validity.

An initial small study (n = 11) has indicated strong interrater reliability (ICC = .80) and good internal consistency ($\alpha = 0.82$). Face validity has been somewhat supported by a pre-practicum to post-practicum test indicating significant skills improvement ($t = 10.8$, $p < .0001$, $r^2 = 0.92$).

Pilot Study.

Although initial support for reliability and validity of the UCSA is good, the small sample size and single evaluation constitute significant limitations. Additionally, interrater reliability checks have been limited to those involved in the development of the instrument and

test-retest comparisons have been limited to pre and post practicum data. To address these limitations, a pilot study was completed with 37 counselors-in-training. These students were recruited from three sections of a Counseling Skills course taught by two different professors and five different graduate assistants. Two different models of counseling skills instruction were employed. One professor taught using Carkhuff's model (2000) and the other professor used Young's Mega Skills model (2009). The data from this pilot study produced an IRR coefficient of ICC = 0.768 with Cronbach's alphas ranging from 0.725 to 0.811. However, a dependent samples *t*-test comparing the mid-term results with the final results did not report a significant difference ($t = 0.704$, $p = 0.484$). This *t*-test fails to provide continuing support for the construct validity of the UCSA.

During this dissertation study, the UCSA was administered electronically through the same website that delivered the training video and the low and high performance videos. A printed copy of the instrument has been included in Appendix C.

Experiment Design

A dual post-test-only control-group design was employed to explore the effect of trained dynamic scoring on interrater reliability and counseling skills performance differentiation. Visually, this model can be represented as the following:

Group UL R----------O_L

Group TL R--------X-------O_L

Group UH R----------O_H

Group TH R--------X-------O_H

With Group UL = the group of untrained raters who score the low counseling skills performance video, Group TL = the group of trained raters who score the low counseling skills performance video using dynamic scoring, Group UH = the group of untrained raters who score the high counseling skills performance video, Group TH = the group of trained raters who score the high counseling skills performance video using dynamic scoring, R = the random group assignments, O_L = the scoring of the low counseling skills performance video, O_H = the scoring of the high counseling skills performance video, and X = the exposure to the dynamic scoring training video.

Variables

The dependent variables measured in this study were the scores assigned by the raters to the low counseling skills performance video (S_L), the scores assigned by the raters to the high counseling skills performance video (S_H), and the interrater reliability coefficient produced by the raters' data (IRR). The independent variables were participant training status (UL and UH = untrained raters, TL and TH = trained raters) and degree held by the participant.

Procedures

Recruitment.

Recruitment was accomplished by invitations extended through the CESNET listserv, emails to potentially interested parties, direct contact with faculty members in various counselor

education and supervision programs, and face-to-face invitations at conferences (state, regional, or national). These invitations included a website link through which one could enter the study. Follow-up invitations were issued through listserv groups two weeks and five weeks into the data collection phase.

Participation in this study was anonymous to the degree that no identity specific information was requested by the demographics form. Participants had the option of identifying themselves for the purpose of receiving a CEU, but such identification was not required for participation. When identifying information was provided for CEU purposes, that information was linked only to the training quiz outcome. Identifying information was not linked to the responses given in the study, and no individual's identifying information has been included in any of the study's reports.

A series of *a priori* G-Power (Faul, Erdfelder, Buchner, & Lang, 2009; Faul, Erdfelder, Lang, & Buchner, 2007) analyses was performed, one for each type of statistical analysis proposed. The results (ANOVA, $\alpha = .05$, power = .80, $f^2 = .25$; bivariate correlation, $\alpha = .05$, power = .80, $r^2 = .50$; independent samples t-test, $\alpha = .05$, power = .80, $d = .50$) indicate a sample of four equal sized groups of $n = 69$ with a total sample size of $n = 276$ as being ideal.

Each potential participant was provided a link to the study's website. Upon entering the web site, the participant was presented with the *Consent To Participate in Research* form. Those who agreed to participate (indicated by clicking on a button embedded in the form) were randomly assigned by the site to one of the four groups (UL, TL, UH, TH). This random assignment produced groups of UL = 17, UH = 21, TL = 16, and TH = 20 for a total of N = 74.

Rater Scoring.

Those assigned to the untrained groups were shown a video of a counseling session and asked to score the performance of the counselor using the UCSA (provided on the screen at the same time as the video being scored). The UL group scored a video where the counselor was instructed to be personable but to demonstrate counseling skills only marginally. The UH group scored a video where the counselor was instructed to do his best to demonstrate all the basic counseling skills. Once the participants completed scoring the video, they submitted their scores using the *Submit* button at the bottom of the form. Participants were then directed to a website where they were offered an opportunity to earn a supervision CEU by providing the necessary personal information, watching the training video about the use of dynamic scoring, and taking the five-question multiple choice quiz. Participants were not required to seek the offered CEU to be a part of the study.

Those assigned to the trained rater groups were shown the training video about the UCSA and the use of dynamic scoring. When they successfully completed the five-question multiple choice quiz, they were presented either the low performance video or the high performance video, depending on group assignment, and asked to score it using the dynamic scoring interface for the UCSA. When the trained rater participants completed scoring the video, they exited the study. If they desired supervision CEU, they were asked to submit the necessary personal information. Participants could withdraw from the study at any time by closing their browser.

Hypotheses and Analyses

To aid in analyzing the data, descriptive statistics were calculated for all groups. Interrater reliability coefficients were also calculated using intra-class correlation for the untrained raters (IRR_U), the trained raters (IRR_T), the raters of the low performance video (IRR_L), and the raters for the high performance video (IRR_H).

This study was designed so that the investigator could test the following hypotheses:

H_{O_1}: There is no significant difference in the mean score on the UCSA (S_{low}, S_{high}, S_{all}) by training status (Untrained raters, Trained raters) or video level (low, high).

This hypothesis was tested using a 2-factor ANOVA (training level X video level).

H_{O_2}: There is no significant difference in IRR coefficient by training status (untrained raters, trained raters).

This hypothesis was tested by first performing a Fisher's *r to z* Transformation on the IRR coefficient (calculated by intraclass correlation) and then testing for significant difference with an independent samples *t*-test.

H_{O_3}: There is no significant difference in IRR_{UR} coefficient by video performance level (low, high).

This hypothesis was tested by first performing a Fisher's *r to z* Transformation on the IRR coefficient (calculated by intraclass correlation) and then testing for significant difference with an independent samples *t*-test.

H_{O_4}: There is no significant difference in IRR_{TR} coefficient by video performance level (low, high).

This hypothesis was tested by first performing a Fisher's *r to z* Transformation on the IRR coefficient (calculated by intraclass correlation) and then testing for significant difference with an independent samples *t*-test.

It was hoped that the distribution of participants across the various demographic categories would be equal enough to allow for a comparative analyses of their scores. However, only the demographic category *Degree Held* was sufficiently equally represented to do comparison testing. This led to the following fifth hypothesis:

H_{O_5}: There is no significant difference in mean UCSA scores by Degree Held.

Conclusion

The purpose of this study was to explore the effects of instrument specific rater training on interrater reliability and low counseling skills performance differentiation. The training consisted of instruction in the use of an instrument and a scoring procedure both of which are original, unpublished products. Every participant, therefore, had equal exposure to these elements. The sample was limited to the population of interest and randomly assigned to the study groups according to a dual post-test-only controlled design. The resultant data was analyzed to compare the scores of trained raters to untrained raters on both high and low performance demonstrations of counseling skills. Statistical comparisons were made of the IRR of trained and untrained raters across the spectrum of high and low performance demonstrations of counseling skills. The analyses and their results are further described in chapter four.

CHAPTER IV: RESULTS

Introduction

This study was designed to explore two areas of counseling skills assessment. The first area was to investigate the effect of instrument-specific rater training on interrater reliability (IRR). The second area was intended to determine the effect of instrument-specific training on the ability of raters to differentiate between high and low counseling skills performances. Participants were randomly assigned to one of four groups in the study. The *UL Group (untrained rater viewing a low performance video)* and the *UH Group (untrained rater viewing a high performance video)* received no training prior to using the Universal Counseling Skills Assessment (UCSA) to score a video recording of a low performing counselor (a counselor not following a model based on basic counseling skills) and a high performing counselor (a counselor following a model based on basic counseling skills), respectively. The *TL Group (trained rater viewing a low performing video)* and the *TH Group* (trained rater performing a high performing video) received 32 minutes of training (via online video instruction) that addressed the basic counseling skills measured by the UCSA and the use of the Dynamic Scoring Interface (DSI) before scoring the same low and high counseling performance videos.

The data collected from the participants included the following demographic identifiers: gender, race, age, highest degree/certification attained, years of professional experience, the counseling skills method by which they were trained, and the counseling skills training method

they would use if they were to teach a counseling skills course. The untrained group completed the USCA, a 12-item Likert-style assessment instrument divided into three subscales (Attending, Basic Listening, and Deepening). The trained groups scored the same videos as the untrained groups by completing the UCSA through the DSI (a computer-based interface that allows scoring of a counselor's performance on a response-by-response basis). The DSI scoring produced composite scores for the same three subscales as the UCSA (Attending, Basic Listening, and Deepening). The data were analyzed as described below, using two-way ANOVA, Fisher's *r* to *z* transformations, and independent samples *t*-tests.

