AN ABSTRACT OF THE DISSERTATION OF

Gideon R. Litherland for the degree of Doctor of Philosophy in Counseling presented on July 29, 2020.

Title: Cross-Validation of Two Supervision Instruments with Counseling Trainees from CACREP-accredited Programs

Abstract approved: _____

Kok-Mun Ng

Supervision is considered a pivotal professional intervention for counselors-in-training as they develop during graduate education and beyond. But, research on clinical supervision suffers from a lack of common instruments that can be utilized across disciplines, including counseling, and national boundaries. While a complex phenomenon to empirically address through research, supervision scholars have called for greater scrutiny in the types of instruments, measurement models, and research designs employed so that the field of supervision may be advanced (Goodyear et al., 2016). Additionally, how supervision-related phenomena are conceptualized, like supervision effectiveness and supervisor competence, within the research is foundational to the development of a robust research ecosystem. In order to contribute to the need for psychometrically robust supervision instruments that are grounded in robust theory, this dissertation project included two supervision measurement cross-validation studies on counselors-in-training in the United States.

In these studies, I examined the psychometric properties of two different supervision instruments that share a similar theoretical conceptualization of supervision: (a) effective supervision (Study 1) and (b) supervisor competence (Study 2). These two measures were

developed in non-U.S. clinical settings but whose psychometric properties and utility have yet to be verified for U.S.-based counseling practitioners. The overarching research question that both studies sought to address was: “Do existing supervision evaluation instruments maintain rigorous psychometric evidence for a sample of CITs from CACREP-accredited programs?” Conceptually, supervision effectiveness and supervisor competence were defined terms of the Proctor Model of Supervision. Each study drew from one sampling of 86 participants who were master’s-level counselors-in-training at CACREP-accredited programs from every region in the United States.

The first study considered the psychometric properties of an instrument, the Manchester Clinical Supervision Scale (MCSS-26), that has been long-utilized to measure clinical supervision effectiveness outside of the U.S. In addition to the overarching research question identified above, Study 1 evaluated item-level performance and instrument-level internal consistency, concurrent validity, and social desirability threats to validity. The MCSS-26 was subjected to item-level analysis using a Generalized Partial Credit Model (GPCM) to explore the item difficulty, discrimination, and satisfaction to theoretical assumptions. Results of the study indicate acceptable instrument-level validity and reliability but poor item-level fitness for multiple items from the original 26-item instrument. Based on sample data, results suggest the revision of the MCSS-26 to a 9-item instrument that more appropriately fits within the item-response theoretical model of analysis for the study sample. Fitness indices for the revised scale suggest a better model fit compared to the fitness indices of the original instrument. Further revision, through continued research, is necessary in order to critically revise the MCSS-26 for use with a US-based counselor-in-training population.

The second study examined an instrument that assesses supervisor competence, the Supervision Evaluation and Supervisor Competence (SE-SC) scale, from the supervisee's perspective. Study 2, similar to Study 1, evaluated the item-level and instrument-level psychometrics of the subscales of the SE-SC. Item performance, internal consistency, concurrent validity, and social desirability threats to validity were all considered. The SE-SC was subjected to item-level analysis based on a GPCM that resulted in difficulty and discrimination parameters while also considering key theoretical assumptions of the model. Data from the current sample indicate acceptable instrument validity and reliability; however, item-level fitness to the model was poor, or "misfitting," for a number of items. Results of Study 2 indicate the need for ongoing refinement of the SE-SC before use with a U.S.-based CIT population. Results further indicate that a 15-item revised SE-SC could be further developed with scrutiny. The revised scale possessed improved fitness indices compared to the original instrument, indicating a better fit to the GPCM.

Supervision instruments that are relevant for U.S.-based CIT are sorely needed and considered critical to the development of the supervision scholarship in the years to come. As two supervision instruments that have been used to assess effectiveness and supervisor competence, the findings from both studies cast doubt on their utility for the population of U.S.-based CIT. Implications for Study 1 and Study 2 are presented with respect to instrument revision/development, counselor education and training, and the common measurement approach in supervision research. Additionally, findings from both studies suggest the urgency of constructing, refining, and developing psychometrically robust supervision instruments that can precisely assess supervision effectiveness and supervisor competence in future research. Overall,

each study contributes to the supervision scholarship by casting doubt on two extant supervision instruments for use with a U.S.-based CIT population.

# Cross-Validation of Two Supervision Instruments with Counseling Trainees from CACREP-accredited Programs

by

Gideon R. Litherland

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of

the requirements for the

degree of

Doctor of Philosophy

Presented July 29, 2020

Commencement June 2021

Doctor of Philosophy dissertation of Gideon R. Litherland presented on July 29, 2020

APPROVED:

---

Major Professor, representing Counseling

---

Dean of the College of Education

---

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

---

Gideon R. Litherland, Author

ACKNOWLEDGEMENTS

I am ever grateful to the many people who have helped, challenged, and supported me along my professional path. This doctoral degree would not have been possible without their energy and investment in my success and growth as a person and professional. These brief acknowledgements do not do the gratitude I have for them justice!

I would like to thank my committee members, Dr. Arien Muzacz, Dr. Abraham Cazares-Cervantes, and Dr. Thomas Miller. I appreciate your constant support, guidance, and time. Each of you have contributed to the completion of my doctoral studies and have supported me to get to this point.

To my mentors and friends, thank you. The grace, compassion, and guidance you have provided me over the years has shaped me in ways words cannot express. I am a better human for knowing each of you and I fully intend to pay it forward. In gratitude Jane Bello-Brunson, Darlene Grega, Dr. Fran Giordano, and Carolyn Schneider.

To my co-chair and mentor, Dr. Thom Field, you jumped into this dissertation project with a clear intention to be facilitative from the get go. I am grateful to you for sharing your expertise and wisdom with such direct clarity and, at times, helping me see another perspective. Thank you for your service.

To my advisor, chair, and mentor, Dr. Kok-Mun Ng, you have taught me more than just how to be a professional. Your humor, generosity of self, and scholarly rigor are a rare combination that I strive to reflect in my ongoing personal development. Thank you for your constant support, feedback, and connection.

To my fellow doctoral candidates and co-conspirators in fostering excellence in counseling, Gretchen Schulthes, Christy Cosper, Rachel Ware Zooi, Nineka Dyson, and Roberta

Miranda, I am indebted to each of you for pushing me to be a more compassionate and critical thinking. You are my village and I do believe it took our paths to cross for me to arrive at this point. Love to you all.

To the Litherlands, Downings, Neimans, and Roberts, thank you for all of the love, patience, and understanding you have sent my way over the years. I would not have been able to complete this degree without you all. You keep me humble and grounded.

To Spencer Neiman, my fiancé and husband-to-be, this accomplishment is as much yours as it is mine. From the beginning, you have encouraged me to imagine what could be instead of just seeing how things are. I would not be where I am today without you in my life. Thank you, love you.

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

TABLE OF CONTENTS (continued)

Page

TABLE OF CONTENTS (continued)

Page

LIST OF TABLES

Table Page

LIST OF APPENDICES

# Chapter 1: Thematic Introduction

As per the requirement of the Ph.D. in Counseling Program at Oregon State University, following the structure of a contemporary, manuscript-style dissertation, the four chapters presented here represent a cross-validation study of two supervision instruments. Chapter 1 presents the context and professional issues that Chapters 2 and 3 attempt to address. In Chapters 2 and 3, the singular focus is a specific supervision instrument that will be subjected to scrutiny. Each chapter represents a stand-alone empirical study. Chapter 2 focuses on the need for psychometrically valid instruments that measure supervision effectiveness. Chapter 3 focuses on the need for psychometrically valid instruments that measure clinical supervisor competency. Connecting Chapters 2 and 3 is their shared operationalization of the Proctor Model of Supervision (Proctor, 2011). The research focus on supervision in counselors-in-training (CIT) in the United States further connects the two research manuscripts. Chapter 4 synthesizes the findings of Chapters 2 and 3 while articulating the main contributions of this dissertation. Key elements discussed further in the present chapter include the state of psychometric validation research in supervision scholarship, latent variables in supervision, the Proctor Model of Supervision, and the training context of sample participants.

## Instrument Development Issues

Instrument development is a critical scientific and ethical issue. Data, results, discussions, and hypothetical rejections are entirely disrupted if the instruments employed within a study possess no evidences of reliability or validity. As DeVellis (2017) notes, “No matter how well designed and executed other aspects of a research endeavor may be, measurement can make or break a study” (p. 246). Scientifically, the methodological design element, that is the instrument, measurement, scale, or assessment, maintains significant potential to detract or contribute to any

conclusions drawn therein. Heppner et al. (2016) note, “strong science is built on strong measures of psychological constructs” (p. 220) while simultaneously, “any psychological construct measured by a scale is culture bound” (p. 223). In examining theoretical constructs and their related latent variables, it is the ethical responsibility of the researcher to provide validation and “seek to understand the culturally based meaning of that scale” (Heppner et al., 2016). Put simply: the relevance and generalizability of any instrument is based on the scrutiny, attention, and chronological persistence of the research to support its use (DeVellis, 2017).

In addition to instrument-level issues, an item-level focus of analysis is critical to ongoing instrument development. Such an item-level focus of analysis directly attempts to refine the measurement model of the instrument, by considering each item, according to how precisely it accurately measures, or captures, the latent construct of concern. Andrich and Marais (2019) describe the importance of item performance, according to item response theory: “…there should be substantive and theoretical reasons for the inclusion of every item in an instrument…” (p. 338). Each item, in an ideal measurement model, satisfies performance expectations according to item-response theory (e.g., difficulty, discrimination). An instrument will only measure as precisely as its items perform, or assess, the latent construct.

Besides the technical aspects of instrument development and psychometrics, there is the ethical aspect to consider. That is, instrument selection is critical insofar as (a) the collected data is an accurate representation of the participant and (b) the instrument has been validated for use with the target population (DeVellis, 2017; Messick, 1995) with (c) items that precisely measure the latent ability/construct of concern (Andrich & Marais, 2019). All these issues are critical to every fields and disciplines that rely on quantitative assessment instruments, including supervision.

**Building an Empirical Body Through Cross-Validation**

The nature of empirical research is the ongoing accumulation of evidence to advance a shared understanding of a phenomenon in question. In this study, the measurement of supervision effectiveness and supervisor competence will be explored. Replication and cross-validation studies are critical to the relevance of supervision research findings because of the highly contextual and dynamic nature of human experience. While supervision scholarship benefits from a methodologically diverse body of research, the development of a corpus of empirical evidence to suggest the relevance and importance of supervision within clinician training, practice, and professional development remains an international and interdisciplinary goal (Ellis et al., 1996; Goodyear et al., 2016). Cross-validation research is critical to this task. Messick (1995) described the essence of cross-validation of instruments as, "The extent to which score meaning and action implications hold across persons or population groups and across settings or contexts is a persistent and perennial empirical question." (p. 741).

Due to the scant attention in supervision research to an instrument's original validation sample/reference group and its resultant psychometrics, researchers appear to rely heavily on an instrument's face validity regardless of the sample characteristics or measurement theory (Ellis et al., 2008). For instance, a methodological challenge presents itself when researching U.S-based master's-level counselors-in-training supervision experiences with an instrument that was first validated on a sample of UK-based doctoral-level psychology trainees. While the end-goal of producing a competent, effective mental health practitioner in the UK and the U.S. may be shared, the context of training and professional milieu are undoubtedly different. Cross-validation research, conducted utilizing established psychometric methods of analysis, presents an opportunity to contribute to such ongoing imperative work within supervision scholarship.

Instrument validation and evaluation relies on psychometrics. Jones and Thissen (2007) offer a clear definition: “Psychometrics, or quantitative psychology, is the disciplinary home of a set of statistical models and methods that have been developed primarily to summarize, describe, and draw inferences from empirical data collected in psychological research.” (p. 21). Such summaries, descriptions, and data analyses are critical to the ongoing validation of supervision instruments. Watkins and Milne (2014) capture the heart of the issue: “That absence of measures has certainly affected advance in studying particular supervision models as well. Any research is only as good as the measures upon which it is based.” (p. 677). In response to the need for psychometrically robust instruments to advance the field of clinical supervision, particularly in relation to cross-cultural supervision research, and to assist with program evaluation, my dissertation project was designed to scrutinize the psychometric properties of two supervision instruments and consider their applicability for use within the clinical training of counselors enrolled in U.S.-based, CACREP-accredited counselor education programs. These instruments focus on supervision effectiveness and supervisor competence.

**Studying Effectiveness and Competence in Supervision**

The study of latent constructs, or hypothesized ideas that are not directly observable, within clinical supervision is fraught with errors and a lack of operationalization (Goodyear et al., 2016). In order to address supervision effectiveness and supervisor competence on a profession-wide scale, instruments are required that can measure the degree of effectiveness of supervision and the level of competence of the supervisor. These two constructs, however, are not directly measurable.

The effectiveness of supervision may be only assessed indirectly. The theoretical question of how to define effectiveness is grist for the mill in the design of the researcher.

Effectiveness of supervision may be measured according to supervisees' wellbeing, or client outcomes, or supervisor rating, or adherence to a theoretical model (e.g., Ladany et al., 2013). Even these supervision effectiveness outcomes may only be assessed indirectly. For example, the Evaluation Process within Supervision Inventory (EPSI) (Lehrman-Waterman & Ladany, 2001) assesses the evaluative aspects of supervision, the Anticipatory Supervisee Anxiety Scale (ASAS) (Ellis, et al., 2014) assesses supervisee anxiety, and the Multicultural Supervision Competencies Questionnaire (Wong & Wong, 2014) assesses supervisor multicultural competency. While supervisee evaluation, anxiety, and multicultural competency are useful variables to address, the preceding instruments are only used to assess how effective supervision is at changing, via increase or decrease, the construct of interest. What is required of supervision research is an a priori theoretical model of (a) the essential elements of effective supervision and (b) a way to measure the impact of effective supervision.

Similarly, supervisor competence may only be assessed indirectly. Competence may seemingly be quickly assessed by review of a supervisor's resume or vita, but to what degree does supervisor competency surface in the supervisory endeavor? Multiple supervision competencies have been developed over the years. Across the helping professions in the U.S., including psychology (Falender et al., 2004; Falender & Shafranske, 2004) and counseling (Association for Counselor Education and Supervision, 1990; Dye & Borders, 1990), to Australia (Psychology Board of Australia, 2018) and the United Kingdom (Roth & Pilling, 2015), models of competent supervision have been articulated. While expert-developed lists of competencies exist, the larger concern is how supervisor competency is operationalized and assessed. To move beyond competency lists, supervisor competency research requires an a priori

theoretical model of (a) what a competent supervisor does within supervision and (b) a way to measure supervisor competence.

The degree to which supervision is effectively and competently delivered is of critical concern as indicated in recent research exploring the experience of post-master's counselors in the U.S. (Cook, 2019) and Ireland (Ellis et al., 2015). Cook (2019) reported upwards of 78% of counselors experience inadequate supervision and 30% experience harmful supervision. These are striking statistics in which it may be fairly concluded that too many counselors experience ineffective supervision delivered by not-so-competent supervisors. In order to address *supervision effectiveness* and *supervisor competence* on a profession-wide scale, instruments are required that can reliably and validly measure the degree of effectiveness of supervision and the level of competence of the supervisor across settings. These two constructs, critically, are not directly measurable so instruments that are psychometrically robust and theoretically-grounded are essential. Thus, we require an organized and theoretically coherent model in order to meaningfully operationalize supervision effectiveness and supervisor competence.

**The Proctor Model**

The Proctor Model of Supervision (Proctor, 2011) provides such a model to conceptualize and understand what essential elements of supervision require measurement. Built out from the early work of Kadushin (1985), Proctor's (2011) model of supervision articulates three functions of supervision. The simplicity and parsimony of Proctor's model is its strength. Namely, supervision serves three purposes that may be broken into three distinct domains: restorative, formative, and normative (Proctor, 2011). The restorative domain concerns emotional processing, experiencing, wellbeing, and supervisee self-awareness. The formative domain concerns the maintenance of supervisee competence, self-reflective capacity, and

effectiveness to provide clinical services. The normative domain concerns professional decorum, ethical and legal responsibilities of the role of counselor, and client management issues. While the three domains are considered complementary and somewhat overlapping, the distinct functions of supervision provide helpful conceptual targets for measurement in assessing the overall utility of supervision. The Proctor Model has demonstrated such utility that it has been used internationally, interprofessionally, and subjected to empirical psychometric scrutiny in Australia and the United Kingdom. To extend and contribute to ongoing supervision research efforts, the theoretical integrity of the Proctor Model is thus worth scrutinizing with a new sample of U.S.-based CITs engaged in clinical supervision.

**Supervision Evaluation in the CACREP Context**

The leading education accreditation body of the counseling profession is the Council for Accreditation of Counseling and Related Education Programs (CACREP). CACREP accredits professional counseling programs offering entry-level counseling specialties such as clinical mental health, school, rehabilitation, and addictions. CACREP maintains standards for supervision within graduate education of future professional counselors engaged in clinical work. Programs that utilize clinical supervision include clinical mental health counseling, rehabilitation counseling, addictions counseling, and marriage and family counseling. In the *2016 CACREP Standards* (CACREP, 2015), accredited programs are required to address the role of supervision within the profession (Section 2.F.1.M), infuse counseling and supervision-related research into the curriculum (Section 2.E), provide opportunities for supervision (Section 1.I), and the evaluation of supervision (Section 4.J, K). Though CACREP requires accredited programs to conduct evaluation of supervision, it has not specified the use of specific assessment instruments or evaluation methods, or provided guidance on the ideal psychometric properties of instruments

utilized. It stands to reason that CACREP expects programs to use evaluation methods that conform to best practice guidelines in the profession.

CACREP-accredited programs are well-incentivized to follow such training standards to maintain accreditation, evaluate program outcomes, and ensure the welfare of trainees. Programs are encouraged to share *Best Practice Guidelines* (Borders et al., 2011) with supervisors as "a large number of counseling professionals who provide clinical supervision are master's-level clinicians who have never received formal supervision training themselves" (Borders et al., 2014, p. 29). Similarly, Luke (2019) noted that many counselor education programs, in order to meet the training and higher education demands, employ supervisors who are part-time instructors, adjunct faculty, or doctoral students with varying degrees of supervision experience. As such, supervisees engaged in clinical supervision may have multiple supervisors, inexperienced supervisors, or find themselves receiving inadequate or harmful supervision. In order for CACREP-accredited programs to comply with the accreditation requirements to engage in robust program evaluation with a view to produce competent professional counselors, programs would need to use meaningful and psychometrically sound methods to evaluate supervision and monitor supervisee-supervisor relationships.

## Dissertation Overview

As clinical supervision research is international in nature (Goodyear et al., 2016; Watkins, 2012; White & Winstanley, 2014), and premised on the basis of reasoning by analogy (Milne, 2006, 2014), it is thus appropriate to consider investigating the psychometric properties of internationally and interprofessionally developed supervision instruments. The two instruments selected for study — the Supervision Evaluation and Supervisor Competency Scale (SE-SC; Gonsalvez et al., 2016) and the Manchester Clinical Supervision Scale-26 (MCSS-26;

Winstanley & White, 2014) — were initially developed and subsequently validated for use within Australian and British contexts. The instruments were developed in accordance with the Proctor Model described earlier. Neither of them has been validated for use in U.S.-based counseling training contexts. Each instrument is described in detail in the following chapters.

**Research Questions**

In order to address the call for increased rigor in supervision research and the quality of available supervision instruments, the main research question connects the following two chapters: Do existing supervision evaluation instruments maintain rigorous psychometric evidence for a sample of CITs from CACREP-accredited programs?

Each instrument was subjected to multiple statistical analysis: Rasch modeling for polytomous responses, internal consistency, and validity. Detailed data analysis plans are described further within the methodology section for each chapter. The research questions for the MCSS-26 (Winstanley & White, 2014) include:

1. Does the MCSS-26 and its subscales possess evidence of internal consistency?
2. Does the MCSS-26 possess item-level fitness?
3. When compared with a measure of training environment, does the MCSS-26 possess evidence of concurrent validity?
4. Does social desirability present a significant threat to the validity of the MSCSS-26?

Chapter 2 (Manuscript 1) details the study to investigate the questions listed above for the MCSS-26. Chapter 3 (Manuscript 2) describes the investigation of the psychometric properties for the SE-SC (Gonsalvez et al., 2016). It represents the second manuscript in this dissertation. Research questions in this manuscript include:

1. Does the SE-SC possess internal consistency?

2. Does the SE-SC possess item-level fitness?
3. When compared with a measure of supervisory relationship, does the SE-SC possess concurrent validity?
4. Does social desirability present a significant threat to validity?

**Research Participants and Sampling Procedures**

Participants for this study hailed from master's-level CACREP-accredited counselor education programs. As the focus of the two manuscripts was on clinical supervision, CITs engaged in clinical supervision were the target sample. Based on clinical experience comparability, CITs in specialties in clinical mental health counseling, addictions counseling, rehabilitation counseling, and marriage and family counseling were contacted and invited to participate in the studies via faculty members in these programs who had direct contact their trainees.

Sampling procedures are described in further detail in the proceeding manuscripts. Of note, one sampling was conducted for both studies. Data gathered from this sample was analyzed for each of the two studies articulated below. Sampling commenced upon approval from Oregon State University's Institutional Review Board (Appendix A). Data collection was conducted online between February 2020 to April 2020 using a secure platform, Qualtrics. A full listing of materials used throughout recruitment, including the entire 130 question survey, is available in Appendix B.

## Conclusion

Clinical supervision remains a difficult target for empirical research because of the multidimensional complexity (Lambert, & Ogles,1997) of the supervision intervention that occurs in a professional relationship (Watkins, 2014) with varying issues requiring attention

(e.g., ethical, legal, clinical, personal, & interpersonal). Supervision research is further complicated by the lack of psychometrically robust instruments (Dawson et al., 2013; Olds & Hawkins, 2014; Watkins, 2012), many of which have not been validated for use across professional groups, clinical settings, and national boundaries. In order to warrant use across professions and settings, supervision instruments require evidence to suggest their reliability, validity, and performance across contexts. DeVellis (2017) articulates this notion succinctly: "As with reliability, validity is not an inherent property of a measurement tool but of the tool in the context of its use. A tool may be valid in one context but invalid in another or when put to a different use." (p. 86). The need for supervision instruments that are useful for professional practice, and based on substantive psychometric evaluation, is critical to advancing an understanding of why supervision is important and how it works. Such calls for an increase in scholarly supervision research rigor are not limited to counseling (Schutt, 2012), but also include allied psychiatric nursing (Buus & Gonge, 2009) and psychiatry (MacDonald & Ellis, 2012). The original research presented herein seeks to contribute to ongoing scholarly efforts to address this deficit within clinical supervision research by executing a cross-validation study on two supervision instruments related to supervision effectiveness and competence, respectively.

This research has direct implications for the training of professional counselors, their supervisors, and future supervision research. Namely, in providing supportive or contrary evidence to suggest the use of psychometrically validated supervision instruments, programs may make more evidence-based decisions for their CITs. As CACREP-accredited programs require the tools to accomplish their self-studies and satisfy accreditation standards, the results of this study may also inform how programs assess, monitor, and evaluate supervisors. Results of this dissertation project are also expected to provide evidence to support or not support the

operationalization of the Proctor Model of supervision in the U.S. counseling training context. Lastly, supervision scholarship and research within an international and cross-cultural perspective will benefit from the current research given the need for (a) psychometrically valid and contextually relevant supervision instruments, (b) greater scrutiny of the utility and effectiveness of supervision, and (c) heightened awareness of methodological issues within supervision research.

## Glossary of Terms

**Concurrent Validity** - "the extent to which test scores have a stronger relationship with criterion (gold standard) [*sic*] measurements made at the time of test administration…" (Boateng et al., 2018, p. 14).

**Cross-Validation Study** - also called model validation; a procedure in which the a priori factor structure of a scale, or the predictive power of a regression equation, is assessed by using the previous a priori theoretical factor structure, or equation, based on to what degree the model holds within the new sample (adapted from Mertler & Vannatta, 2017, p. 346)

**Dendogram** - a classification tree based on the hierarchy of related or "natural" groupings from a hierarchical cluster analysis (Fonesca, 2013, p. 406). The output of a hierarchical cluster analysis.

**Instrument** - "a manifestation of latent constructs; they measure behaviors, attitudes, and hypothetical scenarios we expect to exist as a result of our understanding of the world, but cannot assess directly [sic]." (Boateng et al., 2018, p. 1).

**Item Difficulty –** notation of difficulty = *b.* A psychometric property of an item that indicates the location of the item, or the ease with which the item may be activated according to the latent trait/variable; the higher the difficulty of an item, the more amount – or ability level – required to successfully respond to the item (Boone, 2016).

**Item Discrimination –** notation of discrimination = *a*. A psychometric property of an item that indicates to what degree the item relates to the construct/latent trait. Item discrimination is based on a logarithmic slope function; the steeper the slope, the more discerning, or discriminatory, the item is between those with and without the presence of the latent trait/ability (Edelen & Reeve, 2007).

**Item-Level Fitness** - “items at the more difficult end of the variable [within a Rasch model] should be harder to correctly answer than items at the easy end of the continuum. This should be true for all students answering a set of items regardless of their ability levels. If items do not fit the model, they may measure more than one variable.” (Boone, 2016, p. 5)

**Latent Variable** - “The underlying phenomenon or construct that a scale is intended to reflect...it is latent rather than manifest...the construct is variable rather than constant - that is, some aspect of it, such as its strength or magnitude, changes...” (DeVellis, 2017, p. 24) measurement theory

**Multicollinearity** - “problem created when independent variables are very highly correlated ($r \geq .90$) with each other” (Mertler & Vannatta, 2017, p. 349)

**Polytomous Item Response Model –** a form of item response model (e.g. Graded Response or Generalized Response) in which there an item possesses more than two categories for item responding; also called polychotomous (Boone, 2016).

**Rasch Modeling** - A theory and set of mathematical techniques that “allow nonlinear data to be converted to a linear scale, which then can be evaluated through the use of parametric statistical tests.” (Boone, 2016, p. 7)

**Rescaled Distance** – units of measurement within a dendogram; an index of proximity range with close associations indicated by low numbers (Gonsalvez et al., 2017)

**Response Categories –** a quality of an item (e.g., dichotomous or polytomous); the number of options available to responders to the item – could range between 2 (dichotomous) to more than 2 (polytomous) (e.g., Likert scales with 1-5 indicator options) (Andrich & Marais, 2019).

**Social Desirability** - “a person responding to a test in a manner that he/she feels will present them in a positive light (i.e., faking good)” (Ventimiglia & MacDonald, 2012, p. 487)

**Supervision Competence** - “being qualified, knowledgeable, and able to act in a consistently appropriate and effective manner - reflecting critical thinking, judgement, and decision making - that is in accordance with standards, guidelines, and ethics of the particular profession being practices” (Milne & Watkins, 2014, p. 8) as assessed by the SE-SC.

**Supervision Effectiveness** - the degree to which the intervention of supervision addresses supervisee needs per the Proctor Model (across restorative, formative, and normative domains), as measured by the MCSS-26.

**Supervision/supervisory Relationship** - “a socially embedded educational practice” (Watkins, 2017, p. 204) comprised of the working alliance, transference phenomena, and the real relationship (Watkins, 2015).

**Theoretical Coherency** - property of a scale; “factor loadings are understood in light of their conceptual underpinnings” (Mvududu & Sink, p. 79)

## Abstract

Supervision instruments that assess supervision effectiveness require ongoing scrutiny, testing, and development. As supervision scholars call for a common measurement approach to build a foundation for more advanced, multivariate research designs, this study sought to address this call. We designed a cross-validation study of the Manchester Clinical Supervision Scale (MCSS-26; Winstanley & White, 2014), an instrument frequently used to assess supervision effectiveness, with particular emphasis on applying an item response theory lens of analysis. A total of 86 participants, who were counselors-in-training at CACREP-accredited institutions in the United States, completed an internet-based survey. Results demonstrated acceptable instrument-level validity and reliability psychometrics, but multiple poor-fitting items according to the Generalized Partial Credit Model (GPCM). Based on these results, we offer a revised 9-item version of the MCSS-26. Findings suggest further scrutiny of the MCSS-26 before use with a U.S.-based CIT population.


*Keywords:* supervision effectiveness, MCSS-26, supervision instruments, psychometric evaluation, item-response theory

## Chapter 2: A Validation Study of the Manchester Clinical Supervision Scale

For any field to advance through quantitative inquiry, it requires the use of instruments that operationalize constructs into measurable terms and possess acceptable psychometric properties specific to the population under study (DeVellis, 2017). An instrument's utility is constrained by the context of the available reliability, validity, and precision evidence. Thus, additional and continuous investigation is necessary to advance the utility of the instrument beyond its initial context, such as considering sample representativeness, sample demographics, theoretical coherency, and item-level performance (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). Such continuous effort pertains to the field of supervision as well.

According to supervision scholars, research demonstrating the impact and importance of supervision has several notable flaws. These include methodological concerns (Buus & Gonge, 2009), lack of clear operationalization, few longitudinal studies, and the absence of a "common measurement approach" (Wheeler & Barkham, 2014, p. 380; see Watkins, 2017). Clarifying whether or not an instrument and its items measure the same latent variables, or constructs, across groups is central to building a common measurement approach (Floyd & Widaman, 1995). This is particularly salient given the importance of multicultural assessment validity (Ridley et al., 2008). Though conceptually seeming to be unwieldy, significant efforts have been made to define clinical supervision as a counseling specialty (Borders et al., 2014). The present study represents an attempt to further the discourse on the need for psychometrically sound and contextually applicable supervision-related instruments by investigating the psychometric properties of the MSCC-26 in a sample of U.S.-based counselors-in-training (CIT).

Supervision is an intervention that promotes professional development and skill acquisition within the counseling profession (Borders, 2005). Supervision is a highly contextual and dynamic process (Watkins, 2017) that is tailored to the needs of the system (e.g., agency, training program) (Watkins & Milne, 2014). Further, supervision is (a) an intensive education, (b) a relational intervention, (c) developmentally supportive, and (d) both supervisee and client welfare focused (Bernard & Goodyear, 2014; Milne, 2007). This complexity results in supervision remaining a critical topic of research (Watkins & Milne, 2014), insofar as the effects (e.g., on supervisee, client, community) of its practice remain unclear (White, 2018). To begin to support claims of supervision effectiveness across contexts and systems, supervision evaluation research requires psychometrically robust instruments that have been validated on representative sample groups, known also as reference groups (Goodyear et al., 2016).

Supervision scholars and researchers have articulated concern about the lack of consistently utilized and psychometrically robust instruments to evaluate supervision (Dawson, Phillips, & Leggat, 2013; Ellis et al., 1996; Olds & Hawkins, 2014; Schutt, 2012; Watkins, 2012). Such instruments would lend themselves to the development of a "cumulative and coherent knowledge base [of supervision]" (Wheeler & Barkham, 2014, p. 380) that would facilitate making inferential, predictive, or meta-analytic research of supervision possible. Scholars (Bernard & Goodyear, 2014; Wheeler & Barkham, 2014) attribute this methodological issue to the literature's wide utilization of one-time use instruments developed by researchers for their specific purposes, oftentimes without engaging in robust cross-validation or accounting for measurement invariance (Vandenberg & Lance, 2000). Apt to the present study, "there is little replication of supervision studies, even though there are many cultural differences in supervision across the globe and across modality [sic]" (Wheeler & Barkham, 2014, p. 367). In brief, a need

exists for psychometrically sound instruments that can be utilized within larger professional efforts to organize and focus evidence to bolster claims of effectiveness of supervision (Gonsalvez & Calvert, 2014; Milne & Reiser, 2012) across contexts (e.g., sociocultural milieus, training environments, & work settings) and disciplines (e.g., counseling, nursing). Concerns for supervision effectiveness is particularly salient to counselor education programs as supervision is a signature andragogy of counselor education (Luke, 2019).

**Supervision in Counselor Education**

Supervision is a "mechanism for professional socialization" (Luke, 2019, p. 37) and an avenue for the development of counseling skills, professionalism, and ethical decorum. Supervision of CITs occurs in all professional counselor preparation programs starting in practicum and ending in internship. The premium placed on supervision within the counseling profession is evident in its requirements in training programs, licensure, and accreditation standards (e.g., CACREP, 2015).

The critical role supervision plays necessitates systematic evaluations of its processes and outcomes as evidenced by such requirements in training accreditation standards (CACREP, 2015). These evaluations include, but are not limited to, (a) assessing the quality of supervision, (b) assessing the effect of supervision, (c) gatekeeping inadequate or harmful supervisors, and (d) satisfying accreditation requirements. To achieve such purposes, theoretically sound models and psychometrically robust tools are needed. These tools will also facilitate robust systematic inquiries to inform evidence-based counselor education and supervision practices.

To date, counseling supervision research has fully emerged as a specialty, though methodological issues remain, as alluded to earlier (Watkins, 2012). Researchers have investigated various aspects of supervision, including processes and outcomes from perspectives

of supervisors, supervisees, and observers (e.g., supervision working alliance; Watkins, 2014, 2017). For the purposes of this study, we will briefly discuss major themes of supervision research related to CITs' experience of supervision effectiveness.

While CITs experience a range of worry, anxiety, and imposter syndrome that are developmentally normative (Bernard & Goodyear, 2014; Skovholt & Rønnestad, 1992), nowhere is the trainee more vulnerable than in the supervisory relationship. Nelson and Friedlander (2001) note, "... trainees are vulnerable to poor judgement on the supervisor's part" (p. 385) that may exist without the knowledge of faculty and administration. The concern of supervision evaluation and effectiveness is not solely important for accreditation but is critically linked to supervisee welfare, as well. CITs require effective supervision at this most vulnerable time in their careers as they are being socialized into the profession, engaging in deliberate practice, and acquiring professional skills and habits that will serve them throughout their career. Unfortunately, not all CITs experience effective supervision while in graduate school. For example, Ellis et al.'s (2013) study revealed that 93% of supervisee-participants reported receiving inadequate supervision in their current supervisory relationship, while 35.5% of supervisee-participants reported experiencing harmful supervision in their current supervisory relationship.

In another study, a cross-national comparative analysis between the United States and the Republic of Ireland indicated that many supervisee-participants were receiving inadequate supervision (81%, Republic of Ireland; 75%, United States) and harmful supervision (40%, Republic of Ireland; 25%, United States) at the time of the study (Ellis et al., 2015). Ellis et al. (2015) observe that there is a striking discrepancy between supervisee's perception of inadequate and harmful supervisor behavior and the supervision they receive. In understanding the

supervisor's way of being through the experience of the supervisee, researchers may begin to create a clearer picture of the factors of the supervision relationship that effectively lead to change (Watkins, 2017).

To further illustrate the importance of the supervisee perspective, Gray et al. (2001) interviewed supervisees who experienced counterproductive events and observed the impact on supervisory process and outcomes: "Not only did most of the trainees feel uncomfortable, unsafe, or upset in response to the counterproductive events, but they also deferred to the supervisors' authority, became hypervigilant, nondisclosed, and withdrawn in supervision" (p. 381). As harmful supervision may impact supervisees just as harmful therapy impacts clients (Barlow, 2010; Ellis et al., 2014; Ellis, 2017), the impetus to monitor and detect supervision adequacy and effectiveness rests with the training program (Karpenko & Gidycz, 2012).

Ongoing efforts to further understand and evaluate the learning environment of counseling training programs (see Lau & Ng, 2014; Lau et al., 2019) highlight the importance of assessment frameworks as part of an organizing strategy to improve program and student learning outcomes (Walker & Fraser, 2005). Supervision evaluation, as part of program evaluation, intends to assess whether or not "... the behaviour of the supervisor [leads] to measurable changes in the practice of the supervisee and enhanced outcomes for the recipient of psychological services" (Gonsalvez & McLeod, 2008, p. 84). While client-based supervision outcomes are beyond the scope of our study (see Simpson-Southward et al., 2017), a dimensional perspective of supervision effectiveness provides a useful organizing framework (Milne, 2014; Watkins, 2014).

Multiple questions emerge from the existing literature on supervision. For example, how does supervision affect the supervisee and their development? How do we know if the

supervision provided to CITs is effective or adequate? Given the multifaceted and dynamic nature of supervision, effectiveness may be determined by more than simply client improvement, but also by how supervision meets the needs of the supervisee across restorative, formative, and normative domains. Proctor (2011) proposes a model to address these domains.

## The Proctor Model

The Proctor Model (Proctor, 2011), a widely influential model of clinical supervision (Spence et al., 2001), conceptualizes the "complementary and sometimes contradictory tasks" (p. 25) of supervision with three constructs: *restorative*, *formative*, and *normative*. The restorative domain of supervision addresses the supervisee's wellbeing, resilience, and self-awareness. The formative domain of supervision addresses the self-reflective learning and growing-through-experience nature of supervision. The normative domain of supervision addresses the professional standards, role responsibilities, and ethical concerns that the supervisee experiences. These domains explain a supervisory relationship that is effective for professional performance.

The Proctor Model has been operationalized by White and Winstanley (2010; Winstanley & White, 2011; 2014) and demonstrates evidence of validity across multiple work settings and professions per the ongoing development of the Manchester Clinical Supervision Scale (MCSS-26; Winstanley & White, 2014) in a number of different countries, with the exception of the United States. In articulating trans-theoretical common factors and, specifically, common supervisory tasks, Watkins (2017) identifies six tasks (e.g., nurture facilitative supervisory relationship, develop supervision plan to address supervisee learning needs, provide ongoing monitoring of supervisee progress) that map onto the three domains of the Proctor Model. Due to its utility as a parsimonious framework for conceptualizing effective supervision, the Proctor Model, through the ongoing development of the MCSS-26, has been used as a heuristic model

for supervision across multiple health professions and healthcare settings. Thus, the MCSS-26 seems to be an instrument that can be used trans-professionally, across work and cultural settings.

## Manchester Clinical Supervision Scale

The MCSS-26 (Winstanley & White, 2014) is a revised version of the 36-item MCSS (Winstanley & White, 2011). It is completed by supervisees about their supervision experience to determine the effectiveness of supervision. The MCSS-26 stands out as an instrument that has sound psychometric properties, robust statistical analytic support, translations into seven languages from the original English version, and wide-use in over 100 clinical supervision studies. For example, it was used to (a) compare the effectiveness of supervision between allied health professionals working in large public hospital settings (Snowdon et al., 2016), (b) determine the relationship between supervision effectiveness and patient outcomes for rehabilitation professionals working in inpatient contexts (Snowdon et al., 2019); and (c) assess the relationship between supervision and workplace satisfaction for drug and alcohol counselors (Best et al., 2014). Further, the MCSS-26 has been validated for use within 14 countries across multiple helping professions, namely psychiatric nurses, speech pathologists, dieticians, occupational therapists, podiatrists, social workers, and psychologists, in a variety of settings (e.g., hospital and community based) (Snowdon et al., 2016; Winstanley & White, 2014).

Initially developed and validated by Winstanley (2000) as a 59-item instrument, the MCSS went through several revisions resulting in its current 26-item (6 subscales) format (Winstanley & White, 2011). The scale consists of six subscales under three constructs: restorative (5 items on *trust/rapport*, 5 items on *supervisor advice/support*), formative (4 items on *improved care/skills*, 3 items on *reflection*), and normative (5 items on *importance/value of*

*supervision*, 4 items on *finding time*). Responses to questions are framed with a 5-point Likert scale, ranging from 1 (*Strongly Disagree*), 3 (*No Opinion*), to 5 (*Strongly Agree*), and scored 0 to 4. A total score is computed by summing all 26 items with possible scores ranging from 0 to 104. The higher the score, the higher the level of effectiveness of supervision (Winstanley & White, 2014). Winstanley and White (2014) hypothesize that a score of ≥73 may signal the threshold for effective supervision, or "70% of possible maximum" (p. 392).

Evidence suggests that the MCSS-26 possesses strong validity for use within clinical settings. Rasch analysis of the MCSS-36, based on archival data of nursing ($n$ = 225) and allied health staff ($n$ = 160), resulted in improved model fitness, leading to a revised version by elimination of items that were redundant, misfit, and had low Person Separation Index (PSI) (Winstanley & White, 2011). The MCSS-36, at the time had a seven-factor structure, accounted for 64.4% of the observed variance (Winstanley & White, 2011).

In one study of the original MCSS, reliability coefficients were reported to range from .64 to .88 for subscales (Hyrkäs et al., 2003), while test-retest reliability was reported with intraclass correlation coefficients that ranged from .78 to .87 (Winstanley & White, 2011). Further, Winstanley and White (2014) subjected archival data of the MCSS-26 to a Classification and Regression Tree (CART) analysis to determine which factors would predict a high score. A strong correlation between the MCSS-36 and MCSS-26 ($r$ = 0.975) suggests the utility of the revised scale (Winstanley & White, 2019); but, replication studies in other supervision contexts need to be conducted to verify Winstanley and White's (2011) findings. To date, the MCSS-26 has not been validated for use with the CITs population in the United States.