Descriptive Statistics

A total of 87 potential participants logged into the study's website. Of those 87 entries, 85% of the participants successfully completed the study giving a total $N = 74$. Participants self-reported biological sex (25 males and 49 females), age ranging from 26 to 70 years ($M = 41$, $s = 12$), and professional experience ranging from 1 year to 30 years ($M = 7$, $s = 6$). Of the 74 participants, 14 self-identified as African-Americans, one as Asian, 55 as Caucasian, one as Latina. Three other participants identified their race as *Other*. In the areas of professional identity, 14 held master's degrees in counseling and were certified by their states as clinical supervisors, four held master's degrees in counseling and the ACS certification, 17 were doctoral students in counselor education, and 39 were counselor educators holding terminal degrees. The specific training model experienced by the participants was: 27 trained with Carkhuff's model,

five with Egan's, 15 with Roger's, eight with Ivey's, two with Young's, and 17 with some other unidentified model. When responding to the question about which model participants would use when training others, the responses varied as reported: 18 preferred Carkhuff's model, four would use Egan's, 22 would use Roger's, nine would use Ivey's, nine would use Young's, and 12 would use some other unidentified model. No participant identified as having been trained by Smaby and Maddux's model and no participant chose to use Smaby and Maddux's model to teach skills.

Subscale and total scores were calculated from the untrained groups' UCSA raw scores. An *Attending* (ATT) subscale score was calculated by averaging the entries from the instrument's first five items (Body Language & Appearance, Eye Contact, Minimal Encouragers, Vocal Tone, Verbal Tracking). A *Basic Listening* (BL) subscale score was calculated by averaging items six through ten (Selective Attending, Directions and Encouraging Client To Talk, Paraphrasing, Reflecting Feeling/Meaning, Summarizing). A *Deepening* (DEEP) subscale score was calculated by averaging the last two items of the UCSA (Using Immediacy and Challenging/Pointing Out Discrepancies). Total scores were calculated as an arithmetic mean of the three subscale scores.

Subscale scores were calculated from the trained groups' UCSA/DSI submissions. An *Attending* score was calculated by multiplying each Direct To Talk (DTT) response by the quality rating assigned to it, totaling all the products, subtracting 1/2 point for each missed response (MIS), and averaging this result with the three statically scored items (Body Language & Appearance, Eye Contact, and Vocal Tone). A *Basic Listening* score was calculated by

multiplying each Reflection of Content (ROC), Reflection of Feeling (ROF), Reflection of Meaning (ROM), and Summary (SUM) response by the quality ratings assigned to them, subtracting one point for each Judgmental (JUD), Missed (MIS) response, negative ROC, negative ROF, negative ROM, and negative SUM and dividing the total by the number of positive responses made for each item comprising the subscale. Total scores were calculated as an arithmetic mean of the three subscale scores. Both the scores from the untrained groups and the scores from the trained groups were then converted to *z*-scores for comparative purposes. The calculated ranges, mean raw subscale and total scores, mean subscale and total *z*-scores, and standard deviations are reported by group in Table 1.

Table 1

UCSA Scores and Standard Deviations by Subscale and Group

	ATT			BL			DEEP			TOT		
Group	$\overline{X}_{Raw}$	$\overline{X}_z$	*s*	$\overline{X}_{Raw}$	$\overline{X}_z$	*s*	$\overline{X}_{Raw}$	$\overline{X}_z$	*s*	$\overline{X}_{Raw}$	$\overline{X}_z$	*s*
UL	3.2	-.04	0.2	2.8	-0.9	0.3	2.3	-.03	0.5	2.8	-0.7	0.2
UH	3.5	0.3	0.7	3.7	0.6	0.5	2.7	0.3	1.0	3.4	0.4	0.6
TL†	5.0	0.2	1.1	1.9	-0.9	0.7	1.3	-1.1	0.9	2.7	-0.9	0.6
TH†	4.6	-0.1	1.0	3.0	0.5	0.3	3.9	0.6	0.8	3.9	0.6	0.5

Note. ATT = Attending Subscale; BL = Basic Listening Subscale; DEEP = Deepening Subscale; TOT = Total Score; $\bar{X}_{Raw}$ = Mean Raw Score; $\bar{X}_z$ = Mean *z*-Score; *s* = standard deviation; UL = untrained raters scoring the low performance video; UH = untrained raters scoring the high performance video; TL† = trained raters scoring the low performance video using the DSI; TH† = trained raters scoring the high performance video using the DSI.

Sample Size

A series of *a priori* G-Power (Faul, Erdfelder, Buchner, & Lang, 2009; Faul, Erdfelder, Lang, & Buchner, 2007) analyses indicated an ideal sample size of *N* = 276. These analyses were based on a desired power of at least 0.80 for each test while making assumptions regarding effect size and within group variances. The actual sample size (*N* = 74) fell far short of the ideal (*N* = 276) calling into question whether the planned tests would have sufficient power. However, a review of the *a priori* analysis revealed that a much smaller sample size was acceptable for all of the planned tests except the independent samples *t*-test of hypothesis three. The decision was made to perform the *t*-test analysis of Hypothesis Three and to perform a *post hoc* analysis of actual achieved power.

A *post hoc* analysis of achieved power using the G-Power calculator (Faul, Erdfelder, Buchner, & Lang, 2009; Faul, Erdfelder, Lang, & Buchner, 2007) reported excellent power for the testing of hypotheses one and two (power > 0.90). The power for the independent samples *t*-test used to test hypothesis three (power = 0.16), four (power = 0.40), and five (power = 0.18) was much lower. Had these test resulted in statistically significant outcome, the findings would have been questionable. The structure of hypothesis five was dependent on the distribution of the

participants across the requested demographic categories. As such, this analysis was not part of the *a priori* power analysis.

Assumptions of Related Statistical Analyses

Two-Factor ANOVA

This study's research design called for the use of a two-factor, 2 X 2 ANOVA with participant training level (untrained, trained) and video performance level (low, high) constituting the two independent variables and the dependent variable being defined as the mean total *z*-score for each group. According to Gravetter and Wallnau (2004), Hinkle, Wisersma, and Jurs (2003), and Tabachnick and Fidell (2006), six assumptions are associated with a two-factor ANOVA.

The first assumption is that the dependent variable must be measured and reported as continuous data. All performance scores in this study consist of interval level data meeting the first assumption. The second assumption is that the two independent variables must be categorical in nature and independent of each other. As the two independent variables in this study were participant training level (untrained, trained) and video performance level (low, high) and assignments to these various groups were random, this assumption was met. The third assumption is that there must be independence of observation between the groups. As this study's design is fully nested (no participant was assigned to more than one group and each

group is uniquely constituted), this assumption was met. The fourth assumption is that the data is free from significant outliers. The data were examined using a squared-difference-from-the-mean procedure to identify any outliers. Although there were two subscale scores that qualified as outliers, they did not affect the total score significantly enough to cause the total score to be classified as an outlier. To honor the effort of these participants and the integrity of the overall data set, I chose to retain these scores. The fifth assumption is that the dependent variable is normally distributed across the population. The data were analyzed using the Shapiro-Wilk test for normality and found to be normally distributed ($W = 0.982$, $p = 0.361$). This finding was further confirmed by an examination of the distribution's histogram and a calculation of skewness (-0.115) and kurtosis (-0.634). The final assumption associated with a two-way ANOVA is that all groups' variances are homogeneous. To test this assumption, all cell variances were checked using Levene's test for homogeneity of variance which was not significant ($F = 3.752$, $p = .062$). As the Levene's test for homogeneity of variance tests the null hypothesis, a non-significant result indicates that all variances may be safely assumed to be homogeneous.

Independent Samples *t*-test

The assumptions associated with independent samples *t*-tests are similar to the six assumptions associated with the two-factor ANOVA. The first assumption, that the dependent variable is comprised of continuous data, is the same and has, therefore, been met. The third,

fourth, fifth, and sixth assumptions regarding the independence of observations, lack of significant outliers, a normally distributed dependent variable, and homogeneity of variance between the groups respectively are identical and have, therefore, been met.

The second assumption, however, is different. Unlike the two-way ANOVA, an independent samples *t*-test is used for analyses of data representing one dependent variable and only one independent variable. Therefore, an assumption exists that the one independent variable consists of two categorical levels that are independent. Due to the fully nested study design and the random assignment of participants into study groups, this assumption was met.