## Purpose of the Study

In an attempt to address some of these methodological concerns in supervision research, these two studies seek to systematically validate the MCSS-26 (White & Winstanley, 2014) for use in the United States with master's-level CITs from programs accredited by the Council for Accreditation of Counseling and Related Educational Programs (CACREP). In view of the measurement gaps in the supervision literature discussed above, the present cross-validation research (Messick, 1995) sought to extend the utility of the MCSS-26 by providing psychometric statistics and sample generalizability to the CITs population in the United States. Due to the clinical context of the instrument and the focus of the current study, CITs from clinical mental health counseling, rehabilitation counseling, addictions counseling, and family counseling training programs formed the target sample for our study. It is our belief that CITs from these specialties have more contextual commonality in their supervision experience. The aim of this study is therefore to test the psychometric properties of the MCSS-26 with a representative sample of CITs in the United States who have clinically based supervision experience.

The main research question was: Is the MCSS-26 relevant, reliable, and valid for use with counselors-in-training in the United States? From this overarching question, we sought to address the following research questions:

1. Does the MCSS-26 and its subscales possess evidence of internal consistency?
2. Does the MCSS-26 possess item-level fitness?
3. When compared with a measure of training environment, does the MCSS-26 possess evidence of concurrent validity?
4. Does social desirability present a significant threat to the validity of the MSCSS-26?

The last question came from the belief that power differential is inherently present in supervision relationships with supervisees being in the less-than position and subject to evaluation apprehension (Ellis et al., 2008). As such, it is important to examine if supervisee-completed measures on supervision are susceptible to social desirability as has been demonstrated in multicultural competency research (Gonzalez et al., 2018).

Based on previous research, we hypothesize that item performance of the MCSS-26 can be replicated with a sample of CITs from the United States and there is acceptable evidence of internal reliability and concurrent validity to support the utility of the measure among U.S. CITs. It is our hope that this study will contribute to ongoing supervision research efforts that seek to identify valid instruments for use within supervision practice, evaluation, and development of novice and expert supervisors alike across sociocultural and professional contexts.

## Methods

To address the above research questions, we conducted a cross-validation. A cross-validation study would assist in determining multiple psychometric properties of the MCSS-26. We utilized Rasch modeling analytics for polytomous responses to address the questions.

### Participants

All participants were voluntary adults, aged 18 or older, and satisfied inclusion criteria. Inclusion criteria included individuals who self-identified as a CITs pursuing their master's degree at a CACREP-accredited program (clinical mental health counseling, clinical rehabilitation counseling, addictions counseling, or marriage and family counseling) and were engaged in clinical supervision at the time of the survey.

A total of 135 participant responses were recorded for this study. However, only 86 participant responses were usable and included for analysis due to (a) eligibility criterion

satisfaction and (b) full completion of the study survey. Participants were required to complete all survey questions, as indicated in the informed consent, but were allowed to skip demographic questions to protect anonymity. Appendix C presents full sample demographics. Of those who identified their gender (n = 81, 94% reporting), 71(88%) identified as female, 8 (10%) as male, and 2 (2%) as non-binary. With respect to race (n = 80, 93% reporting), 67 (83%) identified as Caucasian/European/White, 4 (5%) identified as Asian, 3 (4%) identified as Black/African, 3 (4%) identified as Latinx/Hispanic/Spanish, and 3 (4%) identified as Multiracial. The sample demographics were different from demographic statistics for master's students in CACREP accredited programs as reported by CACREP in their 2017 annual report (CACREP, 2018), though with some notable differences. The current sample had a few differences to the overall CACREP master's student population, however. The sample had a greater Caucasian/White representation (83% versus 60% in CACREP annual report; CACREP, 2018) and multiracial identity representation (4% versus 2% in CACREP annual report; CACREP, 2018). Female representation in this sample was also greater at 88% versus 83% in the CACREP annual report (CACREP, 2018). The current sample notably also had decreased Latinx/Hispanic/Spanish representation (4% versus 8% in CACREP annual report; CACREP, 2018), Black/African racial identity (4% versus 19% in CACREP annual report; CACREP, 2018), and male representation (10% versus 17% in CACREP annual report; CACREP, 2018). The annual report indicates data on racial identities including American Indian/Native American, Native Hawaiian/Pacific Islander, Non-resident Alien, and "Other." However, none of these groups were represented in the current sample.

Participants also identified their sexual minority status (*n* = 80, 93% reporting), international student status (*n* = 81, 94% reporting), and age (*n* = 76, 88% reporting). Sixteen

(20%) of participants identified as a sexual minority. Five (6%) identified as international student. Participants ranged in ages from 22 to 69 ($M$ = 30, $SD$ = 8), with 53 (70%) between the ages of 22-29, 16 (21%) between the ages of 30-39), 3 (4%) between the ages of 40-49, 3 (4%) between the ages of 50-59, and 1 (1%) between the ages of 60-69.

### *Program Region, Specialization, and Delivery Method*

Participants hailed from across the country, with various program specializations, and program delivery methods. Some participants reported their program's region ($n$ = 81, 94%). Participants were from all ACES regions: 27 (33%) from NCACES; 25 (31%) from SACES; 16 (20%) from NARACES; 8 (10%) from WACES; and 5 (6%) from RMACES. All participants reported program specialty type (n = 86), with the majority ($n$ = 71, 83%) from clinical mental health counseling and a few from rehabilitation counseling ($n$ = 11, 13%) and marriage and family counseling ($n$ = 4, 5%) comprising the rest of the participants. Compared to the annual report (CACREP, 2018), the clinical mental health counseling specialization was overrepresented by 27% and the marriage and family counseling specialization was overrepresented by 2%. Due to the recent CORE-CACREP merger (Tew-Washburn, 2016), there was no existing data on rehabilitation counseling program enrollment within the annual report. Participants also reported their program delivery methods ($n$ = 86): traditional face-to-face ($n$ = 59, 69%), hybrid ($n$ = 15, 17%), and online ($n$ = 12, 14%).

### *Accrued Clinical Hours*

Participants varied in reported practicum and internship hours accrued with 77 (90%) reporting their direct and indirect combined hours ($M$=372, $SD$=329). In the current sample, 18 (23%) reported less than 100 hours, 17 (22%) reported between 101-200 hours, 4 (5%) reported

between 201-300 hours, 3 (4%) reported between 301-400 hours, and 35 (45%) reported more than 400 hours accrued during training.

***Supervisor Type and Supervision Setting***

Participants identified their supervisor's relationship with the graduate program and the setting where they received supervision upon which they based their responses to the research instrument. Not all participants reported supervisor relationship with the graduate program (*n* = 85, 99%). Supervisors were identified by participants as current faculty (*n* = 15, 18%), site supervisors (*n* = 60, 71%), doctoral students (*n* = 8), or other (*n* = 2, 2%). Participants who identified "other" denoted that they have multiple supervisors across their site and school settings. All participants reported the setting and location of where they received supervision: 29 (34%) at a university clinic, 31 (36%) at an agency or community mental health center, 15 (17%) at a private practice, 4 (5%) at a group practice, and 7 (8%) over telesupervision.

***Supervision Frequency, Duration, and Modality***

Participants reported the frequency, duration, and modality for the supervision they received. All of them reported the frequency and duration of their supervision sessions. A majority of participants (*n* = 78, 91%) reported weekly supervision, 7 (8%) reported biweekly supervision, and 1 (1%) participant reported receiving supervision less than once every three months. Participants reported an average supervision session lasting for 46-60 minutes (*n* = 49, 57%), longer than 60 minutes (n = 21, 24%), between 31-45 minutes (*n* = 12, 14%), between 15-30 minutes (*n* = 3, 3%), and less than 15 minutes (*n* = 1, 1%). Some participants reported supervision modality (*n* = 84): individual supervision (*n* = 52, 62%), a mix of individual and group supervision (*n* = 22, 26%), triadic supervision (*n* = 6, 7%), and group supervision (*n* = 4, 5%).

### *Supervisee and Supervisor Theoretical Orientation*

Every participant reported their theoretical orientation and their supervisor's theoretical orientation ($n$ = 86). As participants were allowed to select multiple theoretical orientations, $n$ counts add up to over 86. Cognitive-behavioral ($n$ = 42, 49%), humanistic ($n$ = 41, 48%), eclectic ($n$ = 26, 30%), interpersonal ($n$ = 23, 27%), systems ($n$ = 13, 15%), psychodynamic ($n$ = 12, 14%), reality/choice theory ($n$ = 4, 5%), dialectical-behavioral ($n$ = 3, 3%), existential ($n$ = 2, 2%), and feminist ($n$ = 2, 2%) were represented within the sample. Additional theoretical orientations included "Adlerian" ($n$ = 1), "attachment" ($n$ = 1), "eye-movement desensitization and reprocessing (EMDR)" ($n$ = 1), "somatic experiencing" ($n$ = 1), "trauma" ($n$ = 1), "social constructivism" ($n$ = 1) and "not determined" ($n$ = 1).

Participants identified their supervisor's theoretical orientation as well. Cognitive-behavioral ($n$ = 43, 50%), humanistic ($n$ = 30, 35%), eclectic ($n$ = 20, 23%), systems ($n$ = 19, 22%), interpersonal ($n$ = 14, 16%), psychodynamic ($n$ = 11, 13%), dialectical-behavioral ($n$ = 3, 3%), and Gestalt ($n$ = 3, 3%). Additional theoretical orientations included "Adlerian" ($n$ = 1), "attachment" ($n$ = 1), "brief solution focused" ($n$ = 1), and "somatic experiencing" ($n$ = 1).

## Procedures

Prior to participant recruitment, we sought approval from the university's Institutional Review Board. Because we do not have direct access to CITs in CACREP-accredited programs, and in order to recruit a sample is that close to be representative of the population of CITs in CACREP-accredited programs, we conducted recruitment through convenience sampling methods electronically. We reached out to program liaisons and faculty teaching in CACREP-accredited programs requesting their help in recruiting participants. As of November 2019, 880 counseling programs are accredited by CACREP. Of these programs, 505 offered specialties in

clinical mental health, rehabilitation, addiction, or family counseling. We created a database of program liaisons and faculty contacts for each of these programs for recruitment purposes. We sent an email to program liaisons and faculty for each of the programs listed in the database. We sent only one email to a program's liaison, regardless of the number of specialties offered. For example, a program may offer two master's degrees - one in family counseling and one in rehabilitation. In this case, we only sent one email to that program's liaison and faculty as they satisfied the inclusion criteria simply by having one specialty in-house and thus offer clinical supervision. We also sent similar recruitment emails to counselor educators who are professional contacts of the research team.

The email to program liaisons and faculty included (a) the scope of the study, (b) research participant informed consent, and (c) a request to share the invitation to participate with currently enrolled counselors-in-training. After two weeks, liaisons and faculty received a follow-up reminder and a gratitude email. We offered, as incentive, participants an opportunity to enter a drawing for 1 of 8 $20 Starbucks gift cards. Participants were informed that their participation is voluntary and anonymous. However, if they chose to enter the gift card drawing, they provided the researchers their email addresses in a separate survey. Participants' responses were not matched to their email address.

Participants completed the study materials via a secure web-based survey platform that uses encryption. The survey was accessible to participants for nine weeks and took about 10-15 minutes to complete, no more than 20 minutes (Revilla & Ochoa, 2017). Over the course of nine weeks, the first author emailed the recruitment invitation at specific intervals (0, 3, 6 weeks). At week nine, a final notice email announced the closure of the survey. The study website included a description of the purpose of this study, participant selection criteria, procedures, consent

information and documents, the survey questionnaires (i.e., demographic questionnaire, MCSS-26). We secured permission to use all survey instruments included in the study from their developers and related publishing companies.

## Measures

The following measures were included in the survey for participants to complete. The research materials were set as forced choice to avoid missing values. We informed participants of this forced choice setting within the informed consent, so that they were well aware of the time demand of the survey.

### *Demographic Questionnaire*

The questionnaire collected self-report data about the participant's gender, age, race, stage of training, counseling specialization, time in program, theoretical orientation (self and supervisor), supervision context, supervision relationship duration, frequency of supervision meetings, and average duration of supervision meetings. These variables are consistent with prior psychometric evaluation studies of supervision instruments (e.g., Gonsalvez et al., 2017; Lehrman-Waterman & Ladany, 2001; Palomo et al., 2010) and supervision evaluation research (Ellis et al., 2013; Lambie et al., 2018). Bambling (2014) and Kemer et al. (2019) highlight the importance of considering supervisee personal characteristics and contextual factors as possible predictors of supervision.

### *Manchester Clinical Supervision Scale*

The Manchester Clinical Supervision Scale (MCSS-26; White & Winstanley, 2014) was described in-depth in a previous section of this article. Participants were asked to respond to the items based on their experience of supervision with a current supervisor. Cronbach's alpha for the current study was .92.

### *Counseling Training Environment Scale*

The 23-item Counseling Training Environment Scale (CTES; Lau et al., 2019) assesses the training environment of counseling and related mental health training programs from the vantage of the supervisee. Initially developed and validated using mixed confirmatory factor analysis (CFA) and item response theory (IRT) methods by Lau et al. (2019), the original validation sample included 277 clinical trainees from accredited programs (CACREP, American Psychological Association, Commission on Accreditation for Marriage and Family Therapy Education, Masters in Psychology and Counseling Accreditation Council, American Art Therapy Association), with the majority of participants (58.8%) hailing from CACREP-accredited programs. A five-factor structure was supported through CFA, with 23 items loading across five subscales microsystem ($n = 5$), mesosystem ($n = 6$), exosystem ($n = 4$), macrosystem ($n = 4$), and chronosystem ($n = 4$). Higher scores indicate positive perceptions of the training environment, lower scores indicate less positive perceptions. Overall scale reliability was reported as $\alpha = 0.92$. Each subscale was reported to maintain adequate reliability: microsystem ($\alpha = 0.75$), mesosystem ($\alpha = 0.77$), exosystem ($\alpha = 0.70$), macrosystem ($\alpha = 0.60$), and chronosystem ($\alpha = 0.81$). In addressing the macrosystem alpha and other psychometric properties of the instrument, Lau et al. (2019) articulate their decision making to retain the subscale based on the systemic and overlapping framework of Bronfenbrenner's (1992) theory. The authors also cautioned researchers to evaluate collinearity closely for this subscale, in future studies. Lau et al. (2019) reported strong test-retest reliability over the course of two weeks ($r = 0.93$, $p < 0.01$, two-tailed). The CTES has been recommended by the developers for program evaluation and monitoring student outcomes throughout the training program (e.g., satisfaction, retention). As a measure of broader training environment factors, subscales of the CTES were expected to

provide evidence of concurrent and divergent validity to the MCSS-26. Cronbach's alpha for the global score for the CTES in the current study was .86.

### *Marlowe-Crowne Social Desirability Scale – Short*

The Marlowe-Crowne Social Desirability Scale – Short – Form A (MCSDSS-A; Reynolds, 1982) assesses participant bias in self-reporting. In their systematic evaluation of multiple short versions of the original MCSDS (Crowne & Marlowe, 1960), Loo and Thorpe (2000) indicated support for Reynolds' (1982) short version of the MCSDS (Forms A and B). In this study, we utilized Form A of the MCSDSS per the scrutiny and evidence considered in Loo and Thorpe's (2000) analysis and shorter parsimony of the scale.

The MCSDSS-A is an 11-item dichotomous scale constructed to assess bias within participant responses. Participants indicate "True/False" in response to multiple statements that target socially desirable responding. For example, "No matter who I'm talking to, I'm always a good listener." After summing scores according to developer guidelines, higher scores indicate evidence of socially desirable responding. For the original MCSD, Reynolds (1982) reported a Kuder-Richardson reliability (KR[20] = .74) and a high Pearson correlation ($r = .91$, $p < .001$) with the MCSDSS. Loo and Thorpe (2000) reported a Cronbach's alpha of .59 for the MCSDSS-A. With the current sample, the MCSDSS-A had a Cronbach's alpha of .72 and a KR[20] of .72.

## Data Preparation Plan

Prior to conducting analyses, based on recommendations in the literature (e.g., Tabachnick & Fildell, 2019), we addressed the accuracy of the data, accounted for missing data, and screened outliers. Mertler and Vannata (2017) also describe the import of data accuracy so that the integrity of statistical conclusions, and the whole study, are ensured.

### *Addressing Accurate Data*

To ensure accuracy of participant data, instruments were completed online, digitally. This reduces the possibility of mis-entering data into a digitized file. To ensure coherency, descriptive statistics were evaluated for plausibility.

### *Addressing Missing Data*

We designed a forced choice survey that would not allow participants to progress without responding to each question presented. As such, we removed the possibility of missing data.

### *Screen for Outliers*

Data more than three standard deviations from the mean were considered outliers and then removed from the reliability analysis, per the standard deviation outlier labeling method (Tabachnick & Fidell, 2019). For item-response analysis such responses are meaningful in determining model fitness, so items were not removed.

## Item-Level Analysis

The Generalized Partial Credit Model (Muraki, 1992; 1993) was adopted for analyzing item-level fitness of polytomous responses. Using maximum likelihood estimations (MLE), the Generalized Partial Credit Model (GPCM) is flexible enough for item performance analysis given the lack of forced assumptions about item discrimination ability and response category intervals. The GPCM explores the latent construct/trait performance according to the measurement model (Muraki, 1992): item-level responsiveness is a function of a latent construct (e.g., ability or agreeableness), or theta (θ), and the difficulty of an item's response structure (threshold categories). Thus, the probability of any items' response categories being selected is a function of the construct's presence in the respondent. The GPCM is a less constrained model without specific intervals of item response categories or any item's difficulty so it tends to result

in a more accurate reflection of the data (Embretson & Reise, 2000) compared to other polytomous models (e.g., GSM). The GPCM, like other item-response models, is premised on a logistic mathematical model of probability. In employing MLE, values of parameters are estimated that "maximizes the probability that this set of responses is observed according to the model" (Andrich & Marais, 2019, p. 113). Items, performing as the model would presume, possess a sequential pattern across response categories, or thresholds (difficulty). In short, as a theta (latent ability) increases (e.g., *moderate agreeableness to the item*) so, too, does probability of selecting a sequentially higher category of responding (e.g., *agree to strongly agree*).

Importantly, the assumptions of the GPCM, and any item-response theory derived model, include unidimensionality and threshold parameters. These parameters were explored for each subscale of the SE-SC as it is theoretically assumed that certain items, identified as a domain/function (e.g., normative, formative, restorative) of the Proctor Model, capture the same latent construct. Exploratory factor analysis (EFA; Ziegler & Hagemann, 2015) was utilized to assess unidimensionality and examine if items were tapping a similar construct (Andrich & Marais, 2019; Baker, 2001; Toland, 2014). A factor loading of .40 or greater was considered acceptable for the EFA. Structural assumptions of the model, previously discussed as threshold parameters or response categories, were examined for sequential responding patterns (e.g., $b_1 = -2$, $b_2 = -1$, $b_3 = 0$…). Items that violated this response structure were considered nonconforming, or in violation of, the model.

Items' unique outfit and infit mean squared were also calculated. Outfit stats were considered unacceptable if they fell outside of 0.6-1.4 (Wright & Linacre, 1994). Of note, outfit stats were primarily used for model assessment because outfit calculations have been discussed

by model experts to be more sensitive to determining misfit as compared to infit stats (Andrich & Marais, 2019).

To assess polytomous model fit, *p*-values of the S-$\chi^2$ and root means square error of approximation (RMSEA) were examined for each item. In reviewing the RMSEA, significance was determined by a gradation of fitness as articulated by Browne and Cudeck (1992): $p > 0.1$ = poor fit, $p < 0.08$ = reasonable fit, and $p < 0.05$ = close fit. The S-$\chi^2$ (Kang & Chen, 2007) for polytomous models produces the degree of similarity between observed and model-based frequencies per response category. To determine mis-fitness for the S-$\chi^2$ a statistically significant value ($p < .05$) is required.

In brief, the GPCM was employed to estimate the following parameters: item location or difficulty (*b*), item discrimination (*a*), and error estimates.

## Results

All calculations were executed within the R environment (R Core Team, 2019) version 1.9.12.31 with psych (Revelle, 2019), mirt (Chalmers, 2012), and ltm (Rizopoulos, 2006) on an iMac running macOS Catalina version 10.15.4. Data of the sample ($n = 86$) possessed no missing data issues due to the forced-response of the instrument questions. Results are presented according to related research question.

### Internal Consistency

Reliability of the MCSS-26 was assessed using Cronbach's $\alpha$ (Cronbach, 1951). Reliability estimate calculated for the MCSS-26 global score was $\alpha = .92$. Those for the subscales .79 for Normative, .82 for Formative, and .89 for Restorative. They all exceeded the recommended .70 (Cortin, 1993).

**Item-Level Fitness**

The following parameters were explored for each item of the MCSS-26 ($M$ = 85, $SD$ = 13): item location or difficulty ($b$), item discrimination ($a$), and error estimates. In order to determine satisfaction of Rasch model assumptions, namely unidimensionality and item independence, items were analyzed according to subscale of the MCSS-26: Normative, Formative, and Restorative. Item-level parameter estimates are presented in Table 1.1 and 1.2. Item trace lines for the MCSS-26 are presented in Appendix D. Test information curves are presented in Appendix E.

***Normative***

The Normative subscale was composed of Items 1, 2, 3, 4, 5, 6, 8, 16, and 20 (White & Winstanley, 2014). Unidimensionality, according to EFA loadings, was acceptable with seven items ranging from .46-.76. Two items—Items 16 and 20—did not load acceptably with .29 and .38, respectively. Nonconforming items, or items that did not possess sequential response categories, included Items 1, 2, 6, 4, and 20. Items that did not fit the model due to underrepresented response thresholds (see Table 1.2) included Items 6, 4, 3, 8, and 16. Item discrimination ($a$) estimates are presented in Table 1.1. Item difficulty ($b$) was calculated across items by taking the mean across item thresholds (e.g. $b_1$, $b_2$; see Table 1.1). Of note, the spread of item discrimination tends to be limited given the response categories ranging from 1-5.

As shown in Table 1.1, items with outfit stats outside of an acceptable range of 0.6-1.4 (Wright & Lincacre, 1994) were identified as misfitting. Within the Normative subscale, Item 4 is considered misfitting with an outfit stat of .396. With a significance level for S-$\chi^2$ set at 0.05 (Chon et al., 2010), items with p-values less than .05 were considered not conforming to the

model (Item 6, $p = .003$). Items' RMSEA $p$-values that poorly fit the model included Item 1 ($p = .089$), Item 3 ($p = .099$), Item 4 ($p = .109$), Item 6 ($p = .135$), and Item 8 ($p = .092$).

***Formative***

The Formative subscale was composed of Items 9, 10, 11, 14, 22, 23, and 26 (White & Winstanley, 2014). Items were determined to be unidimensional with EFA loadings ranging from .49-.75. All items were above the .40 cutoff threshold.

Item response categories for Items 14 and 11 did not fit the GPCM. Underrepresented response thresholds were indicated for Items 9, 10, 22, and 26. Items with outfit stats outside of the acceptable range included Item 22 and Item 26 (Table 1). Error estimates (S-$\chi^2$) for all items indicated conformity to the model. Goodness of fit (RMSEA) values for Items 23 ($p = .00$) and 26 ($p = .00$) indicated a close fit. Items 9 ($p = .072$) and 22 ($p = .058$) reasonably fit the model while Items 10 ($p = .097$), 11 ($p = .126$), and 14 ($p = .180$) poorly fit the model.

***Restorative***

The Restorative subscale was composed of Items 7, 12, 13, 15, 17, 18, 19, 21, 24, and 25 (White & Winstanley, 2014). The majority of items were determined to be unidimensional with EFA loadings ranging from .43-92. However, Item 21 did not reach .40 with a loading of .30. Items 18, 7, 19, 25, and 21 violated the model's assumption of sequential response categories. Underrepresented response categories were indicated for Item 13. Items with outfit stats outside of the acceptable range included Item 17 and Item 25 (Table 1.1). Error estimates (S-$\chi^2$) for Items 24 ($p = .014$) and 18 ($p = .012$) did not conform to the model. Goodness of fit (RMSEA) values for Items 12 ($p = .00$), 13 ($p = .044$), 19 ($p = .459$), and 21 ($p = .00$) indicated a close fit. Items 7 ($p = .071$) and 15 ($p = .083$) reasonably fit the model and Items 17 ($p = .098$), 18 ($p = .143$), 24 ($p = .139$), and 25 ($p = .123$) poorly fit the model.

**Table 1.1**

*MCSS-26 Item Parameters, Ranked by Difficulty*

| Subscale | Item No. | Item Difficulty (Response Categories) | Item Discrimination | MNSQ Outfit | MNSQ Infit |
|---|---|---|---|---|---|
| Restorative ($n$ = 10) | 24 | -0.55 (5) | 1.145 | 0.925 | 1.192 |
| | 12 | -0.91 (5) | 1.023 | 0.937 | 0.986 |
| | 17 | -1.14 (5) | 5.693 | *0.445* | 0.667 |
| | 13 | *-1.15 (4)* | 1.921 | 0.826 | 1.002 |
| | 18 | -1.24 (5) | 1.525 | 0.834 | 0.938 |
| | 15 | -1.24 (5) | 2.301 | 0.728 | 0.917 |
| | 7 | -1.31 (5) | 2.275 | 0.722 | 0.753 |
| | 19 | -1.34 (5) | 2.584 | 0.703 | 0.954 |
| | 25 | -1.44 (5) | 4.922 | *0.414* | 0.932 |
| | 21 | -1.65 (5) | 0.368 | 1.052 | 1.001 |
| Formative ($n$ = 7) | 10 | *-1.17 (4)* | 1.436 | 0.855 | 0.941 |
| | 26 | *-1.31 (3)* | 4.469 | *0.413* | 0.652 |
| | 23 | *-1.33 (4)* | 1.251 | 0.889 | 0.913 |
| | 14 | -1.36 (5) | 2.212 | 0.753 | 1.033 |
| | 11 | -1.365 (5) | 1.356 | 0.893 | 1.018 |

| Subscale | Item No. | Item Difficulty (Response Categories) | Item Discrimination | MNSQ Outfit | MNSQ Infit |
|---|---|---|---|---|---|
| | 9 | *-1.45 (4)* | 1.51 | 0.878 | 0.947 |
| | 22 | *-1.46 (4)* | 2.599 | *0.528* | 0.893 |
| Normative ($n$ = 9) | 1 | -0.83 (5) | 0.564 | 0.929 | 0.940 |
| | 2 | -1.04 (5) | 0.808 | 0.971 | 0.925 |
| | 6 | *-1.06 (4)* | 0.881 | 0.988 | 1.051 |
| | 5 | -1.22 (5) | 1.74 | 0.742 | 0.871 |
| | 4 | -1.23 (5) | 3.468 | *0.396* | 0.830 |
| | 3 | *-1.34 (4)* | 2.131 | 0.771 | 0.848 |
| | 8 | *-1.60 (4)* | 1.57 | 0.846 | 0.933 |
| | 20 | -1.94 (5) | 0.588 | 1.071 | 1.043 |
| | 16 | *-2.93 (3)* | 0.818 | 0.959 | .011 |

*Note.* MNSQ = mean square. Misfits are italicized if MNSQ Outfit < .4.

**Table 1.2**

*MCSS-26 Item Parameters, Response Thresholds*

| Subscale | Item No. | $b_1$ | $b_2$ | $b_3$ | $b_4$ |
|---|---|---|---|---|---|
| Restorative ($n$ = 10) | 24 | -2.099 | -2.197 | -1.989 | -0.113 |
| | 12 | -2.288 | -1.223 | -0.839 | 0.706 |

| Subscale | Item No. | $b_1$ | $b_2$ | $b_3$ | $b_4$ |
|---|---|---|---|---|---|
| | 17 | -1.835 | -1.402 | -1.181 | -0.125 |
| | 13 | -1.777 | -1.568 | -0.1 | NV |
| | 18 | -1.22 | *-1.934* | *-1.594* | -0.192 |
| | 15 | -1.904 | -1.454 | -1.284 | -0.304 |
| | 7 | -1.827 | -1.422 | *-1.657* | -0.348 |
| | 19 | -2.264 | *-1.337* | *-1.634* | -0.115 |
| | 25 | -2.044 | -1.597 | *-1.669* | -0.326 |
| | 21 | -1.299 | *-3.238* | *-2.482* | 0.436 |
| Formative ($n$ = 7) | 10 | -2.187 | -1.734 | 0.405 | NV |
| | 26 | -2.083 | -0.527 | NV | NV |
| | 23 | -3.261 | -1.268 | 0.526 | NV |
| | 14 | -2.166 | -0.943 | *-2.13* | -0.198 |
| | 11 | -1.425 | *-2.545* | *-2.031* | 0.541 |
| | 9 | -2.425 | -2.021 | 0.086 | NV |
| | 22 | -1.929 | -1.914 | -0.526 | NV |
| Normative ($n$ = 9) | 1 | -4.041 | 1.925 | *-2.633* | *1.432* |
| | 2 | -2.703 | 1.394 | *-3.183* | *0.346* |
| | 6 | -0.899 | *-1.866* | -0.417 | NV |
| | 5 | -2.729 | -0.752 | *-1.352* | -0.052 |
| | 4 | -1.467 | *-1.658* | -0.571 | NV |
| | 3 | -2.197 | -1.796 | -0.355 | NV |
| | 8 | -2.227 | -2.183 | -0.395 | NV |
| | 20 | -2.237 | 0.386 | *-4.935* | *-0.973* |
| | 16 | -4.71 | -1.143 | NV | NV |

*Note.* NV = no value.

**Revised Version of the MCSS-26**

On the basis of the item level fitness studies described above, we proposed a revised 9-item version to better fit the GPCM at the subscale and item level. This 9-item version is available in Appendix F. We provide data about fit indices, model parameters, and localized-likelihood information criteria (Akaike's and Bayesian estimates) for the 9-item version in Tables 1.3, 1.4, and 1.5 to demonstrate the superiority of the 9-item version compared to the 26-item version. For interpretation, lower estimates of the AIC and BIC are considered preferable to higher estimates. This is based on the assumption that, when comparing quality of model fitness, that the distance (or value) of the AIC/BIC estimate is considered closer to the "truth," or at a higher probability, when the distance is smaller (Dziak et al., 2020). Thus, the revised 9-item version of the MCSS-26 satisfies item-response theory and model assumptions to a more superior degree than the full 26-item instrument.

**Concurrent Validity**

In determining concurrent validity of the MCSS-26, the CTES ($\alpha$ = .86) was adopted as the instrument of comparison given its previous validation with counselor-in-training samples. Using Pearson's *r* to determine the ratio of covariance between two variables (total score on each instrument; Mertler & Vannata, 2017), the correlation was calculated in the same psych package in R as described previously. Between the MCSS-26 and the CTES, there was a small statistically significant association ($r$ = .18, $p$ = .098) at the $p < .10$ level. Similarly, between the revised 9-item MCSS instrument and the CTES there was a small statistically significant association ($r$ = .21, $p$ = .058) at the $p < .10$ level.

### Assessing Reactivity Threats to Validity

In determining participant-level reactivity as a possible source of threat to validity, a Pearson correlation was calculated between the MCSS-26 and the MCSDSS-A. No statistically significant association between the MCSS-26 and the MCSDSS-A ($\alpha$ = .72) was identified within the sample ($r$ = -0.0031, $p$ = .98). No statistically significant association between the revised 9-item instrument and the MCSDSS-A ($r$ = 0.079, $p$ = .47) was determined either.

## Discussion

The purpose of this study was to test the psychometric properties of the MCSS-26 with a sample of CITs in the US who were receiving clinical supervision at the time of the study. As an instrument that evaluates supervision effectiveness, the MCSS-26 is a prime candidate to consider for practice, training, and research. Thus, this study sought to scrutinize the psychometric item-level properties of the MCSS-26 employing a polytomous item response theory model. A critical aim towards testing the psychometric properties of the MCSS-26 was to determine individual items' fitness to the GPCM model for the current sample. Based on the results presented above, a reduction in items of the MCSS-26, and a recalibration of response categories are indicated. Next we discuss the findings, limitations, and relevant implications for instrument development, counselor education, and supervision research.

### Instrument Validity and Reliability

The MCSS-26 demonstrated reliability ($\alpha$ = .92) for this sample of US-based counselors-in-training. The low correlation between the MCSS-26 and CTES ($r$ = .18) suggests that the two instruments are measuring separate constructs. When considering that the MCSS-26 is a measure of supervision-only, and the CTES is a measure of ecological training environment, it could be

expected that the two instruments would not have a high correlation. In brief, the low correlation between the MCSS-26 and CTES may actually be an indicator of discriminate validity.

With social desirability as a possible threat to validity due to participants' evident power-under role within supervision, our findings yielded no association ($r$ = -0.0031, $p$ = .98) between the MCSS-26 and the MCSDSS-A. With traditional testing psychometrics assessed, we next discuss the MCSS-26's item-level fitness to the GPCM.

**Model Fitness**

From our tests of model fitness, we proposed a revised 9-item version of the MCSS to better fit the data at the subscale and item level. We attended to multiple considerations when revising the MCSS-26 including item revision, elimination, and a short-form creation. As the worst fitting items were identified above, the resulting items of the MCSS-26 that most appropriately fit the model and its assumptions include Items 12, 15, 17 (Restorative), 9, 10, 23 (Formative), and 3, 8, 16 (Normative). Based upon GPCM response fit, as presented in Table 1.3, a 9-item shortened form of the MCSS-26 is endorsed by the sample data. Of note, no response scale system was determined acceptable for all items of the shortened scale. Table 1.4 identifies the fit statistics for the revised scale. Absolute fit statistics, like the RMSEA and the S-$\chi^2$ are particularly sensitive to sample size and the number of items as evidenced by the misfitting values in Table 1.4. However, the GPCM of the revised scale results in smaller values of the Bayesian information criterion (BIC) and the Akaike information criterion (AIC), see Table 1.5. Both the BIC and the AIC are relative information criteria and offer a simple metric of comparison between the original subscale and the revised subscale. Across all subscales, the BIC and AIC were smaller (indicating a better fit) for the revised scale compared to the original MCSS-26. It is worth noting that while our findings suggest a proposed 9-item revision based on

the GPCM, the MCSS-26 requires original developer permission to be revised as a licensed instrument.

We attended to multiple considerations when revising the MCSS-26 including item revision, elimination, and a short-form creation. The classical test theory analysis of validity and reliability psychometrics of the MCSS-26 were beyond the scope of this study. Yet we believe that our findings regarding the GPCM item-level performance of the instrument are supportive of item-level restructuring of the measurement model due to the doubt case on the item-level validity of the 26-item version of the MCSS-26. While the MCSS-26 has accumulated evidence of classical test validity and reliability in many other samples, the findings in the current sample of US-based CITs suggest otherwise when applying an item-response theory level of analysis. At the instrument-level there appears to be some evidence to suggest score validity with the current sample (e.g., Cronbach alpha), the item-level performance, based on this sample, of the MCSS-26 require further scrutiny before it is implemented for use with US-based CITs.

**Table 1.3**

*MCSS-26 Item Parameters, Revised 9-Item Instrument*

| Subscale | Item No. | *a* | $b_1$ | $b_2$ | $b_3$ | $b_4$ |
|---|---|---|---|---|---|---|
| Restorative (*n* = 3) | 12 | 0.801 | -2.659 | -1.297 | -0.932 | 0.748 |
| | 15 | 1.934 | -2.196 | -1.52 | -1.336 | -0.29 |
| | 17 | 6.055 | -2.09 | -1.438 | -1.167 | -0.051 |
| Formative (*n* = 3) | 9 | 2.018 | -2.48 | -1.829 | 0.122 | NV |
| | 10 | 1.471 | -2.352 | -1.715 | 0.427 | NV |
| | 23 | 0.895 | -4.112 | -1.456 | 0.603 | NV |
| Normative (*n* = 3) | 3 | 2.269 | -2.517 | -1.809 | -0.292 | NV |
| | 8 | 1.088 | -2.73 | -2.664 | -0.467 | NV |
| | 16 | 0.757 | -5.071 | -1.203 | NV | NV |

*Note.* NV = no value.

**Table 1.4**

*MCSS-26 Scale Fit Statistics, Revised 9-Item Instrument*

| Subscale | Item No. | *EFA Loading* | MNSQ Outfit | *RMSEA*[*] | S-$\chi^{2*}$ |
|---|---|---|---|---|---|
| Restorative (*n* = 3) | 12 | .54 | .913 | .064 | .246 |
| | 15 | .76 | .689 | *.388*[^] | *.00*[^] |
| | 17 | .96 | *.160* | NaN[^] | NaN[^] |
| Formative (*n* = 3) | 9 | .56 | .568 | *.366*[^] | *.00*[^] |
| | 10 | .60 | .691 | *.203*[^] | *.004*[^] |
| | 23 | .75 | .846 | *.297*[^] | *.00*[^] |
| Normative (*n* = 3) | 3 | .74 | *.393* | NaN[^] | NaN[^] |
| | 8 | .53 | .768 | *.352*[^] | *.00*[^] |
| | 16 | *.30* | .865 | *.191*[^] | *.017*[^] |

*Note*. Misfitting values are italicized. [*]= *p-values*, [^]= $\chi^2$ reported if NaN for S-$\chi^2$

**Table 1.5**

*MCSS-26 and Revised 9-Item Scale with GPCM AIC and BIC*

| Subscale | AIC | BIC |
|---|---|---|
| Restorative Original | 1617 | 1738 |
| Restorative Revised | 580 | 617 |
| Formative Original | 1011 | 1082 |
| Formative Revised | 522 | 551 |
| Normative Original | 1510 | 1606 |
| Normative Revised | 425 | 452 |

*Note*. AIC= Akaike's Information Criterion; BIC = Bayesian Information Criterion

**Limitations**

A primary limitation of the study was the small size of the sample. While the sample size was sufficient to conduct the planned psychometric analyses, further studies with larger samples are needed to validate the revised nine-item version of the MCSS. In determining recruitment strategy and working towards a large sample, multiple barriers and limitations were prevalent. Findings and results should thus be considered in context to the limitations of this study.

Absolute goodness-of-fit tests, like the RMSEA and the S-$\chi^2$ , are sensitive to sample size (Sharma et al., 2005). Estimation parameters are also affected by sample size. Broadly speaking for item response theory applications, the bigger the sample size, the more "fitting" and precise the fit stats and the estimation parameters. Thus, a notable limitation of this study is the precision of the estimation parameters. Parameters may be conceptually grounded, but are not as precise if applied to another sample. Additionally, multiple items (8 of 9) of the Normative subscale are reverse scored items. Items that are reverse scored are notoriously rife with theoretical issues due to their violation of assumption of the presence of an ability/trait and, therefore, often resulting in poor model fits (cf. Weijters et al., 2013). Future research applying item response theory to the MCSS-26, perhaps for large-scale calibration purposes, will need to attend to estimation parameters using a large dataset ($n > 500$). In brief, item fit and parameter estimates reduce the practical generalizability of the study findings to other samples or the population of supervisees at-large.

Despite efforts to recruit a nationally representative sample of CIT from CACREP-accredited programs, the study sample was not comparable in representativeness to the 2017 annual report (CACREP, 2018). As previous research on sample representativeness indicate, baseline sample demographic characteristics may influence a significant portion of outcome

variance (McGlashan et al., 1988) and tend to be a challenge in online survey designs (Vicente & Reis, 2007). Therefore, sample representativeness needs to be addressed in order to develop data credibility and coverage of the target population (Chow, 2002; Ramsey & Hewitt, 2005). The study sample was overrepresented by female and Caucasian/White participants as compared to the CACREP master's student population (CACREP, 2018). Black/African participants were also underrepresented in the study sample. Sample representativeness is a critical limitation as study findings are thus limited in the applicability to CACREP-accredited programs. It is possible that the demographic makeup of the sample might have influenced aggregate responses to the MCSS-26 and CTES. This limitation is important to note for later research attempting cross-study comparison across various samples and instrument use decision making based on this study. While the sample was not representative of the population of CITs described in the 2017 annual report, the sample had a higher representation of Asian CITs and non-binary CITs than in the population, and represented all five ACES regions.