Fisher's *r* to *z* Transformation

Testing the relative strength of correlation between two samples is not as simple as might be assumed. To test for a significant difference between two coefficients, both must first be transformed to a value appropriate for comparison. This process was first suggested by Fisher (1915) and is referred to as Fisher's *r* to *z* transformation. Only two assumptions accompany this procedure. First, the coefficients being transformed must be correlation coefficients limited in their range from -1.000 to 1.000. Second, the correlation coefficients must be independent. The data from this study satisfies the first assumption in that the coefficients to be transformed were calculated by case two intraclass correlation analysis that resulted in correlation coefficients with the specified range. The second assumption was met by the fact that all correlation comparisons were, by study design, between-groups comparisons with no participant being assigned to more

than one group. All *r* to *z* transformations were completed using a *Transformation of r to Fisher's z* table (Kenny, 1987, p. 356). All tests of significant difference between Fisher's *z* values were calculated using Lowry's (2013) *Significance of the Difference Between Two Correlation Coefficients* calculator.

All the assumptions for the use of two-way ANOVAs, independent samples *t*-tests, and Fisher's *r* to *z* Transformations were met and the specified analyses were used to test the study's hypotheses. A variety of descriptive statistics, not associated with any formal assumptions, were also used in various calculations.

Data Analysis

Testing of Hypothesis One

The first hypothesis is stated in the null form: H_{O_1}: There is no significant difference in the mean score on the UCSA by training status of the rater. This hypothesis was based on studies (Lepkowski, Packman, Smaby, & Maddux, 2009; Little, Abney, Packman, Smaby, & Maddux, 2005; McCullough et al., 2003; Urbani et al., 2002) that indicated rater training has an effect on a rater's ability to assess accurately counseling skills. Because that effect might differ depending on the performance level demonstrated on the video recording, a two-way ANOVA was used to test the mean scores assigned by raters (untrained, trained) to two video recordings of counseling sessions (low skills performance, high skills performance). Analysis revealed no main effect of

training level [F(1, 70) = 0.007, p = 0.935, partial η^2 < 0.001], a main effect of video performance level [F(1, 70) = 43.67, p < 0.001, partial η^2 = 0.384], and no significant training level by video level interaction [F(1, 70) = 1.077, p = 0.303, partial η^2 = .015].

Table 2

Training Level X Video Level Two-Way ANOVA for UCSA/DSI Scores

Source	*Df*	*F*	*p*	η^2
(A) TrainLev	1	0.007	0.935	0.00
(B) VidLev	1	43.67	0.000*	0.384
A X B (interaction)	1	1.077	0.303	0.015
Error (w/in groups)	70			

Note. TrainLev = Training level of the rater; VidLev = Performance level of the video being scored.
* = significance at the α = 0.01 level

As the two-way ANOVA is an omnibus test, the results indicate that all of the groups compared are not equal. A series of pairwise comparisons revealed that the only means that differed were the means of the scores assigned to the low performance video compared with the means of the scores assigned to the high performance video which was expected based on the study design. As these analyses did not provide sufficient evidence of a significant difference in means by rater training status, the null for Hypothesis One was not rejected. This indicates that there was not a significant difference between the untrained raters and the trained raters when restricted to either a low or a high performing video. A significant difference was indicated

between the low and high performing videos adding validity to the experiment, but no differentiation was noted between the two levels of raters (untrained, trained).

Testing of Hypothesis Two

Just as the above studies indicated a correlation between rater training and scoring accuracy, research (McCullough et al., 2003; Wang, 2010) has also shown a link between rater training and the reliability of rater scoring. Hypothesis Two [H_{O_2}: There is no significant difference in IRR coefficient by training status (untrained raters, trained raters)] was developed to facilitate the testing of the correlation between rater training and the reliability of rater scoring.

Testing of this hypothesis was performed by first calculating the correlation coefficients (see Table 2) for the groups of the untrained raters [ICC (2, 38) = 0.888, $p < 0.001$] and trained raters [ICC (2, 36) = 0.962, $p < 0.001$]. These coefficients were then transformed to Fisher's *z* values (0.888 to 1.3758, 0.962 to 1.9459) and tested for a significant difference ($z = 2.31$, $p = 0.021$). The critical values in testing Fisher's *z* are ±1.96 for an alpha level set at $\alpha = 0.05$ and ±2.58 for an alpha level set at $\alpha = .01$. The Fisher's *z* result provides support for the alternative hypothesis indicating that there is sufficient evidence of a significant difference in IRR by rater training level when tested at the $\alpha = 0.05$ level. This test result caused a rejection of this study's second null hypothesis indicating that the level of agreement in the trained raters' scores was significantly stronger than the level of agreement in the untrained raters' scores.

Table 3

Interrater Reliability by Group

Group	ICC (2, *k*) coefficient	Significance	Cronbach's α
UL	0.898	$p < 0.001$*	0.907
UH	0.877	$p < 0.001$*	0.941
$TL^{\dagger}$	0.983	$p < 0.001$*	0.989
$TH^{\dagger}$	0.941	$p < 0.001$*	0.964
UR_{All}	0.888	$p < 0.001$*	0.924
TR_{All}	0.962	$p < 0.001$*	0.971

Note. UL = untrained raters restricted to the low performance video; UH = untrained raters restricted to the high performance video; $TL^{\dagger}$ = trained raters restricted to the low performance video using the DSI; $TH^{\dagger}$ = trained raters restricted to the high performance video using the DSI; UR_{All} = all untrained raters; TR_{All} = all trained raters.
* = Significance at the α = .01 level.

Testing of Hypothesis Three

The third hypothesis is stated in the null form: H_{O_3}: There is no significant difference in IRR_U coefficient by video performance level (low, high). This hypothesis and Hypothesis Four were developed to explore whether the reliability of raters' scoring changed depending on the quality of the counseling skills demonstration they were assessing. The testing of Hypothesis

Three was designed to examine the reliability of untrained raters' scoring of the low performance video compared to the reliability of other untrained raters' scoring of the high performance video.

To test this hypothesis, correlation coefficients were calculated for the groups of the untrained raters who scored the low performance video [ICC (2, 17) = 0.898, $p < 0.001$] and for the untrained raters who scored the high performance video [ICC (2, 21) = 0.877, $p < 0.001$]. These coefficients were transformed to Fisher's z values (0.898 to 1.4219, 0.877 to 1.3331) and tested for a significant difference ($z = 0.28$, $p = 0.780$). Because the critical values were not met, these analyses did not provide sufficient evidence of a significant difference in IRR coefficients by video performance level for untrained raters. The results indicated that the third hypothesis should be categorized as a failure to reject the null. This indicates that there was not a significant difference in the IRR coefficients produced by untrained raters who scored the low performance video compared to untrained raters who scored the high performance video.

Testing of Hypothesis Four

As mentioned above, Hypothesis Four was similar to Hypothesis Three. Whereas Hypothesis Three tested the reliability of untrained raters' scoring, Hypothesis Four provides for the formal comparison of the reliability of trained raters' scoring across the demonstrated skills performance spectrum. The fourth hypothesis, formally stated is: H_{O_4}: There is no significant difference in IRR_T coefficient by video performance level (low, high).

To test this hypothesis, the correlation coefficients were calculated for the groups of the trained raters who scored the low performance video [ICC (2, 16) = 0.983, $p < 0.001$] and the trained raters who scored the high performance video [ICC (2, 20) = 0.941, $p < 0.001$]. The resultant coefficients were transformed to Fisher's *z* values (0.983 to 2.2976, 0.941 to 1.7380) and tested for significant difference ($z = 1.72$, $p = 0.085$). As the critical value was not met, the result did not provide sufficient evidence of a significant difference in IRR coefficients by video performance level for trained raters. This indicated that the fourth hypothesis should be categorized as a failure to reject the null, signifying that there was not a significant difference in the IRR coefficients produced by trained raters who scored the low performance video compared to trained raters who scored the high performance video.

Testing of Hypothesis Five

The fifth hypothesis stems from a desire to explore the effect of the various demographic categories on the raters' scoring. In exploring the demographic variables, only the *Degree Held* category was represented with sufficient equality to perform valid comparison testing and that was true only after combining the *Master's (state certified)* ($n = 14$), *Master's (ACS)* ($n = 4$), and *Doctoral Student* ($n = 17$) participants into one new category (*Non-terminal Degree*). This new grouping created two categories (Non-terminal Degree, Terminal Degree) of sufficiently equal size ($n = 35$, $n = 39$) for comparison testing. The resultant hypothesis is stated in the null form: H_{O_5}: There is no significant difference in mean UCSA scores by *Degree Held*.

A Levene's test found these new groups to have sufficiently homogeneous variances ($F = .015$, $p = 0.902$) for testing by an independent samples *t*-test. This test indicated there was no significant difference in mean scores between the two degree groups ($t = 0.044$, $p = 0.300$). The result directed that a failure to reject this study's fifth hypothesis was appropriate. This indicated that there was not a significant difference in scoring recorded by those with non-terminal degrees compared with those who held terminal degrees.

Summary

The study's data met all of the assumptions for the use of two-way ANOVAs, independent samples *t*-tests, and Fisher's *r* to *z* Transformations. These analyses were then employed to test the five hypotheses.

When these analyses were applied to the data, null hypothesis one that explored differences in the mean scores by rater training status was supported. This failure to reject the null indicated that there is no significant difference in the video performance scores assigned by untrained raters and trained raters.

Null hypothesis two, which explored differences in the IRR coefficients by rater training status, was not supported. This rejection of the null indicated a statistically significant difference in the IRR coefficients produced by untrained raters when compared to the IRR coefficients produced by trained raters.

Null hypothesis three, which explored differences in the untrained raters' IRR coefficients by performance level of the video being scored, was supported. This failure to reject the null indicated that this study found no significant difference in the IRR coefficients produced by untrained raters when scoring either the low or high performing videos.

Null hypothesis four, which explored differences in the trained raters' IRR coefficients by performance level of the video being scored, was supported. This failure to reject the null indicated that this study found no significant difference in the IRR coefficients produced by trained raters when scoring either the low performing video or the high performing video.

Null hypothesis five, which explored differences in the mean scores by the degree held by the rater, was supported. Failing to reject the null indicated no significant difference in the scores assigned when comparing raters with non-terminal and terminal degrees.

CHAPTER V: DISCUSSION

Introduction

This study was intended to investigate interrater reliability of counseling supervisors when viewing videotapes of a mock counseling session. Participants were divided into four groups (UL, UH, TL, TH). Group comparisons of mean ratings and interrater reliability were conducted. Two specific research questions derived from the literature were investigated. The first question was focused on exploring the effect of instrument-specific rater training on interrater reliability (IRR). The second question was intended to investigate the effect of that training on the ability of raters to differentiate between high and low counseling skills performances. Statistical analyses of the data revealed no differences between the untrained participants and the trained participants in the scores they assigned to the performance video presented to them. The analyzed data also uncovered no differences in the reliability of the untrained raters' or the trained raters' scoring when comparing scores assigned to the low and high performing videos. Additionally, the data demonstrated no difference between the scores assigned by participants holding master's degrees and those holding terminal degrees. Finally, the data did indicate a significant difference in the reliability of the scoring of untrained raters when compared with trained raters.

This chapter contains a discussion of the study's statistical results, limitations to this study, and suggestions for future research. The chapter concludes with a discussion of how the study's findings may inform counselor education and the rating of counseling skills.

Hypothesis One

Studies have shown that counseling skills assessment, both the assessment of other's skills and self-assessment, can be learned and improved through training (Lepkowski, Packman, Smaby, & Maddux, 2009; Little, Abney, Packman, Smaby, & Maddux, 2005; Urbani et al., 2002). Hypothesis One was designed to explore the effect of brief, video-delivered rater training on raters' ability to score counseling skills from a video recording. This exploration was accomplished by comparing the means of the scores submitted by untrained and trained raters as they assessed the same counseling video performances. The results indicated that there was no difference in the scoring of the untrained raters compared to the scoring of the trained raters.

The results may indicate that the training offered was too brief to produce a difference. The *trained* participants watched a 32-minute training video and completed a five-question quiz on the material. Recent studies (McCullough et al., 2003; Wang, 2010) indicated that as much as 20 hours of training may be necessary before inter-rater reliability improvements were measurable. Although this research stresses the need for longer periods of training, the study context dictated brevity to execute the design. This brevity in the training is recognized as an inherent limitation in this study. More extensive training may produce a different result.

The lack of significant difference in scoring (untrained vs. trained) might also be attributed to the condition that the untrained raters were not *untrained* in the literal sense of the word. All the study participants held at least a master's degree in counseling and all had specific training or experience or both as clinical supervisors. In this study, untrained simply meant they did not receive the instrument-specific rater training associated with this study. Future research may utilize better controls for this confounding variable when exploring proposed differences.

Hypothesis Two

Before rater scoring can be tested for accuracy, consistency must first be demonstrated. Research has shown that rater training affects the reliability of rater scoring (McCullough et al., 2003; Wang, 2010). McCullough's et al. study was of particular interest in the formation of Hypothesis Two as two elements that positively affected rater reliability were identified. The first was increased rater training, and the second was limiting the raters' focus during the scoring process. This study incorporated both of these elements. Although the results of hypothesis one indicate that there is no difference in overall scoring between the raters, indicating that the training had no effect on total rater scores, hypothesis two was focused on the consistency of the raters scoring. Trained raters received 32 minutes of task-specific training prior to scoring a video, and that training prepared them to use the Dynamic Scoring Interface (DSI), a structured procedure for the grading of videos that was designed to limit the rater's focus to a single counselor response at a time. By including both of these elements into this study, I attempted to

test if the training had any consequential effect on interrater reliability or overall differentiation of scores.

Analysis of the data indicated that a significant difference did exist between the reliability of the trained and untrained raters scoring. The effect size was large (d = 1.15) and the test power was strong (power > 0.90) signifying that the difference between the two groups' IRR coefficients was not due to measurement error or chance. This result supports previous findings (McCullough et al., 2003; Schanche et al., 2010; Wang, 2010) and indicates that brief training combined with narrowing the focus of the judge to rating single counselor responses may present counseling programs with alternate methods for improving consistency of educators' and site supervisors' assessment of counseling skills.

One result of this study is specifically important. The high correlations of both sets of raters are encouraging for the use of the UCSA. Though not as strong as the IRR of the trained raters' data [ICC (2, 36) = 0.962], the untrained raters' data produced a strong reliability coefficient [ICC (2, 38) = 0.888]. The UCSA has been used routinely in a program evaluation context. The unpublished findings of these continual evaluations resulted in only moderate IRR statistics (ICC = 0.688, Meacham & Stoltz, 2012). The data produced in this study is much stronger, providing empirical support for the use of the instrument outside the context of program evaluation. However, caution must be used in interpreting this result. Having a larger number of raters (compared to a program evaluation environment) may have contributed to the instrument's strong performance by virtue of the fact that larger sample sizes tend to increase the

ICC coefficient. Additionally, recent revisions to the Likert rating descriptors (included in this study) may also have contributed to the improved performance of the instrument.

Hypotheses Three and Four

Hypotheses Three and Four were similar in design and intended purpose. Both examined whether the reliability of scoring varied depending on the level of counseling performance presented to the raters. These hypotheses explored the reliability of the untrained and trained raters' scoring respectively. An examination of the coefficients indicated that there were no differences in the reliability of the untrained or trained raters' scoring regardless of the video they were presented. All raters, untrained and trained, demonstrated the ability to score both low and high performances of counseling skills with high reliability.

These results combine well with the findings from Hypothesis Two to support the use of the UCSA in counseling programs. Whether assessing beginning counseling students' skills (who would tend to perform at a lower level) or those of students completing internship (who would be expected to perform at a higher level), the UCSA and the UCSA administered through the DSI appear to produce strong reliability.

Hypothesis Five

Because the level of rater training has been shown to affect rater performance (McCullough et al., 2003; Schanche et al., 2010; Wang, 2010) and educational attainment is intuitively linked to level of training, Hypothesis Five was designed to explore the effect of a rater's educational level on his or her ability to score counseling skills. This exploration was accomplished by dividing the participants into two groups based on educational demographics. A comparison of the group means indicated that there was no difference in the scoring of the two groups.

This result signals that recent changes to states' qualifications for certifying supervisors and the enhanced qualifications for receiving Approved Clinical Supervisor (ACS) certification have created a class of master's level supervisors whose assessment of counseling skills is consistent with counselor educators'.

Limitations of the Study

This study was designed to test the stated hypotheses while using established research methodology to ensure the internal and external validity of the results. Specific limitations of the study are discussed below.

The entirety of this study was conducted over the internet and was, therefore, subject to the inherent limitations of the medium – namely the lack of participant screening and monitoring. While the study's website link was unpublished and invitations to participate were extended only through listservs and emails to likely qualified participants, no screening of

participants occurred and no password protection was used to control entrance to the study. This means that unqualified people who happened upon the website could enter the study and participants who had completed the study could reenter. Data integrity checks were programmed into the website to protect the database from outside influence, but no control could be exercised over who entered the website. The possibility remains, therefore, that unqualified participants might have entered the website or qualified participants could have entered the study more than once. The data were inspected with these weaknesses in mind and no blatantly anomalous data was discovered.

A second limitation was the extreme brevity of the study's training. Other studies (McCullough et al., 2003; Schanche et al., 2010; Wang, 2010) conducted over longer time periods have provided 20 to 32 hours of rater training. This study was limited to one 32 minute training video in combination with the focus limiting effects of the DSI. Future studies might benefit from a graduated training schedule to explore the effects of longer periods of training on raters using the DSI.

Another limitation is that data for this study were collected for only 52 days. A longer data gathering period might have provided for more participants, which according to the law of large numbers (Ross, 2009), would improve the generalizability of the results and provide a more equal distribution across demographic characteristics.