A barrier to recruitment efforts was the pass-through method of contacting CACREP Liaisons at all CACREP-accredited programs and requesting them to forward the recruitment material. Without direct outreach to potential participants, this circuitous method may have impacted the sample size. Study recruitment may also have been impacted by external environmental events such as the global pandemic of COVID-19 (Zhou et al., 2020). In the middle of recruitment outreach, on the week of the second planned contact and the third final contact, global and national anxiety were heightened (McGinty et al., 2020) and clinical training efforts across counseling programs were disrupted to varying degrees (CACREP, 2020). Notwithstanding the above-mentioned limitations, findings in this study contribute to a larger

body of evidence suggesting the ongoing refinement and revision of the MCSS-26 for use with trainees.

## Implications and Recommendations

From the results of this study, we propose a revision of the 26-item MCSS to a 9-item instrument for counselor education. This study also bears implications for the ongoing construction of a common measurement approach for quantitative supervision research methodology.

### *Instrument Revision*

This study is the first to consider item-level fitness of the MCSS-26 to a Generalized Partial Credit Model (GPCM) with a US-based sample. Historically, item response theory treatments of the MCSS-26 have relied on archival data from international repositories. The results suggest a possible revised solution to fit the GPCM model across the three subscales of Restorative (Items 12, 15, 17), Formative (Items 9, 10, 23), and Normative (Items 3, 8, 16). The suggested revised scale, as seen in Appendix E test information curves, possesses greater precision. Further, the revised scale also possesses clearer discrimination ability across items (Table 1.3). If an item and its conceptual peers are more cleanly able to discriminate the presence of a supervisee's agreeableness to the construct being assessed, then the more clearly supervisors, administrators, or researchers will be able to detect effective supervision – or ineffective supervision!

The main benefit of the revised 9 item MCSS-26 (Appendix F) may be in the feasibility of its integration into supervision evaluation or research methodologies. Shorter scales that are less time-consuming present less of a burden to participants and a smoother data collection strategy for administrators and researchers (Ziegler et al., 2014). Based on preliminary

correlation between the MCSS-26 and the revised scale ($r = .93$, $p < .00001$) a high-level of correlation was achieved, which is critical for ongoing instrument revision and development. Internal reliability of the revised scale was also determined to be excellent ($\alpha = .92$). Further research is needed to verify the validity and utility of this revised measure in U.S.-based counselors as well as non-U.S.-based clinical populations.

### *Counselor Education Programs*

Current supervision evaluation research aims to assess if a supervisor's interventions produce measurable change in the supervisee and the supervisee's practice. Applied, or clinical, training is the assumed responsibility of counselor education programs, regardless of specialty. In order to systematically assess if the supervision being offered is effective, programs require valid, reliable, and precise tools. As few psychometrically sound instruments exist for supervision effectiveness evaluation (Ellis et al., 2008; Watkins & Milne, 2014), it is incumbent upon counseling researchers to develop such instruments for use in training programs and in clinical practice. It is thus important that counselor educators have robust tools to select and implement in supervision effectiveness evaluation of site supervisors, faculty supervisors, and supervision-of-supervision (metavision). Research (Cook, 2019; Ellis et al., 2015) continues to routinely demonstrate that harmful and inadequate supervision occur at less than acceptable rates. Our proposed 9-item MCSS may present an option, upon further research and refinement, for programs seeking to assess supervision effectiveness during clinical training.

### *Advancing a Common Measurement Approach for Supervision Research*

The current study is the first to explore the MCSS-26 item-level fitness to a polytomous Rasch model with a US-based counselor-in-training sample. The MCSS-26 is already an internationally recognized supervision instrument (Winstanley & White, 2014), the advent of

using the MCSS-26 with a US-based sample is novel in and of itself. While the MCSS-26 has not been normed or validated for use with a US-based population yet, it might have great utility for future research. Our study contributes to the international effort to build a common measurement approach within supervision research architecture. However, our findings raise questions about measure's psychometric properties for use in its current form in U.S. counseling training programs. Our findings highlight the caution for adopting measures across cultures and settings without first systematically examining their psychometric properties for the population on which they are to be applied (DeVellis, 2017). Further research is needed to verify our findings as well as verifying the cross-cultural and cross-setting psychometric properties of the measure prior to considering the MCSS-26 as being an empirically supported common measure for supervision research to facilitate international supervision research collaborations and cross-cultural comparative studies.

Further, in order to break the cycle of single-study instruments within US-based supervision scholarship, it is incumbent upon US-based supervision researchers to contribute to the international scholarship that is focused on bolstering claims of the effectiveness of supervision across work contexts, disciplines, and modalities of clinical mental health care. The counseling profession originated in the United States and there are ongoing efforts to introduce and support the counseling profession in other countries (e.g., NBCC, 2020). In regard to psychometric measurements specifically, many instruments that were developed in the U.S. are confined to use within the U.S. Instruments developed in other countries have not been commonly adopted within the United States. As the counseling profession continues to develop a global footprint, a two-way synergistic relationship is needed whereby the U.S. counseling profession reviews and utilizes instruments and approaches developed outside of the United

States. The MCSS-26 is an instrument developed in the United Kingdom, utilized internationally, that has strong psychometric properties that could be useful for U.S. counseling. For example, with additional research, CACREP-accredited programs in the U.S. may find value from using the MCSS-26 to evaluate the quality of supervision in their programs. However, this study indicates more work on the ongoing development and refinement of the MCSS-26 for use with a US-based sample of counselors-in-training in CACREP-accredited programs is needed.

As supervision research requires an increasingly diverse methodological body of work, supervision effectiveness evaluation instruments that are precise and relevant to the population of inquiry are a necessity. Indeed, as the counseling profession internationalizes and formalizes professional association collaborations (Ng, 2012; Ng et al., 2012; NBCC, 2020), counseling research is well-positioned to contribute to the scholarly international and interdisciplinary supervision body of research. One such critical effort is the construction of a common measurement approach insofar as quantitative methodologies and statistical analysis may advance from descriptive to structural to predictive. Clear and shared measurement models are the key to international research collaboration for supervision scholars.

The MCSS-26 could have excellent utility for counselor education programs - both master's and doctoral-level programs. Across its multiple iterations (Winstanley & White, 2011; 2014) the MCSS-26 has held up to robust statistical scrutiny and demonstrated utility time and time again, making it one of the most valuable instruments in the field of supervision research. As the field moves towards a common measurement approach, the MCSS-26 will require large datasets, constant validation, performance assessment, and theoretical scrutiny (Ziegler et al., 2014). In brief, though our findings indicate caution against considering the MCSS-26 as a

common measurement in supervision research, our study represents an attempt toward an international effort to build a common measurement approach in supervision research.

### *Multicultural Implications*

Given the limitation of the sample's representativeness of the larger CACREP enrollee population, the instrument in its current form as well as it revised, 9-item format should be carefully used with supervisees who are of diverse racial background. The MCSS-26, while explicitly measuring supervision effectiveness, does not include a multicultural construct within its underlying theoretical foundation. Thus, we believe the MCSS-26 is not an appropriate instrument if multicultural considerations, dynamics, or outcomes are a critical element of supervision effectiveness assessment. Perhaps, in order to more precisely capture supervision effectiveness, future iterations and revisions of the MCSS-26 could include a multicultural component. Of course, the addition of a theoretical construct to the instrument may constitute a divergence from the underlying Proctor Model. However, if the MCSS-26 is to be responsive to ongoing supervision effectiveness evaluation needs, and multicultural competency is considered a foundational element for supervisee clinical performance, then the MCSS-26 requires a multiculturally responsive element.

## Conclusion

We examined the psychometric properties of the MCSS-26, with particular attention given to a GPCM. Based on our findings, we propose a 9-item version of the MCSS-26. Our findings suggest the need for continued refinement and development of the MCSS-26 with US-based samples in clinical training. We believe that our study has contributed significantly toward the development of the MCSS-26 and the discourse on the issues and challenges related to finding common measurement approach for quantitative supervision research.

# References

Althubaiti, A. (2016). Information bias in health research: Definition, pitfalls, and adjustment methods. *Journal of Multidisciplinary Healthcare*, *9*(1), 211-217. https://doi.org/10.2147/JMDH.S104807

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Andrich, D., & Marais, I. (2019). *A course in Rasch measurement theory: Measuring in the educational, social and health sciences*. Springer.

Baker, F.B. (2001). *The basics of item response theory*. College Park, MD: ERIC Clearinghouse on Assessment and Evaluation.

Bambling, M. (2014). Creating positive outcomes in clinical supervision. In C.E. Watkins, Jr. & D.L. Milne (Eds.), *Wiley international handbook of clinical supervision* (pp. 445-457). Oxford, United Kingdom: Wiley.

Barlow, D. H. (2010). Negative effects from psychological treatments: A perspective. *American Psychologist, 65*(1), 13–20. https://doi.org/10.1037/a0015643

Bernard, J.M., & Goodyear, R.K. (2014). *Fundamentals of clinical supervision* (5th edition). Boston, MA: Merrill.

Best, D., White, E., Cameron, J., Guthrie, A., Hunter, B., Hall, K., ... & Lubman, D. I. (2014). A model for predicting clinician satisfaction with clinical supervision. *Alcoholism Treatment Quarterly*, *32*(1), 67-78. https://doi.org/10.1080/07347324.2014.856227

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L.

(2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health, 6*(1). https://doi.org/10.3389/fpubh.2018.00149

Bond, T. G. & Fox, C. (2015). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (3rd ed.). New York, NY: Routledge.

Borders, L. D. (2005). Snapshot of clinical supervision in counseling and counselor education: A five-year review. *The Clinical Supervisor, 24*(1-2), 69-113. https://doi.org/10.1300/J001v24n01_05

Borders, L. D., Glosoff, H. L., Welfare, L. E., Hays, D. G., DeKruyf, L., Fernando, D. M., & Page, B. (2014). Best practices in clinical supervision: Evolution of a counseling specialty. *The Clinical Supervisor*, *33*(1), 26-44. https://doi.org/10.1080/07325223.2014.905225

Bronfenbrenner, U. (1992). Ecological systems theory. In R. Vasta (Ed.), *Annals of child development. Six theories of child development: Revised formulations and current issues* (pp. 187–249). London, UK: Jessica Kingsley.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York, NY: Guilford Press.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological methods & research*, *21*(2), 230-258. https://doi.org/10.1177/0049124192021002005

Buus, N., & Gonge, H. (2009). Empirical studies of clinical supervision in psychiatric nursing: A systematic literature review and methodological critique. *International Journal of Mental Health Nursing, 18*(4), 250-264. https://doi.org/10.1111/j.1447-0349.2009.00612.x

Chalmers, R., P. (2012). mirt: A Multidimensional Item Response Theory Package for the R

Environment. *Journal of Statistical Software, 48*(6), 1-29. https://doi.org/10.18637/jss.v048.i06

Chon, K. H., Lee, W. C., & Dunbar, S. B. (2010). A comparison of item fit statistics for mixed IRT models. *Journal of Educational Measurement*, *47*(3), 318-338. https://doi.org/10.1111/j.1745-3984.2010.00116.x

Chow, S. L. (2002). Issues in statistical inference. *History and Philosophy of Psychology Bulletin*, *14*(1), 30-41.

Copeland, P., Dean, R. G., & Wladkowski, S. P. (2011). The power dynamics of supervision: Ethical dilemmas. *Smith College Studies in Social Work*, *81*(1), 26-40. https://doi.org/10.1080/00377317.2011.543041

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of applied psychology*, *78*(1), 98. https://doi.org/10.1037/0021-9010.78.1.98

Council for the Accreditation of Counseling and Related Educational Programs (CACREP). (2015). *2016 CACREP Standards*. Alexandria, VA: Author.

Council for the Accreditation of Counseling and Related Educational Programs (CACREP). (2018). CACREP vital statistics 2017: Results from a national survey of accredited programs. Alexandria, VA: Author.

Council for Accreditation of Counseling and Related Educational Programs [CACREP]. (2020). CACREP Response to COVID-19. https://www.cacrep.org/news/cacrep-statement-on-covid-19/

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334. https://doi.org/10.1007/BF02310555

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of consulting psychology*, *24*(4), 349-354. https://doi.org/10.1037/h0047358

Dawson, M., Phillips, B., & Leggat, S. (2013). Clinical supervision for allied health professionals: A systematic review. *Journal of Allied Health*, *42*(2), 65-73. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/23752232

DeVellis, R. F. (2017). *Scale development: Theory and applications* (5th ed). Los Angeles: Sage Publications.

Dziak, J. J., Coffman, D. L., Lanza, S. T., Li, R., & Jermiin, L. S. (2020). Sensitivity and specificity of information criteria. *Briefings in bioinformatics*, *21*(2), 553-565. https://doi.org/10.1093/bib/bbz016

Efstation, J. F., Patton, M. J., & Kardash, C. M. (1990). Measuring the working alliance in counselor supervision. *Journal of Counseling Psychology, 37*(3), 322–329. https://doi.org/10.1037/0022-0167.37.3.322

Ellis, M. V. (2017). Narratives of harmful clinical supervision. *The Clinical Supervisor*, *36*(1), 20-87. https://doi.org/10.1080/07325223.2017.1297752

Ellis, M. V., Berger, L., Hanus, A. E., Ayala, E. E., Swords, B. A., & Siembor, M. (2014). Inadequate and harmful clinical supervision: Testing a revised framework and assessing occurrence. *The Counseling Psychologist*, *42*(4), 434-472. https://doi.org/10.1177/0011000013508656

Ellis, M. V., Creaner, M., Hutman, H., & Timulak, L. (2015). A comparative study of clinical supervision in the Republic of Ireland and the United States. *Journal of Counseling Psychology*, *62*(4), 621-631. https://doi.org/10.1037/cou0000110

Ellis, M. V., D'Iuso, N., & Ladany, N. (2008). State of the art in the assessment, measurement, and evaluation of clinical supervision. In A. K. Hess, K. D. Hess, & T. H. Hess, (Eds.), *Psychotherapy supervision: Theory, research, and practice* , (2nd ed., pp. 473 – 499 ). New York: Wiley.

Ellis, M. V., Ladany, N., Krengel, M., & Schult, D. (1996). Clinical supervision research from 1981 to 1993: A methodological critique. *Journal of Counseling Psychology, 43*(1), 35 - 50. https://doi.org/10.1037/0022-0167.43.1.35

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment*, *7*(3), 286-299. https://doi.org/10.1037/1040-3590.7.3.286

Gonsalvez, C. J., & Calvert, F. L. (2014). Competency-based models of supervision: Principles and applications, promises and challenges. *Australian Psychologist*, *49*(4), 200-208. https://doi.org/10.1111/ap.12055

Gonsalvez, C. J., Hamid, G., Savage, N. M., & Livni, D. (2017). The supervision evaluation and supervisory competence scale: Psychometric validation. *Australian Psychologist*, *52*(2), 94-103. https://doi.org/10.1111/ap.12269

Gonsalvez, C. J., & McLeod, H. J. (2008). Toward the science-informed practice of clinical supervision: The Australian context. *Australian Psychologist*, *43*(2), 79-87. https://doi.org/10.1080/00050060802054869

Gonzalez, J., Barden, S. M., & Sharp, J. (2018). Multicultural competence and the working alliance as predictors of client outcomes. *The Professional Counselor*, *8*(4), 314-327. https://doi.org/10.15241/jg.8.4.314

Goodyear, R. K., Borders, L. D., Chang, C. Y., Guiffrida, D. A., Hutman, H., Kemer, G., ... &

White, E. (2016). Prioritizing questions and methods for an international and interdisciplinary supervision research agenda: Suggestions by eight scholars. *The Clinical Supervisor*, *35*(1), 117-154. https://doi.org/10.1080/07325223.2016.1153991

Gray, L. A., Ladany, N., Walker, J. A., & Ancis, J. R. (2001). Psychotherapy trainees' experience of counterproductive events in supervision. *Journal of Counseling Psychology*, *48*(4), 371-383. https://doi.org/10.1037/0022-0167.48.4.371

Hyrkäs, K., Appelqvist-Schmidlechner, K., & Paunonen-Ilmonen, M. (2003). Translating and validating the Finnish version of the Manchester Clinical Supervision Scale. *Scandinavian Journal of Caring Sciences*, *17*(4), 358-364. https://doi.org/10.1046/j.0283-9318.2003.00236.x

Kang, T., & Chen, T. T. (2007). An Investigation of the Performance of the Generalized SX 2 Item-Fit Index for Polytomous IRT Models. *ACT Research Report Series, 2007*-1. ACT, Inc.

Karpenko, V., & Gidycz, C. A. (2012). The supervisory relationship and the process of evaluation: Recommendations for supervisors. *The Clinical Supervisor*, *31*(2), 138-158. https://doi.org/10.1080/07325223.2013.730014

Kemer, G., Sunal, Z., Li, C., & Burgess, M. (2019). Beginning and expert supervisors' descriptions of effective and less effective supervision. *The Clinical Supervisor*, *38*(1), 116-134. https://doi.org/10.1080/07325223.2018.1514676

Kyriazos, T. A. (2018). Applied psychometrics: Sample size and sample power considerations in factor analysis (EFA, CFA) and SEM in general. *Psychology*, *9*(08), 2207-2230. https://doi.org/10.4236/psych.2018.98126

Lambie, G. W., Mullen, P. R., Swank, J. M., & Blount, A. (2018). The Counseling Competencies

Scale: Validation and Refinement. *Measurement and Evaluation in Counseling and Development, 51*(1), 1-15. https://doi.org/10.1080/07481756.2017.1358964

Lau, J., & Ng, K.-M. (2014). Conceptualizing the counseling training environment using Bronfenbrenner's ecological theory. *International Journal for the Advancement of Counselling, 36*(4), 423–439. http://doi.org/10.1007/s10447-014-9220-5

Lau, J. M., Ng, K. -M., & Vallett, D. B. (2019). The counseling training environment scale: Initial development and validity of a self-report measure to assess the counseling training environment. *Measurement and Evaluation in Counseling and Development*, *52*(4), 255-273. https://doi.org/10.1080/07481756.2019.1595813

Lehrman-Waterman, D., & Ladany, N. (2001). Development and validation of the evaluation process within supervision inventory. *Journal of Counseling Psychology, 48*(2), 168–177. https://doi.org/10.1037/0022-0167.48.2.168

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association, 83*, 1198–1202. https://doi.org/10.1080/01621459.1988.10478722

Loo, R., & Thorpe, K. (2000). Confirmatory factor analyses of the full and short versions of the Marlowe-Crowne Social Desirability Scale. *The Journal of Social Psychology*, *140*(5), 628-635. https://doi.org/10.1080/00224540009600503

Luke, M. (2019). Supervision in the counselor education context. In J.E.A Okech & D.J. Rubel (Eds.), *Counselor education in the 21st century.* (pp. 35-52). Alexandria, VA: American Counseling Association.

McCambridge, J., Witton, J., & Elbourne, D. R. (2014). Systematic review of the Hawthorne

effect: New concepts are needed to study research participation effects. *Journal of Clinical Epidemiology*, *67*(3), 267-277. https://doi.org/10.1016/j.jclinepi.2013.08.015

McGinty, E. E., Presskreischer, R., Han, H., & Barry, C. L. (2020). Psychological Distress and Loneliness Reported by US Adults in 2018 and April 2020. *JAMA*. https://doi.org/10.1001/jama.2020.9740

McGlashan, T. H., Carpenter Jr, W. T., & Bartko, J. J. (1988). Issues of design and methodology in long-term followup studies. *Schizophrenia Bulletin*, *14*(4), 569-574. https://doi.org/10.1093/schbul/14.4.569

Mertler, C. A., & Vannatta, R. A. (2017). *Advanced and multivariate statistical methods: Practical application and interpretation* (5th ed.). Glendale, CA: Pyrczak.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749. https://doi.org/10.1037/0003-066X.50.9.741

Milne, D. (2014). Beyond the “acid test”: A conceptual review and reformulation of outcome evaluation in clinical supervision. *American Journal of Psychotherapy*, *68*(2), 213-230. https://doi.org/10.1176/appi.psychotherapy.2014.68.2.213

Milne, D., & Reiser, R. P. (2012). A rationale for evidence-based clinical supervision. *Journal of Contemporary Psychotherapy, 42*(3). https://doi.org/10.1007/s10879-011-9199-8

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied psychological measurement, 16*(2), 159-176. https://doi.org/10.1002/j.2333-8504.1992.tb01436.x

Muraki, E. (1993). Information functions of the generalized partial credit model. *Applied*

*Psychological Measurement, 17*(4), 351-363. https://doi.org/10.1002/j.2333-8504.1993.tb01538.x

Mvududu, N. H., & Sink, C. A. (2013). Factor analysis in counseling research and practice. *Counseling Outcome Research and Evaluation*, *4*(2), 75-98. https://doi.org/10.1177/2150137813494766

National Board for Certified Counselors Foundation [NBCC] (2020). International Capacity Building. https://www.nbccf.org/programs/international

Nelson, M. L., & Friedlander, M. L. (2001). A close look at conflictual supervisory relationships: The trainee's perspective. *Journal of Counseling Psychology*, *48*(4), 384-395. https://doi.org/10.1037/0022-0167.48.4.384

Ng, K. M. (2012). Internationalization of the counseling profession and international counseling students. *International Journal for the Advancement of Counselling*, *34*(1), 1-4. https://doi.org/10.1007/s10447-012-9147-7

Ng, K. M., Choudhuri, D. D., Noonan, B. M., & Ceballos, P. (2012). An internationalization competency checklist for American counseling training programs. *International Journal for the Advancement of Counselling*, *34*(1), 19-38. https://doi.org/10.1007/s10447-011-9141-5

Olds, K., & Hawkins, R. (2014). Precursors to measuring outcomes in clinical supervision: A thematic analysis. *Training and Education in Professional Psychology*, *8*(3), 158. https://doi.org/10.1037/tep0000034

Palomo, M., Beinart, H., & Cooper, M. J. (2010). Development and validation of the Supervisory Relationship Questionnaire (SRQ) in UK trainee clinical psychologists. *British Journal of Clinical Psychology*, *49*(2), 131-149. https://doi.org/10.1348/014466509X441033

Proctor, B. (2011). Training for the supervision alliance. In J. R. Cutcliffe, K. Hyrkas, & J. Fowler (Eds.), *Routledge Handbook of Clinical Supervision: Fundamental International Themes*. London: Routledge.