The final and perhaps most significant limitation of the study was created by the confluence of four variables (training received in the study, training received outside of the study, familiarity with assessment instruments that employ a Likert-style scoring system, novelty

of the DSI). That untrained raters (due to their professional training and use of various assessment instruments) had more preparation for the task assigned them than did the trained raters (who due to the novelty of the DSI were limited to the 32 minutes of training they received in the study) is likely. The combination of training (previous training vs. study associated training) and scoring environment (Likert-style scoring vs. DSI scoring) may have resulted in untrained raters who actually had more training than did the trained raters. Future studies will need to avoid this situation.

Implications for Future Research

The first hypothesis was designed to test the effect of rater training on rater scoring and indicated there was no difference between the groups. Other studies have shown significant differences in rater scoring after longer rater training (Fitch, Gillam, & Baltimore, 2004; Wang, 2010). Future research may test variable training times to investigate the effect that training time has on rater scoring. Such a study might reveal both a minimum necessary level of training for raters and the point at which further training no longer produces improvement in rater performance.

The strong IRR performance of the UCSA (administered in both its traditional, static form and through the DSI) warrants further exploration. This study constitutes the first testing of the revised Likert scale descriptors. The revision was an attempt to tie the raters' judgment concretely to a common decision point. The rater decision called for was whether a skill had

been sufficiently demonstrated to indicate that the student should proceed to practicum. Whether this revision of the descriptors contributed to the instrument's strong showing is a matter for future study.

Because the trained raters differed by experimental condition (use of the DSI) from the untrained raters, the question arises, how much of the difference in group IRR is attributable to the training and how much is due to the use of the DSI. Answering that question with any degree of certainty is beyond the scope of this study. However, as the McCullough et al. (2010) and Wang (2010) studies recommend between 20 and 32 hours of training before reaching strong levels of IRR and considering that this study provided only 32 minutes of training, the use of the DSI may have had a strong effect on the reliability of the trained raters' scoring. The finding of significant differences in IRR statistics supports the construct that brief, focused training may have a positive effect on interrater reliability. Parsing out the effect of each would be a significant contribution to the professional literature.

Implications for Counselor Education

Acknowledging that the study has limitations, there are significant findings that inform counselor education. The result of testing Hypothesis One indicates that extremely brief (32 minutes) on-line training, no matter how specific, does not appear to improve the validity of rater scoring to acceptable levels. Thus, specific online training may need to be lengthened to levels suggested by the professional literature. However, the findings from testing Hypothesis Two do

show that brief, specific training and a scoring procedure that limits rater focus can result in improved interrater reliability. Based on these results, faculties might consider the combination of on-line training that includes a scheme for narrowing the raters' focus to improve the rater-training needs of faculty, doctoral students, and site supervisors. These efforts should thereby improve program evaluation through strengthened IRR.

The results of testing Hypotheses Three and Four indicate that the reliability of rater scoring did not vary depending on the performance level of the counseling skills being assessed. However, the testing procedure revealed that the UCSA, whether administered statically or dynamically, can produce very strong interrater reliability. These findings provide evidence that counselor educators and clinical supervisors may have more confidence in the use of the UCSA as another data point when making student evaluations of counseling skills.

Finally, the results of testing Hypothesis Five demonstrated that master's level supervisors and counselor educators assessed counseling skills similarly. While this finding tends to support the standards that identify qualified supervisors, a contradiction seems to exist with program evaluation data (Meacham & Stoltz, 2012) that shows site supervisors score counseling students' work significantly higher than do counselor educators. The UCSA forms submitted by site supervisors often report students as performing at the highest possible level (all 5's across the instrument's 12 items) while the students' professors score the same work considerably lower. The results of testing Hypothesis Five might be presented to site supervisors during their training to: (a) help inform them regarding the instrument's use, (b) acknowledge the expertise of

certified supervisors, and (c) impress upon them the importance of developmentally appropriate assessment and evaluation to students' development.

Conclusion

This study was designed to explore the effect of instrument-specific rater training on interrater reliability. Trained raters, restricted to a structured scoring system (DSI), produced significantly more reliable data than did untrained raters. How much of the demonstrated improvement was attributable to the training and how much was associated with the DSI is a matter for future research.

Additionally, this study was designed to examine the effect of instrument-specific rater training on raters' ability to differentiate between low and high counseling skills performances. The data demonstrated no significant differences between the untrained and trained raters. As studies testing much more extensive training regimens (Fitch, Gillam, & Baltimore, 2004; McCullough et al., 2010; Wang, 2010) have found evidence connecting training with improved rater scoring, this remains an area of active inquiry. This study's results may help shape future inquires into the link between rater training and rater scoring.

The findings of this study inform not only future research but are applicable to current counselor education. This study offers educators new effective, efficient training mechanisms that make productive use of current technology. The results also support the use and continued

development of an instrument (UCSA) specifically designed to structure the decision making process that counselor educators employ to move a student into more advanced clinical work.

REFERENCES

REFERENCES

Allen, D. (Ed.). (1967). *Microteaching: A description*. Palo Alto, CA: Stanford University Teachers Education Program.

Altman, D. (2000, January). *A review of experimentwise type I error: Implications for univariate post hoc and for multivariate testing*. Paper presented at the annual meeting of the Southwest Education Research Association. Retrieved from http://ericae.net/ft/tamu/altpaper.pdf

American Counseling Association. (2005). *Code of ethics and standards of practice*. Alexandria, VA: Author.

Andersen, T. (1991). *The reflecting team: Dialogues and dialogues about dialogues*. New York: Norton.

Baker, S. B., Daniels, T. G., & Greeley, A. T. (1990). Systematic training of graduate-level counselors: Narrative and meta-analytic reviews of three major programs. *The Counseling Psychologist, 18*, 355-421.

Bandura, A. (1991). Social cognitive theory of self-regulation. *Organizational Behavior and Human Decision Processes, 50*, 248-287.

Bandura, A. (1997). Self-efficacy: The exercise of control. New York: Freeman.

Barrett-Lennard, G. T. (1962). Dimensions of therapist response as causal factors in therapeutic change. *Psychological Monographs, 76*, 1-36.

Bernard, J. M., & Goodyear, R. K. (2009). *Fundamentals of clinical supervision*. Upper Saddle River, NJ: Pearson Education.

Boyd, J. (1978). *Counselor supervision: Approaches, preparation, practices*. Muncie, IN: Accelerated Development.

Byrne, A., & Hartley, M. (2010). Digital technology in the 21st century: Considerations for clinical supervision in rehabilitation education. *Rehabilitation Education, 24*, 57-67.

Carkhuff, R. R. (1969a). *Helping and human relations: A primer for lay and professional helpers (Vol. 1)*. New York: Holt, Rinehart, & Winston.

Carkhuff, R. R. (1969b). *Helping and human relations: A primer for lay and professional helpers (Vol. 2)*. New York: Holt, Rinehart, & Winston.

Carkhuff, R. R. (1969c). Helper communication as a function of helpee affect and content. *Journal of Counseling Psychology, 16*, 126-131.

Carkhuff, R. R. (1972). Credo of a militant humanist. *Personnel and Guidance Journal, 51*, 237-242.

Carkhuff, R. R. (2000). *The art of helping in the 21st century* (9th Ed.). Amherst, MA: Human Resource Development Press.

Carkhuff, R. R. (2009). *Trainer's guide for: The art of helping* (9th Ed.). Amherst, MA: Human Resource Development Press.

Carkhuff, R. R., & Burstein, J. W. (1970). Objective, therapist and client ratings of therapist-offered facilitative conditions of moderate- to low-functions therapists. *Journal of Clinical Psychology, 26*, 394-395.

Carkhuff, R. R., & Cannon, J. C. (1969). The effect of rater level of functioning and experience upon the discrimination of facilitative conditions. *Journal of Consulting Psychology, 33*, 189-194.

Carkhuff, R. R., Collingwood, T., & Renz, L. (1969). The effects of didactic training upon trainee level of discrimination and communication. *Journal of Clinical Psychology, 25*, 460-461.

Carkhuff, R. R., Friel, T., & Kratochvil, D. (1970). The differential effects of sequence of training in counselor-responsive and counselor-initiated conditions. *Counselor Education and Supervision, 9*, 106-108.

Carroll, M. (1996). *Counseling supervision: Theory, skills, and practice*. London: Cassell.

Collingwood, T., Renz, L., & Carkhuff, R. R. (1969). The effects of client confrontation upon levels of immediacy offered by high and low functioning counselors. *Journal of Clinical Psychology, 25*, 224-225.

Conver, B. J. (1944). Studies in phonographic recordings of verbal material: The completeness and accuracy of counseling interview reports. *Journal of General Psychology, 30*, 181-203.

Council for Accreditation of Counseling and related Educational Programs. (2009). *2009 Standards*. Retrieved from http://www.cacrep.org/so009standards.html

Crews, J., Smith, M. R., Smaby, M. H., Maddux C. D., Torres-Rivera, E., Casey, J. A. (2005). Self-monitoring and counseling skills: Skills-based versus interpersonal process recall training. *Journal of Counseling & Development, 83*, 78-85.

Daniels, T. (2003). Overview of research on microcounseling: 1967-present. In A. E. Ivey & M. B. Ivey (Eds.), *Intentional interviewing and counseling: Your interactive resource.* Pacific Grove, CA: Brooks/Cole.

Daniels, J. A., & Larson, L. M. (1992). The impact of performance feedback on counseling self-efficacy and counselor anxiety. *Counselor Education and Supervision 41*, 120-130.

Dawes, R. (1994). *House of cards: Psychology and psychotherapy built on myth.* New York: The Free Press.

Eriksen, K., & McAuliffe, G. (2003). A measure of counselor competency. *Counselor Education & Supervision, 43,* 120-133.

Fitch, T., Gillam, L., & Baltimore, M. (2004). Consistency of clinical skills assessment among supervisors. *The Clinical Supervisor, 23,* 71-81.

Falender, C. A., Cornish, J. A. E., Goodyear, R. K., Hatcher, R., Kaslo, N. J., Leventhal, G., Shafranske, E., Sigmon, S., Stoltenberg, C., & Grus, C. (2004). Defining competencies in psychology supervision: A consensus statement. *Journal of Clinical Psychology, 60,* 771-785.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41,* 1149-1160.

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.

Fisher, R.A. (1915). Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population. *Biometrika, 10,* 507–521.

Fitch, T., Gillam, L, & Baltimore, M. (2004). Consistency of clinical skills assessment among supervisors. *The Clinical Supervisor, 23,* 71-81.

Forsyth, D. R., & Ivey, A. E. (1980). Microtraining: An approach to differential supervision. In A. K. Hess (Ed.), *Psychotherapy supervision: Theory, research and practice,* 242-261. New York: Wiley.

Freeman, B., & McHenry, S. (1996). Clinical supervision of counselors-in-training: A nationwide survey of ideal delivery, goals, and theoretical influences. *Counselor Education & Supervision, 36,* 144-58.

Gravetter, F. J., & Wallnau, L. B. (2004). *Statistics for the behavioral sciences* (6th ed.). Belmont, CA: Thomas Learning.

Grencavage, L. M., & Norcross, J. C. (1990). Where are the commonalities among the therapeutic common factors? *Professional Psychology: Research and Practice, 21*, 372-378.

Green, S. B., & Salkind, N. J. (2005). *Using SPSS for Windows and MacIntosh: Analyzing and understanding data.* Upper Saddle River, NJ: Pearson Education.

Hawley, L. D. (2006). Reflecting teams and microcounseling in beginning counselor training: Practice in collaboration. *Journal of Humanistic Counseling, Education and Development, 45*,198-207.

Hiebert, B., Uhlemann, M. R., Marshall, A., & Lee, D. L. (1998). The relationship between self-talk, anxiety, and counselling skill. *Canadian Journal of Counselling, 32*, 163-171.

Hinkle, D. E., Wisersma, W., & Jurs, S. G. (2003). *Applied statistics for the behavioral sciences* (6th ed.). New York: Houghton Mifflin.

Hess, A. K. (1980). Training models and the nature of psychotherapy supervision. In A. K. Hess (Ed.), *Psychotherapy supervision: Theory, research, and practice* (pp. 15-28). New York: John Wiley & Sons.

Holloway, E. L. (1995). *Clinical supervision: A systems approach*. Thousand Oaks, CA: Sage Publications.

Horvath, A. O. (1981). *An exploratory study of the working alliance: Its measurement and relationship to outcome. Unpublished doctoral dissertation,* University of British Columbia, Vancouver, British Columbia, Canada.

Ivey, A. E. (1973). Microcounseling: The counselor as trainer. *The Personal and Guidance Journal, 51,* 311-316.

Ivey, A. E., & Authier, J. (1971). *Microcounseling: Innovations in interviewing, counseling psychotherapy, and psychoeducation*. Springfield, IL: Charles C. Thomas.

Jennings, L., Goh, M., Skovholt, T. M., & Banerje-Stevens, D. (2003). Multiple factors in the development of the expert counselor and therapist. *Journal of Career Development, 30,* 59-72.

Kagan, N., Krothwohl, D. R., and Miller, R. (1963). Stimulated recall in therapy using video tape: A case study. *Journal of Counseling Psychology, 10,* 237-243.

Kenny, D. A. (1987). *Statistics for the social and behavioral sciences* [Adobe PDF]. Retrieved from http://davidakenny.net/doc/statbook/kenny87.pdf

Kiresuk, T. J., Smith, A., & Cardillo, J. E. (1994). *Goal attainment scaling: Applications, theory, and measurement*. Hillsdale, NJ: Lawrence Erlbaum.

Lambert, M. J. (1989). The individual therapist's contribution to psychotherapy process and outcome. *Clinical Psychology Review, 9,* 469-485.

Lepkowski, W. J., Packman, J., Smaby, M. H., & Maddux, C. (2009). Comparing self and expert assessments of counseling skills before and after skills training, and upon graduation. *Education, 129,* 363-371.

Lictenberg, J.W. (1997). Expertise in counseling psychology: A concept in search of support. *Educational Psychology Review, 9,* 221–238.

Little, C, Packman, J., Smaby, M. H., & Maddux, C. D. (2005). The skilled counselor training model: Skills acquisition, self-assessment, and cognitive complexity. *Counselor Education & Supervision, 44*, 189-200.

Lowry, R. (2013). Vassarstats: Website for statistical computation. Retrieved from http://www.vassarstats.net/rdiff.html

Love, G. (1988, November). *Understanding experimentwise error probability*. Paper presented at the annual meeting of the Mid-South Educational Research Association, Louisville. (ERIC Document Reproduction Service No. ED 304 451)

Matarazzo, R. G., Phillips, J. S., Wiens, A. N., & Saslow, G. (1965). Learning the art of interviewing: A study of what beginning students do and their pattern of change. *Psychotherapy: Theory, research and Practice, 2*, 49-60.

Matarazzo, R. G., Wiens, A. N., & Saslow, G. (1966). Experimentation in the teaching and learning of psychotherapy skills. In L. A. Gottschalk & A. Auerbach (Eds.), *Methods of research in psychotherapy*, 597-635. New York: Appleton-Century-Crofts.

McCullough, L., Kuhn, N., Andrews, S., Valen, J., Hatch, D., & Osimo, F. (2003). The reliability of the Achievement of Therapeutic Objectives Scale (ATOS): A research and teaching tool for psychotherapy. *Journal of Brief Therapy, 2*, 75-90.

McNeil, B. W., Stoltenberg, C. D., & Romans, J. S. (1992). The Integrated Developmental Model of supervision: Scale development and validation procedures. *Professional Psychology: Research & Practice, 20*, 329-333.

Meacham, P. Jr., & Stoltz, K. B. (2012, September). *Improving the robustness of counseling skills assessment: A case for intraclass correlation case 3 analysis*. Paper presented at the Association for Assessment in Counseling and Education, Orlando, FL.

Miller, W. R., & Mount, K. A. (2001). A small study of training in motivational interviewing: Does one workshop change clinician and client behavior? *Behavioural and Cognitive Psychotherapy, 29,* 457–471.

Miller, W. R., Yahne, C. E., Moyers, T. B., Martinez, J., & Pirritano, M. (2004). A randomized trial of methods to help clinicians learn motivational interviewing. *Journal of Consulting and Clinical Psychology, 72,* 1050–1062.

Paladino, D. A., Barrio-Minton, C. A., & Kern, C. W. (2011). Interactive training model: Enhancing beginning counseling student development. *Counselor Education & Supervision, 50,* 189-206.

Pierce, R. (1968). Graduate training of facilitative counselors: The effects of individual supervision. *Journal of Counseling Psychology, 17,* 210-215.

Ramseyer, G. C. (1979). Testing the difference between dependent correlations using the Fisher Z. *Journal of Experimental Education, 47,* 307-310.

Reivich, R., & Geertsma, R. (1969). Observational media and psychotherapy training. *Journal of Nervous and Mental Disorders, 148,* 310-327.

Rogers, C. R. (1957). The necessary and sufficient conditions of therapeutic personality change. *Journal of Consulting Psychology, 21,* 95-103.

Ronnestad, M. H., & Skovholt, T. M. (1993). Supervision of beginning and advanced graduate students of counseling and psychotherapy. *Journal of Counseling and Development, 71,* 396-405.

Rubel, E. C., Sobell, L. C., & Miller, W. R. (2000). Do continuing education workshops improve participants' skills? Effects of a motivational interviewing workshop on substance-abuse counselors' skills and knowledge. *The Behavior Therapist, 23,* 73-77, 90.

Russell-Chapin, L. E. (2000). The Counselling Interview Rating Form: A teaching and evaluation tool for counsellor education. *British Journal of Guidance & Counselling*, *28,* 115-124.

Saitz, R., Sullivan, L.M., and Samet, J. (2000). Training community-based clinicians in screening and brief intervention for substance abuse problems. *Substance Abuse*, *2*, 21-31.