R Core Team. (2019). *R: A language and Environment for Statistical Computing.* R Foundation for Statistical Computing. https://www.R-project.org

Ramsey, C. A., & Hewitt, A. D. (2005). A methodology for assessing sample representativeness. *Environmental Forensics*, *6*(1), 71-75. https://doi.org/10.1080/15275920590913877

Revelle, W. (2019). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois. R package version 1.9.12

Reynolds, W. M. (1982). Development of reliable and valid short forms of the Marlowe-Crowne Social Desirability Scale. *Journal of clinical psychology*, *38*(1), 119-125. https://doi.org/10.1002/1097-4679(198201)38:1<119::AID-JCLP2270380118>3.0.CO;2-I

Ridley, C. R., Tracy, M. L., Pruitt-Stephens, L., Wimsatt, M. K., & Beard, J. (2008). *Multicultural assessment validity: The preeminent ethical issue in psychological assessment.* In L. A. Suzuki & J. G. Ponterotto (Eds.), *Handbook of multicultural assessment: Clinical, psychological, and educational applications* (p. 22–33). Jossey-Bass.

Rizopoulos, D. (2006). ltm: An R package for Latent Variable Modelling and Item Response Theory Analyses. *Journal of Statistical Software, 17*(5), 1-25. http://www.jstatsoft.org/v17/i05/

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art.

*Psychological Methods*, *7*(2), 147-177. https://doi.org/10.1037/1082-989X.7.2.147

Schmitt, T. A. (2011). Current methodological considerations in exploratory and confirmatory factor analysis. *Journal of Psychoeducational Assessment*, *29*(4), 304-321. https://doi.org/10.1177/0734282911406653

Schutt, M. A. (2012). Replication and extension of Ellis, Ladany, Krengel, and Shult (1996); Clinical supervision and research from 1981 to 1993: A methodological critique. (Doctoral dissertation). Retrieved from https://preserve.lehigh.edu/cgi/viewcontent.cgi?article=2297&context=etd

Sharma, S., Mukherjee, S., Kumar, A., & Dillon, W. R. (2005). A simulation study to investigate the use of cutoff values for assessing model fit in covariance structure models. *Journal of Business Research*, *58*(7), 935-943. https://doi.org/10.1016/j.jbusres.2003.10.007

Simpson-Southward, C., Waller, G., & Hardy, G. E. (2017). How do we know what makes for "best practice" in clinical supervision for psychological therapists? A content analysis of supervisory models and approaches. *Clinical Psychology & Psychotherapy*, *24*(6), 1228-1245. https://doi.org/10.1002/cpp.2084

Skovholt, T. M., & Ronnestad, M. H. (1992). Themes in therapist and counselor development. *Journal of Counseling & Development*, *70*(4), 505-515. https://doi.org/10.1002/j.1556-6676.1992.tb01646.x

Snowdon, D. A., Leggat, S. G., Harding, K. E., Boyd, J., Scroggie, G., & Taylor, N. F. (2018). The association between effectiveness of clinical supervision of allied health professionals and improvement in patient function in an inpatient rehabilitation setting. *Disability and rehabilitation*, 1-10. https://doi.org/10.1080/09638288.2018.1518493

Snowdon, D. A., Millard, G., & Taylor, N. F. (2016). Effectiveness of clinical supervision of

allied health professionals: A survey. *Journal of Allied Health*, *45*(2), 113-121. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/27262469

Spence, S. H., Wilson, J., Kavanagh, D., Strong, J., & Worrall, L. (2001). Clinical supervision in four mental health professions: A review of the evidence. *Behaviour Change, 18*(3), 135–155. http://doi.org/10.1375/bech.18.3.135

Tabachnick, B., & Fidell, L. (2019). *Using multivariate statistics* (7th ed.). Boston, MA: Pearson.

Tew-Washburn, S. (2016). The CACREP/CORE Merger: Implications for the Rehabilitation Counseling Identity. *Alabama Counseling Association Journal*, *41*(1).

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4-70. https://doi.org/10.1177/109442810031002

Vicente, P., & Reis, E. (2007, October). Methodological issues in online surveys. In *Proceedings of the IADIS International Conference on WWW/Internet* (Vol. 2, pp. 173-176).

Walker, S. L., & Fraser, B. J. (2005). Development and validation of an instrument for assessing distance education learning environments in high education: The Distance Education Learning Environments Survey (DELES). *Learning Environments Research, 8*(3), 289–308. http://doi.org/10.1007/s10984-005-1568-3

Watkins, C. E. (2012). Psychotherapy supervision in the new millennium: Competency-based, evidence-based, particularized, and energized. *Journal of Contemporary Psychotherapy*, *42*(3), 193-203. https://doi.org/10.1007/s10879-011-9202-4

Watkins Jr, C. E. (2014). The supervisory alliance: A half century of theory, practice, and

research in critical perspective. *American Journal of Psychotherapy*, *68*(1), 19-55. https://doi.org/10.1176/appi.psychotherapy.2014.68.1.19

Watkins Jr, C. E. (2017). Convergence in psychotherapy supervision: A common factors, common processes, common practices perspective. *Journal of Psychotherapy Integration*, *27*(2), 140-152. https://doi.org/10.1037/int0000040

Watkins, C. E., & Milne, D. L. (2014). Clinical supervision at the international crossroads: current status and future directions. In C.E. Watkins, Jr. & D. L. Milne (Eds.), *Wiley international handbook of clinical supervision* (pp. 673-696). Oxford, United Kingdom: Wiley.

Weijters, B., Baumgartner, H. & Schillewaet, N. (2013). Reversed Item Bias: An Integrative Model. *Psychological Methods, 18*(3), 320-334. https://doi.org/10.1037/a0032121

Wheeler, C. E., & Barkham, D. L. (2014). A Core Evaluation Battery for Supervision. In C.E. Watkins, Jr. & D.L. Milne (Eds.), *Wiley international handbook of clinical supervision* (pp. 367-385). Oxford, United Kingdom: Wiley.

White, E. (2018). Measuring clinical supervision: How beneficial is yours and how do you know? *Journal of advanced nursing*, *74*(7), 1437-1439. https://doi.org/10.1111/jan.13529

White, E., & Winstanley, J. (2010). A randomised controlled trial of clinical supervision: Selected findings from a novel Australian attempt to establish the evidence base for causal relationships with quality of care and patient outcomes, as an informed contribution to mental health nursing practice development. *Journal of Research in Nursing*, *15*(2), 151-167. https://doi.org/10.1177/1744987109357816

White, E., & Winstanley, J. (2014). Clinical supervision and the helping professions: An

interpretation of history. *The Clinical Supervisor*, *33*(1), 3-25. https://doi.org/10.1080/07325223.2014.905226

Winstanley, J. (2000). Manchester clinical supervision scale. *Nursing Standard*, *14*(19), 31-32. https://doi.org/10.7748/ns.14.19.31.s54

Winstanley, J., & White, E. (2011). The MCSS-26: revision of the Manchester Clinical Supervision Scale using the Rasch Measurement Model. *Journal of Nursing Measurement*, *19*(3), 160-178. http://dx.doi.org/10.1891/1061-3749.19.3.160

Winstanley, J., & White, E. (2014). The Manchester Clinical Supervision Scale: MCSS-26. In C.E. Watkins, Jr. & D.L. Milne (Eds.), *Wiley international handbook of clinical supervision* (pp. 386-401). Oxford, United Kingdom: Wiley.

Winstanley, J. & White, E. (2019). MCSS-26 User Manual © Version 6.0. White Winstanley Ltd, Cheshire, United Kingdom.

Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist, 34*(6), 806-838. https://doi.org/10.1177/0011000006288127

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square-fit values. *Rasch Measurement Transactions, 8,* 370.

## Abstract

Supervision researchers have called for the development of a common measurement approach within supervision instrument development. In order to assess supervisor competency, scholars require instruments that can precisely measure the competence of the supervisor. In order to address the ongoing need for psychometrically robust supervision instruments that assess supervisor competence, we designed a cross-validation study of the Supervision Evaluation and Supervisor Competence (SE-SC; Gonsalvez et al., 2017) scale. The psychometric properties of the instrument and its items was examined using a polytomous response model, the Generalized Partial Credit Model (GPCM). Participants ($n$ = 86) were counselors-in-training at CACREP-accredited institutions in the United States who were currently engaged in supervision at the time of web-based survey. Data from this sample indicated acceptable instrument-level validity and reliability psychometrics. However, item-response analysis yielded many items that did not fit within the GPCM, indicating the need for a revised instrument. Based on these results, we proposed a revised 15-item version of the SE-SC. Our findings suggest the need for more testing and development of the SE-SC if it were to be employed with CIT.


*Keywords:* supervisor competence, SE-SC, supervision instruments, psychometric evaluation, item-response theory

## Chapter 3: Measuring Supervisor Competence with Counselors-in-Training

The utility of any quantitative scale that seeks to measure certain human functioning relies on the evidence supportive of its use with a particular population in a particular setting. This requirement applies to quantitative supervision research as well. However, in recent years, supervision scholars and researchers have routinely expressed concern about the quantitative methodological rigor of supervision research (Wheeler & Richards, 2007), specifically the lack of consistently utilized and psychometrically robust instruments to facilitate investigation and evaluation of supervision (Dawson et al., 2013; Olds & Hawkins, 2014; Schutt, 2012; Watkins, 2012b, 2018). Scholars (Bernard & Goodyear, 2014; Wheeler & Barkham, 2014) attribute this issue to the wide use of one-time use instruments developed by researchers for their specific purposes and identify this as an important methodological weakness. Further, replication studies of these instruments across populations are lacking, thereby inhibiting population comparison research and resulting in no further evidence to suggest the continued use of these instruments given the many contextual differences in supervision across the globe (Wheeler & Barkham, 2014).

Hence, there is a need for replication research to examine supervision evaluation instruments in order to advance the development of a "cumulative and coherent knowledge base [of supervision]" (Wheeler & Barkham, 2014, p. 380) that would allow for robust research alongside deployment in practice settings (Gonsalvez & Calvert, 2014). The present study represents an attempt to address the above-identified concerns. Specifically, we sought to systematically validate the Supervision Evaluation and Supervisor Competency Scale (SE-SC; Gonsalvez et al., 2017) for use with master's-level counselors-in-training (CIT) in the United

States from programs accredited by the Council for Accreditation of Counseling and Related Educational Programs (CACREP).

**Supervisor Competence**

As supervision is a distinct professional service that requires competency based on education, training, and experience (Falender, 2014), many practical permutations of supervision preparation exist across countries and disciplines (Watkins, 2012a). Regardless of work setting, discipline, or country, existing broad agreement suggests that supervisors steward the development of supervisees, ensure public welfare, and, ultimately, serve as gatekeepers to the profession (Bernard & Goodyear, 2014; Rønnestad et al., 2019). Thus, assessing supervisors' competence is critical in order to account for (a) suitability for the profession; (b) ability to integrate knowledge, skills, and attitudes (Rubin et al., 2007); and (c) ability to serve as a developmental, ethical, and supportive role model for supervisees (Allan et al., 2016).

Rubin et al. (2007) define competencies as elements of competence that "... involved the whole person and are teachable, observable, measurable, containable, practical, derived by experts, flexible and transferable across settings, and continually reevaluated and redefined" (p. 454). Key to note is that competence is not the idealized standard (Gonsalvez & Calvert, 2014). It is the "minimum acceptable standard for independent practice" (p. 204) and is continually developed with practice over time. Supervision competencies serve as a meaningful trans-theoretical measuring stick to which evaluation of supervisor performance may be compared. They form a foundation of evidence-based supervision practice (EBSP) that aims to enhance supervisee progress while simultaneously enhancing client care (Watkins, 2012b). EBSP complements evidenced-based clinical services (EBCS) insofar as the main objectives of

improving client care, safeguarding client welfare, and delivering effective services overlap (Milne & Reiser, 2012).

Evaluation of supervisor performance, then, is one among many ingredients that serves as an opportunity for feedback, accountability, and reflection within the supervision relationship for both the supervisor and supervisee (Borders, 2014). In their systematic literature review of evidence-based supervisor training, Milne et al. (2011) observed that feedback was the most frequent activity in supervisor education and development. In the absence of regular, meaningful feedback, supervisors and supervisors of supervisors risk drifting from quality and focused supervision. Further, evaluation of supervisor competence bears great significance for counselor education programs because supervisors play a critical role in these programs (Luke, 2019).

Supervisors working in counselor education programs require feedback in order to improve and extend their supervision skills. However, barriers to meaningful feedback and supervision evaluation abound in training programs. Gonsalvez and McLeod (2008) suggest that the power difference in the supervisory relationship "makes it likely that such feedback is systematically biased" (p. 84). Without systematic and formal evaluations of supervision, Gonsalvez and McLeod (2008) note that "supervisors can continue to provide supervision for many years without receiving an objective and fair appraisal" (p. 84). Additionally, as is sometimes the case, counselor education doctoral students (CEDS) and community-based professional counselors who do not have in-depth training in supervision may serve as supervisors to CITs (Luke, 2019). Even new counselor educators "lack a depth and breadth to their supervisory expertise" (Luke, 2019, p. 44) and would benefit from regular feedback on their work. Thus, monitoring supervisor competence represents an important training environment

quality issue, in addition to a professional gatekeeping ethical responsibility (American Counseling Association, 2014).

To do so, training program administrators would require time-saving assessment tools that are theoretically sound and empirically based to assess the quality and competence of supervision staff. Competency-based approaches are presumed to increase transparency, objectivity, and ecological validity of evaluation processes (Gonsalvez & Calvert, 2014). The SE-SC (Gonsalvez et al., 2017), developed based on the Proctor model of supervision (Proctor, 2011), appears to offer a promising measure for use to assess supervisor competence. However, there is yet any reported psychometric properties based on North American CITs to support its research and practice utilities in the American counseling training setting.

**Supervision Evaluation and Supervisory Competence Scale**

Influential in its parsimony and operationalization of supervisory tasks, the Proctor Model (Proctor, 2011) articulates "complementary and sometimes contradictory tasks" (Spence et al., 2001, p. 25). These tasks comprise three domains that are crucial for supervision to be effective: *restorative*, *formative*, and *normative.* These tasks are considered essential competencies for supervisors to demonstrate in order for supervision to be beneficial to the CITs. The restorative domain of supervision primarily concerns with the wellbeing, resilience, and self-awareness of CITs as they navigate the "emotional burden of practice" (Snowdon et al., 2016, p. 114). The formative domain of supervision, focused on fostering self-reflection and learning through experience, attends to CITs' development and maintenance of high-quality care. Lastly, the normative domain concerns the key professional standards, legal and role responsibilities, and ethical concerns that arise for the CIT in supervision. With considerable theoretical and empirical support backing its use for practice and research (Dawson et al., 2013; Kilminster &

Jolly, 2010), the Proctor Model provides a clear, parsimonious model to assess supervisor competence across the restorative, formative, and normative domains.

Gonsalvez et al. (2017) conceptualize the SE-SC based on the Proctor Model. The SE-SC is completed by supervisees about their experience of their supervisors' competence in supervision. Initially developed and validated, the 31-item SE-SC assesses supervision effectiveness (3 items), supervision satisfaction (3 items), and specific supervisor competencies (25 items). Supervisor competence are assessed across six subscales based on the three constructs of the Proctor Model: *restorative* (A1 openness, caring, and support; A5 restorative), *formative* (A6 reflective practitioner competencies, reflection), and *normative* (A2 supervisor knowledge and expertise as therapist; A3 supervision planning and management; A4 goal-directed supervision). Responses to questions are framed with a 7-point Likert-type scale, ranging from 1 (*not at all, strongly disagree*) to 7 (v*ery much so, strongly agree*) and scored with values of 1-7, with higher values interpreted as better outcomes or competence present.

The scale produces two sets of scores: (a) six subscale scores (openness, caring and support, *n items* = 5; supervisor knowledge and expertise as therapist, *n* = 2; supervision planning and management, *n* = 4; goal-directed supervision, *n* = 2; restorative competencies, *n* = 3; reflective practitioner competencies, *n* = 6) and (b) an overall score, comprising the mean of the supervision satisfaction and effectiveness subscales. Each subscale score is determined by taking the mean of all the individual item scores. The overall score, comprised of six items, is determined by the average of three items assessing supervision effectiveness and three items assessing supervision satisfaction. The six subscales each contain a set of items to assess specific competencies. The original scale developers suggest a score of 6 or above in a subscale "as a measure of supervisory excellence" (Craig Gonsalvez, personal communication, June 13, 2019).

The SE-SC was originally validated for use with 142 doctoral and master's students in psychology in Australia. Evidence from a cluster analysis suggests evidence of score validity for the SE-SC. Using a hierarchical clustering statistical technique, the scale developers used rescaled distance (RD) units to determine an a priori cluster structure of the 22 competency items (Gonsalvez et al., 2017). The resulting dendogram articulated three clusters of items based on the tightness of association. The A-cluster's (6 subscales) reliability coefficients were reported to range from .75 - .92 (Gonsalvez et al., 2017). The B-cluster's, a relaxed RD parameter resulting in 3 subscales, test-retest reliability was reported to range from .81- .93 (Gonsalvez et al., 2017). Test-retest reliability, however, was only determined with a subset ($n$ = 20) of the sample. The SE-SC subscale A1 (Openness, Caring, and Support) possesses good convergent validity with another supervisory alliance measure (SWAI-Rapport, $r$ = 0.82; Efstation et al., 1990). The final version of the SE-SC consists of 26 items: 4 items assessing overall effectiveness and 22 competency items across three subscales.

Given the recent development of the SE-SC, few research publications exist that demonstrate its utility in the field. However, as a relatively new instrument that is specifically derived from supervision practice and competency literature, the SE-SC presents one potential tool in the design and evaluation of competent and evidence-based counseling supervision services.

## Purpose of Study

Gonsalvez et al. (2017) called for further examination of the SE-SC's psychometric properties and replication across different samples. This is particularly critical because the scale was developed based on 142 psychology graduate students in Australian and its item-level performance has not been examined. Thus, this research seeks to investigate its psychometric

properties in a representative sample of CITs in the United States. Findings may potentially extend the utility of the measure beyond its initial population and setting. Given the relevance of the scale to clinical contexts and supervision settings, CITs working towards their degrees in clinical mental health counseling, rehabilitation counseling, addictions counseling, and family counseling were the target population in this psychometric validation study.

"Is the SE-SC reliable and valid for use with CITs in the United States?" forms the main research question. Specifically, our study addresses the following questions:

1. Does the SE-SC possess internal consistency?
2. Does the SE-SC possess item-level fitness?
3. When compared with a measure of supervisory relationship (SWAI-T), does the SE-SC possess concurrent validity?
4. Does social desirability present a significant threat to validity?

As supervisees find themselves in power-under positions within the supervision relationship and are subject to evaluation apprehension (Copeland et al., 2011; Ellis et al., 2008), we consider it critical to explore the possibility of supervisee-completed measures as ample opportunities for socially desirable responding. Thus, we included Question 4 above to examine how items of the SE-SC would correlate with a measure of social desirability.

We hypothesized that item-level performance of the SE-SC would fit a polytomous Rasch model based on a sample of CIT from the United States and there would be acceptable evidence of internal reliability and concurrent validity support the utility of the measure among U.S. CIT. The evaluation of supervision, in particular supervisor competency, and the development of robust, psychometrically valid instruments remain critical areas for future

research. We hope this study informs future supervision competency research by exploring the psychometric properties of the SE-SC.

## Method

A cross-validation study was determined as best suited to meaningfully address the research questions. In the present cross-validation study, we assessed the psychometric properties of the SE-SC using Rasch modeling analytics. Using a one-time sampling strategy of a large number of participants, this study focuses on assessing multiple psychometric properties.

### Participants

In order to be eligible for participation in the study, participants were required to satisfy all inclusion criteria. Inclusion criteria included voluntary adults, age 18 or older, who self-identified as a master's-level counselor-in-training enrolled at a CACREP-accredited program in either clinical mental health counseling, clinical rehabilitation counseling, addictions counseling, or marriage and family counseling. Participants were also required to be currently engaged in clinical supervision as part of their clinical training experience.

We collected a total of 135 participant responses. Of this total, 86 participant responses were used in final analysis. Responses removed from analysis were due to either participant (a) ineligibility to participate due to self-identified criterion or (b) incomplete completion of the study survey. Participants were asked to respond to each survey question within the informed consent. Participants were allowed, however, to not respond to specific demographic questions to protect anonymity – an intentional design choice in order invite full disclosure on instruments of inquiry. For full sample demographics, see Appendix C. With respect to gender (n = 81, 94% reporting), the sample consisted of 71(88%) females, 8 (10%) males, and 2 (2%) non-binary persons. With respect to race (*n* = 80, 93% reporting), participants identified as

Caucasian/European/White (*n* = 67, 83%), Asian (*n* = 4, 5%), Black/African (*n* = 3, 4%), Latinx/Hispanic/Spanish (*n* = 3, 4%), and Multiracial (*n* = 3, 4%) in the sample.

The annual report (CACREP, 2018) categories were utilized to compare sample representativeness to the 2017 annual report for master's-level counselors-in-training. Between the sample and the population, a few differences are important to note. The current sample had an increase in female representation (88% versus 83% in CACREP annual report; CACREP, 2018), Caucasian/White representation (83% versus 60% in CACREP annual report; CACREP, 2018), and multiracial identity representation (4% versus 2% in CACREP annual report; CACREP, 2018). This sample had a marked decrease in Black/African representation (4% versus 19% in CACREP annual report; CACREP, 2018) and Latinx/Hispanic/Spanish representation (4% versus 8% in CACREP annual report; CACREP, 2018). Males were also underrepresented compared to the annual report (10% versus 17% in CACREP annual report; CACREP, 2018). Data on American Indian/Native American, Native Hawaiian/Pacific Islander, Non-resident Alien, and "Other" racial groups are presented in the annual report but none of these groups were represented in the current sample.

Data on sample sexual minority status (*n* = 80, 93% reporting), international student status (*n* = 81, 94% reporting), and age (*n* = 76, 88% reporting) were also collected. With respect to sexual minority status, 16 (20%) of participants identified as a sexual minority. With respect to international student status, 5 (6%) participants identified as an international student. Sample ages ranged from 22 to 69 (*M* = 30, *SD* = 8), with 22-29 (*n* = 53, 70%) comprising the largest group, followed by 30-39 (*n* = 16, 21%), 40-49 (*n* = 3, 4%), 50-59 (*n* = 3, 4%), and 60-69 (*n* = 1, 1%). Additional sample characteristics are presented herein to further any generalizability conclusions drawn from this study.

### *Program Region, Specialization, and Delivery Method*

Participants reported enrollment in programs across the country. Program specializations and delivery methods varied within the sample. Not all participants reported their program's region ($n$ = 81, 94%), but of those who did NCACES ($n$ = 27, 33%), SACES ($n$ = 25, 31%), NARACES ($n$ = 16, 20%), WACES ($n$ = 8, 10%), and RMACES ($n$ = 5, 6%) were all represented. Across program specialty types, participants reported ($n$ = 86) enrollment in clinical mental health counseling ($n$ = 71, 83%), rehabilitation counseling ($n$ = 11, 13%), and marriage and family counseling ($n$ = 4, 5%). All participants reported their program delivery methods ($n$ = 86). Traditional ($n$ = 59, 69%), hybrid ($n$ = 15, 17%), and online ($n$ = 12, 14%) program delivery methods were represented in the sample.

### *Accrued Clinical Hours*

Participants varied in reported practicum and internship hours accrued. Some participants reported ($n$ = 77, 90%) their combined hours ($M$ = 372, $SD$ = 329): 18 (23%) reported less than 100 hours, 17 (22%) reported between 101-200 hours, 4 (5%) reported between 201-300 hours, 3 (4%) reported between 301-400 hours, and 35 (45%) reported more than 400 hours accrued during training.

### *Supervision Setting and Supervisor Type*

Participants identified the setting where they received supervision and the supervisor's relationship to their graduate counseling program. Every participant reported the setting of supervision ($n$ = 86, 100%): university clinic ($n$ = 29, 34%), agency or community mental health center ($n$ = 31, 36%), private practice ($n$ = 15, 17%), group practice ($n$ = 4, 5%), and via telesupervision ($n$ = 7, 8%).

Most participants (*n* = 85, 99%) reported their supervisor's relationship with the counseling program. Participants reported supervisors were current faculty (*n* = 15, 18%), site supervisors (*n* = 60, 71%), doctoral students (*n* = 8), or other (*n* = 2, 2%). Participants identified "other" and communicated they had multiple supervisors across their site and program settings.

### *Supervision Frequency, Duration, and Modality*

Participants also reported the frequency, duration, and modality of supervision. Every participant reported the frequency and duration of supervision sessions. Participants reported weekly supervision (*n* = 78, 91%), biweekly/every two weeks (*n* = 7, 8%), and less than once every three months (*n* = 1, 1%). A majority of participants (*n* = 49, 57%) reported an average supervision session lasting for 46-60 minutes, with 21 (24%) participants reporting supervision sessions lasting for longer than 60 minutes, 12 (14%) reporting 31-45 minutes, 3 (3%) reporting 15-30 minutes, and 1 (1%) reporting less than 15 minutes. Not all participants reported supervision modality (*n* = 84). Individual supervision was most common (*n* = 52, 62%), with a mix of individual and group supervision next most common (*n* = 22, 26%), followed by triadic supervision (*n* = 6, 7%), and group supervision (*n* = 4, 5%).

### *Supervisee and Supervisor Theoretical Orientation*

All participants (*n* = 86) reported their theoretical orientation and their supervisor's theoretical orientation. Survey permissions were set so that participants could select theoretical orientations, so *n* counts add up to over 86. From most to least common within the sample for participant theoretical orientation: cognitive-behavioral (*n* = 42, 49%), humanistic (*n* = 41, 48%), eclectic (*n* = 26, 30%), interpersonal (*n* = 23, 27%), systems (*n* = 13, 15%), psychodynamic (*n* = 12, 14%), reality/choice theory (*n* = 4, 5%), dialectical-behavioral (*n* = 3, 3%), existential (*n* = 2, 2%), and feminist (*n* = 2, 2%). Other orientations included "Adlerian" (*n* = 1), "attachment" (*n* =

1), “eye-movement desensitization and reprocessing (EMDR)” (*n* = 1), “somatic experiencing” (*n* = 1), “trauma” (*n* = 1), “social constructivism” (*n* = 1), and “not determined” (*n* = 1) for participants.

Supervisor theoretical orientation, from most to least common, included: cognitive-behavioral (*n* = 43, 50%), humanistic (*n* = 30, 35%), eclectic (*n* = 20, 23%), systems (*n* = 19, 22%), interpersonal (*n* = 14, 16%), psychodynamic (*n* = 11, 13%), dialectical-behavioral (*n* = 3, 3%), and Gestalt (*n* = 3, 3%). Other supervisor theoretical orientations included “Adlerian” (*n* = 1), “attachment” (*n* = 1), “brief solution focused” (*n* = 1), and “somatic experiencing” (*n* = 1).

**Procedure**

Prior to participant recruitment, the study was approved by the university’s Institutional Review Board. As the focus of this study was instrument validation on a sample of CITs, deliberate efforts were made to invite participation based on enrollment in a counseling program. A sample was gathered from counseling professionals who self-identified themselves as master’s-level CITs currently receiving clinical supervision.

At the time of writing this procedure, November 2019, 880 counseling programs tracks were accredited by CACREP per the online directory of programs. A select number of these programs (*n* = 505) offer degree-specialties in clinical, rehabilitation, family, or addictions counseling. For recruitment purposes, we created a database of contact information for every program’s liaisons and faculty members. We sent only one email to a program, even if that program contains multiple degrees with specialties. For example, if a program offers two master’s degrees, one in clinical mental health and one in addiction, we only sent one mail to that program. We sent an email to the liaison for each program listed in the database with program

faculty receiving the same message via carbon copy. We also sent similar recruitment email to our personal contacts, counselor educators who had direct contact with their CITs.

The email to program liaisons and faculty contained the scope of the study, research questions, and the informed consent for potential participants. We invited them to share the email with their currently enrolled counselors-in-training. As an incentive to participants, they were invited to enter a drawing for 1 of 8 $20 Starbucks gift cards was offered. Participation was voluntary and anonymous. After two weeks from the original email, liaisons and program faculty were sent a follow-up reminder and a thank you note. To increase the sample size, we encouraged participants to forward the message to potential participants. If participants chose to enter for the gift card drawing, they entered their email address in a different survey. Responses to the research survey were not matched to their email addresses.

The secure web-based survey platform to which potential participants used encryption to protect participants' identifying information. Participants had access to a description of the purpose of this study, participant selection criteria, procedures, consent information and documents, the survey questionnaires (i.e., demographic questionnaire, SE-SC, SWAI-T), and a reminder of their rights as a volunteer participant. Prior to taking the online survey, participants were asked to review the consent information provided, indicate their agreement to participate, and complete the survey. All supervision instruments included in the survey questionnaire were authorized for use by instrument creators and developers.

The survey was available online for nine weeks. The first author sent recruitment emails at 3-week intervals (0, 3, 6 weeks). Three weeks after the final reminder, we closed the survey. The survey took approximately 10-15 minutes. Per consultation of the literature, surveys should not exceed 20 minutes otherwise survey completion rates drop (Revilla & Ochoa, 2017).

## Measures

### *Demographic Questionnaire*

The questionnaire gathered participant's self-report data including gender, age, race, stage of training, counseling specialization, time in program, theoretical orientation (self and supervisor), supervision context, supervision relationship duration, frequency of supervision meetings, and duration of supervision meetings. Such personal characteristics may serve as variables of possible outcomes (Bambling, 2014) and are consistent with supervision research (Ladany & Muse-Burke, 2001; Lambie et al., 2018). Of importance, previous psychometric validation efforts of supervision instruments (e.g., Lehrman-Waterman & Ladany, 2001; Palomo et al., 2010).

### *Supervision Evaluation and Supervisory Competence Scale*

The Supervision Evaluation and Supervisory Competence Scale (SE-SC; Gonsalvez et al., 2017) was discussed in-depth in a previous section of this article.

### *Supervision Working Alliance Inventory-Trainee Version*

The 19-item Supervision Working Alliance Inventory-Trainee Version (SWAI-T; Efstation et al., 1990) measures the quality of the supervisory relationship from the perspective of the supervisee. All items load across two subscales: (a) *Rapport* with the supervisor and (b) *Client-focused* nature of supervision sessions. The SWAI-T utilizes a 7-point Likert-type scale, with responses ranging from 1 (*Almost Never*) to 7 (*Almost Always*). Each subscale maintains adequate reliability (Rapport, $\alpha = .90$; Client focus, $\alpha = .77$; Efstation et al., 1990). Scale developers utilized the Supervisory Styles Inventory (SSI, Friedlander & Ward, 1984) to provide evidence of convergent validity with the Rapport subscale (Attractive, $r = .78$, $p \leq .001$) and the Client-focus subscale (Task-oriented subscale, $r = .52$, $p \leq .001$). As the SWAI-T has been

utilized frequently within supervision research (Watkins, 2014) and was utilized by authors in the SE-SC's development, the SWAI-T provides a useful tool to evaluate convergent validity. Cronbach's alpha for the current study was .96.

### *Marlowe-Crowne Social Desirability Scale Short*

In their systematic evaluation of multiple short versions of the original Marlowe-Crowne Social Desirability Scale (MCSDS; Crowne & Marlowe, 1960), Loo and Thorpe (2000) indicated support for Reynolds' (1982) short version of the MCSD (Forms A and B). In this study, we utilized Form A of the Marlowe-Crowne Social Desirability Scale Short (MCSDSS-A) per the scrutiny and evidence considered in Loo and Thorpe's (2000) analysis and shorter parsimony of the scale. Form A of the MCSDSS is an 11-item scale in which participants indicated "True/False" in response to the statement. For example, "No matter who I'm talking to, I'm always a good listener." After summing scores according to developer guidelines, higher scores indicate evidence of socially desirable responding. Reynolds (1982) reported Kuder-Richardson reliability (KR[20] = .74) and high correlation with the original MCSD ($r$ = .91, $p$ < .001). Cronbach's alpha for the MCSDSS-A, as presented in Loo and Thorpe (2000), was .59. For the current study, the MCSDSS-A had a KR[20] of .72 and a Cronbach's alpha of .72.

## Data Preparation Plan

Multiple data issues require attention before analysis including missing data, data accuracy, satisfaction of statistical assumptions, and managing outliers (Tabachnick & Fidell, 2019). Mertler and Vannata (2017) note that data accuracy is critical to ensure the integrity of the conclusions drawn from the data analysis.

### *Addressing Accurate Data*

Instruments will be completed digitally, and not by hand, thus contributing to the accuracy of participant-based data. While this might eliminate the possibility of researchers mis-entering data into a digitized data file, further steps were taken to account for possible participant entry errors, such as frequent outliers. Descriptive statistics will be evaluated for coherency and plausibility.

### *Addressing Missing Data*

Missing data is a rampant issue in data analysis (Tabachnick & Fidell, 2019). To address the possibility of missing data, participants interfaced with a survey that would not allow progression without responding to each question presented. Thus, accounting for missing data was not a critical step within data preparation.

### *Screen for Outliers*

Tabachnick and Fidell (2019) describe the standard deviation outlier labeling method, used here within this study, that assumes a normal distribution. Data more than three standard deviations from the mean was considered an outlier and eliminated from the reliability analysis. Items were not removed from item-response analysis as such responses are meaningful in determining model fitness.

### *Screen for Multicollinearity*

Multicollinearity correlations were identified and addressed according to theoretical assumptions. Multicollinearity occurs when high correlations ($r \geq .90$; Mertler & Vannata, 2017) occur between variables. For this study, in which simple exploratory factor analytic procedures are utilized, multicollinearities are not inherently troublesome, as they may be in regression

analysis, because of possible theoretical overlap, similarity, or "tapping" related constructs (Tabachnick & Fidell, 2019).

**Item-Level Analysis**

In examining item-level fitness, data were analyzed using item response theory for polytomous responses, specifically the Generalized Partial Credit Model (Muraki, 1992; 1993) and maximum likelihood estimations (MLE). MLE "maximizes the probability that this set of responses is observed" (Andrich & Marais, 2019, p. 113) per the GPCM. The Generalized Partial Credit Model (GPCM) was adopted due to its flexibility in initial analysis of polytomous item response structures and its lack of assumptions about item discrimination parameters. Assuming unknown intervals between response categories in model determination lends itself to analyzing the presumed latent performance of each item and its related measurement model (Muraki, 1992). Such a less constrained model thus results in estimates for each item that are a more accurate reflection of the data (Embretson & Reise, 2000). The GPCM, like other Rasch-based or derived models, runs on a logistic mathematical model of probability. Succinctly, item-level responsiveness is determined by the latent trait (ability or agreeableness), or theta, and the difficulty (threshold categories) of an item. The probability of any response being selected is a function of the trait's presence in the respondent. Items performing as expected within the response category system possess a sequential pattern across response thresholds (difficulty). Succinctly, Toland (2014) notes, "This means that each increasing category is more likely to be selected than previous response categories as one moves along the latent trait continuum" (p. 138). As a theta (latent ability) increases (e.g., *high agreeableness to the item*) so, too, does probability of selecting a sequentially higher category of responding (e.g., *strongly agree*).

Critically, assumptions of unidimensionality and threshold parameters were explored for each subscale. As each subscale of the SE-SC purports to tap a different construct, each is examined and reported according to the items within. To assess unidimensionality (Ziegler & Hagemann, 2015), exploratory factor analysis (EFA) was utilized to examine if subscale items were tapping the same single construct (Andrich & Marais, 2019; Baker, 2001; Toland, 2014). In order to satisfy unidimensionality within an EFA, loading of .40 or greater was considered acceptable. To assess threshold parameter, or structural assumptions of the model, difficulty thresholds were examined for sequential responding (e.g., $b_1 = -2$, $b_2 = 1$, $b_3 = 0$…). Items violating such difficulty threshold structures were considered in violation of the model. Items outfit and infit mean squared (outfit) were also calculated for model fit. Outfit stats outside of an acceptable range of 0.6-1.4 (Wright & Lincacre, 1994) were identified as misfitting. Outfit stats were used as the assessment statistic as outfit calculations tend to be more sensitive compared to infit stats, which tend to show less misfit. Further, in assessing polytomous model fit, item *p*-values of the S-$\chi^2$ and root means square error of approximation (RMSEA) were examined for significance. For the RMSEA, a gradation of fitness is articulated by Browne and Cudeck (1992), where values greater than 0.1 indicate a poor fit, values less than 0.08 indicate a reasonable fit, and values less than 0.05 indicate a close fit. The S-$\chi^2$ calculates the degree of similarity between observed and model-based (predicted) response frequencies per category (Kang & Chen, 2007). Mis-fitness for the S-$\chi^2$ is indicated by a statistically significant value ($p < .05$) and is sensitive to sample size. In sum, using the GPCM, the following parameters were explored for each item of the SE-SC: item location or difficulty (*b*), item discrimination (*a*), and error estimates.

## Results

Results are presented according to the order of the research questions. Calculations were executed within the R environment (R Core Team, 2019) version 1.9.12.31 with psych (Revelle, 2019), mirt (Chalmers, 2012), and ltm (Rizopoulos, 2006).

### Internal Consistency

Instrument reliability was assessed using Cronbach's $\alpha$ (Cronbach, 1951). An alpha level of 0.70 or greater was adopted as the acceptability level to determine reliability (Cortina, 1993). Reliability estimates calculated for the SE-SC ($\alpha$ = .97) and the subscales Normative ($\alpha$ = .94), Formative ($\alpha$ = .94), and Restorative ($\alpha$ = .92) yielded adequate consistency across the SE-SC.

### Item-Level Fitness Parameters

In order to meet model assumptions of unidimensionality and item independence, items were analyzed according to subscale of the SE-SC ($M$ = 150, $SD$ = 28): Normative ($M$ = 44, $SD$ = 10.5), Formative ($M$ = 35, $SD$ = 7.5), and Restorative ($M$ = 48.5, $SD$ = 8). Item trace lines of the SE-SC subscales are located in Appendix G. Appendix H presents the test information curves for each subscale.