Saunders, S. M., Howard, K. I., & Orlinsky, D. E. (1989). The Therapeutic Bond Scales: Psychometric characteristics and relationship to treatment effectiveness. *Psychological Assessment, 1, 323-330.*

Schanche, E., Nielsen, G. H., McCullough, L., Valen, J., & Mykletun, A. (2010). Training graduate students as raters in psychotherapy process research: Reliability of ratings with the Achievement of Therapeutic Objectives Scale (ATOS). *Nordic Psychology, 62*, 4-20.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlation: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420-428.

Smaby, M. H., & Maddux, C. D. (2011). *Basic and advanced counseling skills: The skilled counselor training model*. Belmont, CA: Brooks/Cole.

Smaby, M. H., Maddux, C. D., Schaefle, S., & Packman, J. (2006). *Counseling skills training, client perceptions, and goal attainment*. Unpublished manuscript, University of Nevada at Reno, NV.

Smaby, M. H., Maddux, C. D., Torres-Rivera, E., & Zimmick, R. (1999). A study of the effects of a skills-based versus a conventional group counseling training program. *The Journal for Specialists in Group Work, 24,* 152-163.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32,* 752–760.

Steward, R. J., Breland, A., &. Neil, D. M. (2001) Novice supervisees' self-evaluations and their perceptions of supervisor style. *Counselor Education & Supervision*, *41*, 131-141.

Stein, D. M., & Lambert, M. J. (1995). Graduate training in psychotherapy: Are therapy outcomes enhanced? *Journal of Counseling Psychology, 63,* 182-196.

Stevens, H. S., Dinoff, B. L., & Donnenworth, E. E. (1998). Psychotherapy training and theoretical orientation in clinical psychology programs: A national survey. *Journal of Clinical Psychology, 54*, 91-96.

Tabachnick, B. G., & Fidell, L. S. (2006). *Using multipartite statistics* (5th ed.). London: Allyn and Bacon.

Torres-Rivera, E., Wilbur, M. P., Maddux, C. D., Smaby, M. H., Phan, L. T., & Roberts-Wilbur, J. (2002). Factor structure and construct validity of the counselor skills personal development rating form. *Counselor Education & Supervision, 41*, 268-278.

Tracey, T. J., Clidden, C. E., & Kokotovic, A. M. (1988). Factor structure of the counselor rating form – short. *Journal of Counseling Psychology, 35*, 330-335.

Truax, C. B., Carkhuff, R. R., & Douds, J. (1964). Toward an integration of the didactic and experiential approach to training in counseling and psychotherapy. *Journal of Counseling Psychology, 11*, 240-247.

Urbani, S., Smith, M. R., Maddux, C. D., Smaby, M. H., Torres-Rivera, E., & Crews. J. (2002). Skills based training and counselor self-efficacy. *Counselor Education & Supervision,* 42, 92-106.

Walters, S. T., Matson, S. A., Baer, J. S., & Ziedonis, D. M. (2005). Effectiveness of workshop training for psychosocial addiction treatments: A systematic review. *Journal of Substance Abuse Treatment, 29*, 283-293.

Wang, B. (2010). On rater agreement and rater training. *English Language Teaching, 3*, 109-112.

Wang, H. (2008). A study on raters' interpretation and application of the rating criteria in TEM4-Oral. *Theory and Practice of Foreign Languages Teaching, 2*, 33-39.

Weinrach, S. G. (1987). Microcounseling and Beyond: A Dialogue with Allen Ivey. *Journal of Counseling & Development*, 65, 532.

Wen, Q., Liu, X., & Jin, L. (2005). Native and nonnative judgments of Chinese learners' English public speaking ability. *Foreign Language Teaching and Research, 37*, 337-342.

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

Wilbur, M. P., Roberts-Wilbur, J., Morris, J. R., Betz, R. L., & Hart, G. M. (1991). Structured group supervision: Theory into practice. *The Journal for Specialists in Group Work, 16*, 91-100.

Williams, A. (1995). *Visual and active supervision: Roles, focus, technique*. New York: W. W. Norton.

Young, M. E. (2009). *The art of helping: Building blocks and techniques* (4th Ed.). Upper Saddle River, NJ: Pearson Education.

Zimmick, R., Smaby, M. H., & Maddux, C. D. (2000). Improving the use of a group counseling scale and related model to teach theory and skills integration. *Counselor Education & Supervision, 39*, 284-295.

APPENDIX

APPENDIX A

Consent to Participate in an Experimental Study

Title: The Effect of Instrument Specific Training on Interrater Reliability and Counseling Skills Performance Differentiation

Investigators

Paul Meacham, Jr., M.Ed., NCC
Doctoral Candidate
Department of Leadership and Counselor Education
University of Mississippi
10213 Hwy 51, Courtland MS, 38620
(662) 561-5528

Kevin B. Stoltz, Ph.D.
Department of Leadership and Counselor Education
109 Guyton Hall
The University of Mississippi
(662) 915-5376

Description

We are interested in helping counselor educators and clinical supervisors improve the quality of their assessment of students' and supervisees' counseling skills. In this study, we have two primary interests. We want to know: (a) to what extent dynamic scoring affects interrater reliability (IRR) when scoring demonstrated counseling skills via video recordings and (b) to what extent dynamic scoring affects raters' ability to identify those counselors whose demonstrated skills fall on the lower end of the performance spectrum.

All participants will be asked to connect to a website and complete a demographics form providing us with basic non-identifying personal and professional information. Participants will also be asked to view a video of a counselor (in session with a standardized client) and score the counselor's performance according to a provided 12-item instrument (the Universal Counseling Skills Assessment, UCSA). Participants will also have the opportunity to earn ½ of a Supervision Continuing Education Unit (CEU) by viewing a training video regarding the assessment of counseling skills and completing a short test on the material covered in the training video. Participants may ask questions about the study before or after participation and may withdraw from the study at any time. Approximately one hour will be required to complete the study including the continuing education training and testing.

Risks and Benefits

A short test is required for those who desire the CEU and testing triggers a measure of anxiety for some people. However, testing for the CEU is not necessary to participate in the study, and we encourage you to talk with us outside of the study about any uncomfortable anxiety produced by the testing experience. We do not think there are other risks involved in completing study. In addition, you may receive the personal satisfaction of knowing you have contributed to an honest

effort to advance our profession in the service of our fellow man/woman and one Supervision CEU at no financial cost.

Cost and Payments
Participants' successful completion of the training video and test will earn them one CEU at no financial cost. There are no costs to the participants or payments made for participating in this study.

Confidentiality
Absolute confidentiality of data provided through the Internet cannot be guaranteed due to the limited protections of Internet access. Please be sure to close your browser when finished so no one will be able review the answers you have provided to the study. No names or other identifying information will be collected or attached to scoring forms. If you desire the CEU, you will be asked to provide identifying information on a separate form which will be associated with the test answers you provide but will not be associated with the demographic information given earlier or with the scoring data collected in the study. If you chose not to pursue the CEU, no identifying information will be collected at any time.

Right to Withdraw
You have the right to refuse to participate in the study or to withdraw from the study at any time. You may contact either of the researchers if you have a question about not completing the study after you have started. If you choose to withdraw after you have completed the scoring of the counseling video, removal of your data from the study will not be possible as no identifying information will be associated with the data you generate. If you choose to withdraw, you will be provided with a link where you can still view the training video and complete the test to earn a ½ supervision CEU without financial cost.

The researchers may terminate your participation in the study without your consent and for any reason, such as protecting your safety or protecting the integrity of the research data. If the researchers terminate your participation, you will still be provided a link where you can view the training video and complete the test to earn the ½ Supervision CEU without financial cost.

IRB Approval
The University of Mississippi's Institutional Review Board (IRB) has reviewed this study and determined that it fulfills the human research subject protection obligations required by state and federal law and University policies. If you have any questions, concerns, or reports regarding your rights as a participant of research, please contact the IRB at (662) 915-7482.

Statement of Consent
I have read the above information. I have been given a copy of this form. I have had an opportunity to ask questions, and I have received answers. I consent to participate in the study.

Signature of Participant Date

Signature of Investigator Date

NOTE TO PARTICIPANTS: DO NOT SIGN THIS FORM
IF THE IRB APPROVAL STAMP ON THE FIRST PAGE HAS EXPIRED.

APPENDIX B

Demographics Form

1. Age: _____

2. Gender ☐ Male ☐ Female

3. Race

☐ African-American
☐ Asian
☐ Caucasian
☐ Latina/Latino
☐ Middle-Eastern
☐ Native American/First People
☐ Pacific Islander
☐ Other _____

4. Professional Standing

☐ Master's in counseling (state certified supervisor)

☐ Master's in counseling (ACS)

☐ I have completed a doctoral level clinical supervision course

☐ Ph.D. or Ed.D. in Counselor Education or closely related field

5. Years of Experience as a Counselor Educator and/or Clinical Supervisor: _____

6. By what model were you taught counseling skills?

☐ Carkhuff's Human Resource Development Model
☐ Egan's Skilled Helper Model
☐ Roger's Client-centered Model
☐ Ivey's Microcounseling Model
☐ Smaby & Maddux's Skilled Counselor Training Model
☐ Young's Megaskills Model
☐ Other _____

7. If you were to teach a course in basic counseling skills next semester, what model of skills instruction would you use?

- ☐ Carkhuff's Human Resource Development Model
- ☐ Egan's Skilled Helper Model
- ☐ Roger's Client-centered Model
- ☐ Ivey's Microcounseling Model
- ☐ Smaby & Maddux's Skilled Counselor Training Model
- ☐ Young's Megaskills Model
- ☐ Other _____

APPENDIX C

Universal Counseling Skills Assessment

Please rate student performance on each skill listed using the following ratings:	**N/O** = Not observed but not required by the situation **1** = Not observed when required by the situation **2** = Demonstrated insufficiently to proceed to practicum/internship **3** = Demonstrated sufficiently to proceed to practicum/internship but inconsistently **4** = Demonstrated sufficiently and consistently enough to proceed to practicum/internship **5** = Demonstrated at a level normally indicative of one who has at least completed internship

	Specific Skill	Rating					
Attending	1. **Body Language & Appearance** *Maintains open, relaxed, confident posture. Leans forward when talking. Maintains professional dress.*	NO	1	2	3	4	5
	2. **Eye Contact** *Maintains appropriate eye contact.*	NO	1	2	3	4	5
	3. **Minimal Encouragers** *Uses prompts (uh huh, okay, right) to let the client know s/he is heard. Uses silence helpfully. Uses nods and body gestures to encourage client to talk.*	NO	1	2	3	4	5
	4. **Vocal Tone** *Matches the sense of the session and session goals. Vocal tone communicates caring and connection with the client.*	NO	1	2	3	4	5
	5. **Verbal Tracking** *Staying on topic that client presents. Repeats key words or phrases.*	NO	1	2	3	4	5

Attending Tot. _____ # of Qualified Ratings _____ Attending Avg. _____

	Specific Skill	Rating					
Basic Listening	6. **Selective Attending** *Selectively attend to key aspects of client communication.*	NO	1	2	3	4	5
	7. **Directions and Encouraging Client to Talk** *Uses "tell me more…about" or similar statements that encourage the client to talk about the specific aspects of the client's communication.*	NO	1	2	3	4	5
	8. **Paraphrasing (Reflections of Content)** *Uses brief, accurate, & clear rephrasing of what the client has expressed.*	NO	1	2	3	4	5
	9. **Reflecting Feeling/Meaning** *States succinctly the feeling experienced by the client (You feel ___ or You feel ___ when ___) Feeling/Meaning statements are personalized*	NO	1	2	3	4	5
	10. **Summarizing** *Makes statements at key moments in the session that capture the overall sense of what the client has been expressing.*	NO	1	2	3	4	5

Basic Listening Tot. _____ # of Qualified Ratings _____ Basic Listening Avg. _____

Please rate student performance on each skill listed using the following ratings:	**N/O** = Not observed but not required by the situation **1** = Not observed when required by the situation **2** = Demonstrated insufficiently to proceed to practicum/internship **3** = Demonstrated sufficiently to proceed to practicum/internship but inconsistently **4** = Demonstrated sufficiently and consistently enough to proceed to practicum/internship **5** = Demonstrated at a level normally indicative of one who has at least completed internship

	Specific Skill	Rating					
Deepening	11. **Using Immediacy** *Recognizes here-and-now feelings, expressed verbally or nonverbally, of the client or the counselor. Can be related to the counselor-client relationship ("As we talk about ___, I sense you are feeling ___. I'm feeling ___ about how you are viewing the problem right now.")*	NO	1	2	3	4	5
	12. **Challenging/Pointing Out Discrepancies** *Expresses observations of discrepancies ("You expect yourself to do ___ when facing the problem of ___ but you do ___ instead. When this happens, you feel about yourself.")*	NO	1	2	3	4	5

Deepening Tot. # of Qualified Ratings Deepening Avg.

Total Score _____ Total # of Qualified Ratings _____ Total Avg. _____

Comments

APPENDIX D

Skills Quiz [With Pop-up Responses]

1. A high quality *reflection of content* (paraphrase):
    A. Repeats the contents of the client's expressions using the client's exact words as much as possible. [This answer is not correct because a reflection of content is most effective when it expresses the content of the client's expressions in a "fresh way" using different words. Please click "Ok" and attempt the question again.]
    B. Feeds back to the client the content of what he or she has just said but in a restated form.
    C. Challenges clients by causing them to confront the incongruences in their narrative. [This answer is not correct. It is a good working definition of the skill of challenging the client but does not describe a reflection of content. Please click "Ok" and attempt the question again.]
    D. None of the above. [This answer is not correct. Please click "Ok" and attempt the question again.]
2. Counselors know they have made an accurate *reflection of feeling* when:
    A. Their response is interchangeable with the feeling expressed by client.
    B. Their response causes a cathartic outpouring of tears. [This answer is not correct. A reflection of feeling may evoke an emotionally charged response by the client, but it is also true that it may not. The accuracy of the reflection is not proportional to nor measured by the level of client's emotional response. Please click "Ok" and attempt the question again.]
    C. Their response causes the client to think about their situation in a different way. [This answer is not the most correct response. A reflection of feeling focuses on the

affective and does not typically produce cognitive reframing. Please click “Ok” and attempt the question again.]

D. Their response amazes the client with an almost mystic like ability to read the inner person. [This answer is not correct. A high quality reflection of feeling may cause the client to feel heard and understood, but it does not cause rational people to ascribe mystical powers to the counselor. Please click “Ok” and attempt the question again.]

3. A *reflection of meaning* focuses on:

A. The hidden truth the client is afraid to face. [This answer is not correct. Sometimes the meaning of a set of circumstances may be hidden to a client, but that is not always the case. Furthermore, when the meaning of a situation is hidden from a client it is not necessarily due to fear. Please click “Ok” and attempt the question again.]

B. The personal import of the content of the client’s life and their emotional reaction to that content.

C. The reframed perspective clients should adopt to make the most out of their situations. [This answer is not correct. A reflection of mean focuses more on what is, and less on might become. Please click “Ok” and attempt the question again.]

D. The true gestalt composed of all the individual events of the client’s life. [This answer is not correct. A reflection of meaning is limited to the content and emotions expressed by a client in-session. The client brought these elements into the session for a reason. The focus of a reflection of meaning is to facilitate the exploration of

the importance of these elements to the client. Please click “Ok” and attempt the question again.]

4. A confrontation is at its most effective when the helper:

   A. Confronts the client with his or her mistakes in a candid way. [This answer is not correct. It is not for the counselor to sit in judgment over what parts of a client’s life may or may not be mistakes. Please click “Ok” and attempt the question again.]

   B. Confronts the client with gentleness and compassion. [This answer is not correct. There are times that gentleness may be helpful and other times when it may serve to obscure the presence of incongruences in the client’s circumstances. Please click “Ok” and attempt the question again.]

   C. Confronts the client using direct, behavioral observations of the lack of congruency in his or her responses to a particular situation.

   D. Confronts the client regarding his or her maladaptive responses to life’s encounters. [This answer is not correct. The purpose of a confrontation is not to set the counselor in opposition to the client. Rather, the purpose of a confrontation is to help the client see that certain elements of his or her life are at odds with each other. Please click “Ok” and attempt the question again.]

5. Which of the following skills incorporates an element of counselor self-disclosure?

   A. Reflection of content [This answer is not correct. A reflection of content returns to the client the expressed elements of the client’s life. Please click “Ok” and attempt the question again.]

   B. Summarizing [This answer is not correct. A summary encapsulates the contents and feelings of the clients along with the meaning he or she attaches to those elements. It

is focused on the session not on the counselor. Please click “Ok” and attempt the question again.]

C. Confrontation [This answer is not correct. A confrontation focuses on various elements of the client’s life and the how those elements are incongruent with each other. Please click “Ok” and attempt the question again.]

D. Immediacy

VITA

Paul Meacham, Jr. was born in San Diego, CA, on February 27, 1963 to Paul and Nydia Meacham. He attended the Morrision, TN Elementary School from the first through the eighth grade. At the beginning of the ninth grade, Paul moved to the Warren County High School from which he graduated with honors in 1981. In the fall of 1981, Paul began his studies at the University of Tennessee – Knoxville. Paul majored in Mechanical Engineering but left school at the end of the first year due to a lack of interest and a lack of performance.

Paul resumed his college career in 2002 at Southern Christian University, completing a BA in Biblical Studies in 2005 and graduating with highest honors. In 2007, Paul entered the graduate counseling program at the University of Mississippi graduating with a M.Ed. in 2009. In 2009, Paul began the Doctoral program for Counselor Education and Supervision at the University of Mississippi. His research interests include counseling skills instruction and assessment, research methodology, and statistical analyses. During his final year in the PhD program, Paul accepted a position as an Assistant Professor at Amridge University, which he will begin upon graduation.