**Table 2.1**
*SE-SC Item Parameters, Ranked by Item Difficulty*

| Subscale | Item No. | Item Difficulty (Response Categories) *b* | Item Discrimination *a* | MNSQ Outfit | MNSQ Infit |
|---|---|---|---|---|---|
| Restorative ($n$ = 8) | 7 | *-.35 (6)* | 1.715 | *0.441* | 0.752 |
| | 20 | -1.02 (7) | 1.836 | 0.804 | 0.985 |
| | 8 | *-1.06 (5)* | 5.62 | *0.495* | 0.703 |
| | 19 | *-1.1* | 2.158 | 0.677 | 1.033 |

| Subscale | Item No. | Item Difficulty (Response Categories) *b* | Item Discrimination *a* | MNSQ Outfit | MNSQ Infit |
|---|---|---|---|---|---|
| | | *(6)* | | | |
| | 9 | *-1.15* *(6)* | 2.394 | 0.774 | 1.044 |
| | 6 | *-1.21* *(5)* | 2.418 | 0.676 | 1.084 |
| | 18 | -1.42 (7) | 0.928 | 0.955 | 0.986 |
| | 5 | *-1.45* *(5)* | 1.715 | 0.868 | 0.981 |
| Formative (*n* = 6) | 24 | -0.98 (7) | 2.764 | 0.679 | 0.960 |
| | 26 | *-1.03* *(6)* | 3.323 | *0.585* | 0.908 |
| | 25 | -1.03 (7) | 2.031 | 0.789 | 0.905 |
| | 22 | *-1.04* *(6)* | 2.965 | 0.652 | 0.872 |
| | 23 | -1.22 (7) | 3.55 | 0.636 | 0.814 |
| | 21 | -1.36 (7) | 2.853 | 0.624 | 0.924 |
| Normative (*n* = 8) | 17 | -0.54 (7) | 1.972 | 0.804 | 0.921 |
| | 16 | *-0.58* *(6)* | 1.893 | 0.835 | 0.901 |
| | 15 | -0.71 (7) | 3.007 | 0.678 | 0.786 |
| | 14 | -0.84 (7) | 1.982 | 0.783 | 0.891 |
| | 12 | -1.02 (7) | 2.571 | 0.705 | 0.928 |

| Subscale | Item No. | Item Difficulty (Response Categories) *b* | Item Discrimination *a* | MNSQ Outfit | MNSQ Infit |
|---|---|---|---|---|---|
| | 13 | *-1.03 (6)* | 2.146 | 0.704 | 0.913 |
| | 11 | -1.27 (7) | 1.167 | 0.858 | 1.042 |
| | 10 | -1.3 (7) | 1.193 | 0.995 | 1.010 |

*Note.* MNSQ = mean square. Misfits are italicized if MNSQ Outfit < .4.

**Table 2.2**
*SE-SC Item Parameters, Response Thresholds*

| Subscale | Item No. | $b_1$ | $b_2$ | $b_3$ | $b_4$ | $b_5$ | $b_6$ |
|---|---|---|---|---|---|---|---|
| Restorative (*n* = 8) | 7 | -1.915 | -1.618 | -1.188 | -0.763 | -0.095 | NV |
| | 20 | -1.49 | *-1.995* | -0.943 | *-0.954* | -0.748 | -0.017 |
| | 8 | -1.765 | -1.136 | -0.952 | -0.376 | NV | NV |
| | 19 | -1.805 | -1.316 | *-1.36* | -0.822 | -0.213 | NV |
| | 9 | -1.295 | *-2.046* | *-1.345* | -0.824 | -0.226 | NV |
| | 6 | -2.344 | -1.123 | -0.927 | -0.435 | NV | NV |
| | 18 | -2.154 | -1.954 | -1.595 | -1.546 | -0.187 | *-1.093* |
| | 5 | -2.142 | -1.675 | -1.529 | -0.453 | NV | NV |
| Formative (*n* = 6) | 24 | -2.033 | -1.527 | -1.143 | -1.005 | -0.388 | 0.203 |
| | 26 | -2.086 | -1.369 | -1.093 | -0.528 | -0.085 | NV |
| | 25 | -1.967 | -1.868 | -0.521 | *-1.248* | -0.448 | -0.172 |
| | 22 | -2.136 | -1.247 | -1.017 | -0.676 | -0.107 | NV |
| | 23 | -2.767 | -1.587 | -1.398 | -1.15 | -0.558 | 0.166 |
| | 21 | -2.693 | -2.185 | -1.336 | -1.197 | -0.579 | -0.169 |
| Normative (*n* = 8) | 17 | -0.805 | *-1.678* | -0.482 | *-0.71* | -0.112 | *-0.537* |
| | 16 | -1.747 | -0.905 | -0.527 | 0.016 | 0.276 | NV |

| Subscale | Item No. | $b_1$ | $b_2$ | $b_3$ | $b_4$ | $b_5$ | $b_6$ |
|---|---|---|---|---|---|---|---|
| | 15 | -1.68 | -1.492 | -0.956 | -0.315 | *-0.518* | 0.682 |
| | 14 | -1.215 | *-1.588* | *-0.836* | -0.979 | -0.59 | 0.17 |
| | 12 | -1.822 | *-1.538* | -1.66 | -0.821 | -0.533 | 0.24 |
| | 13 | -1.887 | -1.531 | -1.324 | -0.475 | 0.076 | NV |
| | 11 | -1.982 | -1.572 | *-2.117* | *-0.64* | -0.905 | -0.414 |
| | 10 | -2.31 | -1.477 | -1.371 | -0.895 | *-1.44* | -0.278 |

*Note.* NV = no value.

***Normative***

The Normative subscale included Items 10, 11, 12, 13, 14, 15, 16, and 17 (Gonsalvez et al., 2017). Within the Normative subscale, all items were determined to be unidimensional with loading greater than the .40 in an exploratory factor analysis (loadings ranged from .68 - .90). Unidimensionality was also determined by reviewing the infit and outfit mean square statistic. As shown in Table 2, items with outfit stats outside of an acceptable range of 0.6-1.4 (Wright & Lincacre, 1994) were identified as misfitting. Item difficulty (*b*) was calculated using the average of the item thresholds (e.g., $b_1$, $b_2$…) (see Table 2.1) in order to ease comparability of difficulties. However, item difficulty thresholds were examined for sequential responding (e.g. $b_1$ = -2, $b_2$ = -1, $b_3$ = 0…) with nonconforming items identified in Table 2.2. Nonconforming items, that is items with response categories that did not fit the response category sequence (e.g., 1 = *Not at all, Strongly disagree*; 7 = *Very much so, Strongly agree*) included Items 17, 15, 14, 12, 11, 10. Underrepresented response categories for Items 16 and 13 were indicated according to the GPCM. Item discrimination (a) estimates are presented in Table 2. Of note, the spread of item discrimination tends to be limited given the response categories ranging from 1-7.

Error estimates were calculated using S-$\chi^2$ for polytomous Rasch models. Setting significance level for S-$\chi^2$ at 0.05 (Chon et al., 2010), such error estimates aid in determining conformity to the model by reviewing p-values. Items with p-values less than .05 assessed as misfitting, including Item 11 ($p$ = .015) and item 15 ($p$ = .12). RMSEA p-values within the Normative subscale resulted in Item 11 ($p$ = .123) and Item 15 ($p$ = .151) poorly fitting the model, while Items 12 ($p$ = .074), 13 ($p$ = .081), and 16 ($p$ = .056) reasonably fit the model, and Items 10 ($p$ = .00), 14 ($p$ = .00), and 17 ($p$ = .00) closely fit the model.

***Formative***

The Formative subscale included Items 21, 22, 23, 24, 25, and 26 (Gonsalvez et al., 2017). All items satisfied the assumption of unidimensionality per the exploratory factor analysis, with loadings greater than .40 and ranging from .83-.89. Item 25 was the only nonconforming item with respect to response structure categories. Underrepresented response categories for Items 26 and 22 were identified. Item 26 outfit estimate was outside the acceptable range for fitness (Table 2). Error estimates (S-$\chi^2$) for all items indicated conformity to the model. Goodness of fit (RMSEA) values for Items 21 ($p$ = .126) and 22 ($p$ = .119) indicated poor fit. All other items reasonably fit or closely fit the model.

***Restorative***

The Restorative subscale was determined as Items 5, 6, 7, 8, 9, 18, 19, and 20 (Gonsalvez et al., 2017). All items were determined to be unidimensional with exploratory factor analysis loadings ranging from .67-.93. All item loadings exceeded the .40 cutoff threshold.

Item response categories for Items 20, 19, 9, and 18 did not conform to the expected response model. Underrepresented response categories were identified for Items 7, 8, 19, 9, 6, and 5. Outfit estimates for Item 7 and Item 8 (Table 2) fell outside the acceptable range. Error

estimates (S-$\chi^2$) for Items 5 ($p$ = .028) and 9 ($p$ = .042) did not conform to the model. Goodness of fit (RMSEA) values for items 6 ($p$ = .00) indicated a close fit. Items 8 ($p$ = .084), 19 ($p$ = .062), and 20 ($p$ = .088) reasonably fit the model and Items 5 ($p$ = .142), 7 ($p$ = .101), 9 ($p$ = .132), 18 ($p$ = .101) poorly fit the model. The worst-fitting Restorative subscale items were Item 5 and Item 9.

**Revised Version of the SE-SC**

Based on the satisfaction of model assumptions and data fitness estimates (Tables 2.3, 2.4, 2.5), a revised version included 11 items. The 11-item revised version of the competency subscales of the SE-SC has a better fit compared to the original 22 item version. In examining the superiority of data fit to the GPCM, Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC) were estimated, in Table 2.6 and 2.7, with lower values indicating improved model fit (Dziak et al., 2020). Both AIC and BIC assume that the estimated value is the distance between the unknown true likelihood of the model and the fitted likelihood of the model; thus, the smaller the distance the "closer to the truth" of the model fit. The smaller AIC and BIC associated with the GPCM of the revised subscales, compared to the GPCM of the original subscales, support the conclusion that the revised subscales better fit the data than the original subscales of the SE-SC.

As the competency subscales were the primary focus of scrutiny, as was the case in Gonsalvez et al.'s (2017) work, the final revised version (Appendix I) of the SE-SC contains 15 items total: 11 competency items (across 3 subscales) and 4 "overall" items. Preliminary Pearson correlation between the revised subscale-only 11-item SE-SC and the original SE-SC was .98 ($p$ < .0001). Preliminary Pearson correlation between the revised 15-item SE-SC and the original SE-SC was .98 ($p$ < .0001).

### Concurrent Validity

To furnish concurrent validity of the SE-SC, the SWAI-T was utilized given its previous validation with trainees engaged in supervision (Gonsalvez et al., 2017). A Pearson's *r* correlation between the original SE-SC and the SWAI-T was .80 ($p < .0001$), indicating a large association between them, providing evidence of concurrent validity. A Pearson's *r* correlation between the revised 11-item subscale version of the SE-SC and the SWAI-T was .78 ($p < .0001$), and .77 ($p < .0001$) with the 15-item revised version.

### Assessing Reactivity Threats to Validity

A Pearson correlation was calculated between the SE-SC (with and without outlier response) and the MCSDSS-A to assess for participant reactivity as a possible threat to validity. No statistically significant association was identified within the sample between the SE-SC ($r[86] = .04333$, $p = .7$; $r\ [85] = .0575$, $p = .61$) and the MCSDSS-A ($\alpha = .72$; KR[20] = .72). No statistically significant association was indicated between the 11-item revised SE-SC subscales ($r[86] = .006$, $p = .954$; $r\ [85] = .02$, $p = .874$), or the 15-item revise SE-S ($r[86] = -0.004$, $p = .969$; $r\ [85] = .006$, $p = .953$).

## Discussion

The purpose of this study was to investigate the psychometric properties of the SE-SC with a sample of CITs in the United States. The SE-SC was designed to be completed by supervisors about their supervisors' competence and effectiveness in supervision. Results of the study support the ongoing development of SE-SC with specific considerations for item revision, item removal, and response category revision. Below, findings are discussed further alongside limitations and future implications.

**Instrument Validity and Reliability**

Reliability ($\alpha$ = .97) and validity of the instrument were established according to instrument development and revision practices (DeVellis, 2017). Initial evidence of concurrent validity was established, with outliers removed, using the SWAI-T ($r$ = .80, $p$ < .0001), which was also previously utilized in the development of the SE-SC (Gonsalvez et al., 2017). Additionally, we wanted to attend to a possible threat to instrument validity, due to sample characteristics reporting about an authority figure, by measuring social desirability. No statistically significant association ($r$ = .04333, $p$ = .70) was found between the SE-SC and the MCSDSS-A. While our findings provide evidence of validity and reliability established for the current form of the SE-SC, item-level fitness to the GPCM seems to tell a different story.

**Model Fitness**

Critically, this study tested item-level fitness to a polytomous item-response model (GPCM). Model fitness assessment of the SE-SC sought to explore the item-level measurement structure with the current sample of CITs. Items of the SE-SC that were identified as not conforming to the model, such as those identified in Table 2.1 and 2.2, require further scrutiny and research to evaluate their performance. But, the small sample size of the current study limits generalizability and, particularly, goodness-of-fit value calculation. Nonetheless, results from this sample indicate multiple areas of nonconformity to the GPCM, necessitating a closer look at which items most closely fit the model. Given the importance of satisfying model assumptions and then exploring fitness, item revision decision making was considered in such an order.

Items of the SE-SC were examined for assumption violation first and fitness second. Based on the results of this sample, the resulting items that most appropriately fit the GPCM included Items 5, 6, 7, 8 (Restorative), 21, 22, 23, 24, 26 (Formative), and 13 and 16

(Normative). Of note, while scholars typically recommend a three-item minimum to capture a single latent variable (DeVellis, 2017), items on the Normative subscale did not satisfy basic response category assumptions. This, considered a limitation, is discussed later. Upon item satisfaction of model assumptions, as presented in Table 2.3, an 11-item shortened form of the SE-SC is endorsed by the sample data. However, a mutually acceptable category response scale system was not achieved, which could be attributed to sample size. Further, as presented in Table 2.4, item fitness estimates range from acceptable to special quantities, such as NaNs (Not a Number). Such NaNs indicate incalculable outputs of the model; namely, 0/0. Such an output may indicate an "overfitness" or too perfect of a fit of the model. Future item response analysis of the SE-SC will require larger samples to explore appropriate threshold categories, item estimates, and fitness to item measurement models.

**Table 2.3**
*SE-SC 11-Item Parameters, Response Thresholds (with and without outliers*)*

| Subscale | Item No. | *a* | $b_1$ | $b_2$ | $b_3$ | $b_4$ | $b_5$ | $b_6$ |
|---|---|---|---|---|---|---|---|---|
| Restorative (*n* = 4) | 5 | 1.859 | -2.399 | -1.691 | -1.484 | -0.397 | NV | NV |
| | 6 | 2.359 | -2.59 | -1.142 | -0.899 | -0.395 | NV | NV |
| | 7 | 9.255 | -2.076 | -1.596 | -1.201 | -0.709 | -0.031 | NV |
| | 8 | 4.96 | -1.847 | -1.164 | -0.943 | -0.322 | NV | NV |
| Formative (*n* = 5) | 21 | 2.516 | -2.557 | -2.186 | -1.373 | -1.258 | -0.581 | -0.191 |
| | 22 | 3.168 | -2.09 | -1.275 | -1.029 | -0.671 | -0.1 | NV |
| | 23 | 4.167 | -2.658 | -1.599 | -1.408 | -1.138 | -0.552 | 0.166 |
| | 24 | 2.898 | -1.991 | -1.548 | -1.168 | -1.006 | -0.388 | 0.207 |
| | 26 | 2.802 | -2.054 | -1.42 | -1.154 | -0.525 | -0.107 | NV |
| Normative (*n* = 2) | 13 | 7.027 | -2.12 | -1.671 | -1.265 | -0.436 | 0.237 | NV |
| | 16 | 0.993 | -2.273 | -0.936 | -0.492 | 0.234 | *-0.024* | NV |

Note. *no difference indicated in parameter estimates for outlier account ($n$ = 85 v. $n$ = 86); NV = No Value

**Table 2.4**
*SE-SC 11-Item Fit Statistics (without outlier)*

| Subscale | Item No. | *EFA Loading* | MNSQ Outfit | *RMSEA** | S-$\chi^2$* |
|---|---|---|---|---|---|
| Restorative ($n$ = 4) | 5 | .74 | .792 | .109 | .136 |
| | 6 | .78 | .702 | .177 | .012 |
| | 7 | .97 | .193 | NaN | NaN |
| | 8 | .89 | .444 | NaN | NaN |
| Formative ($n$ = 5) | 21 | .84 | .655 | .052 | .298 |
| | 22 | .89 | .613 | .118 | .089 |
| | 23 | .91 | .565 | NaN | NaN |
| | 24 | .87 | .659 | .151 | .087 |
| | 26 | .86 | .627 | .082 | .182 |
| Normative ($n$ = 2) | 13 | .76^ | .083 | NaN** | NaN** |
| | 16 | .76^ | .830 | .522** | 0** |

Note. *=*p-values*; ^=*alpha;* **=$\chi^2$ *p-value; EFA = exploratory factor analysis*

**Table 2.5**
*SE-SC 11-Item Fit Statistics (with outlier)*

| Subscale | Item No. | *EFA Loading* | MNSQ Outfit | *RMSEA** | S-$\chi^2$* |
|---|---|---|---|---|---|
| Restorative ($n$ = 4) | 5 | .74 | .792 | NaN | NaN |
| | 6 | .78 | .702 | .176 | .012 |
| | 7 | .97 | .193 | NaN | NaN |
| | 8 | .89 | .444 | NaN | NaN |
| Formative ($n$ = 5) | 21 | .84 | .643 | .225 | .021 |

| Subscale | Item No. | *EFA Loading* | MNSQ Outfit | *RMSEA** | S-$\chi^2$* |
|---|---|---|---|---|---|
| | 22 | .89 | .611 | .119 | .085 |
| | 23 | .91 | .568 | NaN | NaN |
| | 24 | .87 | .653 | .144 | .097 |
| | 26 | .86 | .619 | .085 | .167 |
| Normative (*n* = 2) | 13 | .76^ | .083 | NaN** | NaN** |
| | 16 | .76^ | .830 | .541** | 0** |

Note. **=p-values*; ^=*alpha; **=*$\chi^2$ *p-value*

**Table 2.6**

*SE-SC Original Version and SE-SC 11-Item Revised with GPCM AIC and BIC without outlier*

| Subscale | AIC | BIC |
|---|---|---|
| Restorative Original | 1401 | 1516 |
| Restorative Revised | 619 | 670 |
| Formative Original | 1162 | 1260 |
| Formative Revised | 952 | 1030 |
| Normative Original | 1769 | 1901 |
| Normative Revised | 482 | 509 |

*Note*. AIC = Akaike's Information Criterion;
BIC = Bayesian Information Criterion

**Table 2.7**

*SE-SC Original Version and SE-SC 11-Item Revised with GPCM AIC and BIC with outlier*

| Subscale | AIC | BIC |
|---|---|---|
| Restorative Original | 1401 | 1516 |
| Restorative Revised | 635 | 686 |
| Formative Original | 1163 | 1261 |

| Subscale | AIC | BIC |
|---|---|---|
| Formative Revised | 969 | 1050 |
| Normative Original | 1769 | 1902 |
| Normative Revised | 504 | 533 |

*Note*. AIC = Akaike's Information Criterion;
BIC = Bayesian Information Criterion

**Instrument Revision**

This study is the first of its kind to examining the item-level fitness of the subscales of the SE-SC, according to a polytomous Rasch model, with a US-based CIT sample. The present study supports a variable response category system for the subscales of the SE-SC and an abbreviated 11-item version (Restorative = Items 5, 6, 7, 8; Formative = Items 21, 22, 23, 24, 26; and Normative = Items 13 and 16). As presented in Table 2.3, the spread of discrimination parameters (*a*) for items in the revised scale is broader than in Table 2.1; meaning that items are collectively able to better assess/detect responses across the continuum of response categories. The revised 11-item SE-SC, as presented in Appendix I, is a briefer, less time-consuming scale with more utility that may be more easily employed across clinical, training, and research settings. A preliminary Pearson correlation between the revised subscale-only 11-item SE-SC and the original SE-SC ($r = .98$, $p = < 0.0001$) was acceptable. Preliminary Pearson correlation between the revised 15-item SE-SC and the original SE-SC ($r = .98$; $p = < 0.0001$) was also acceptable. Internal reliability, calculated using Cronbach's alpha without the outlier from original data, of the revised 11-item SE-SC ($\alpha = .94$) and the 15-item SE-SC ($\alpha = .96$) were excellent. Critically, alpha levels are expected to be high in the development of instruments as future iterations of instruments are expected to capture the same theoretical constructs as previous iterations. However, revised or short scales cannot be constructed within one study.

Rather, scale revision takes multiple studies, various methodologies for examination, and significant amounts of data to suggest rigorous reconstruction (Ziegler et al., 2014).

Future item-response models, in addition to classical test theory models, will shed further light on the performance of the SE-SC as researchers may be able to use developer suggested evaluation parameters. Findings presented within this study require verification and replication as the presented revised subscales cast doubt on the utility of the instrument to holistically assess supervisor competence for a US-based sample. In the case of further item revision and refinement, a significant amount of data would be required that could be scrutinized using classical test theory techniques alongside item response theory techniques. For example, to assess supervisor competence, SE-SC instrument developers suggested an acceptable score of 6 or greater (personal communication, Craig Gonsalvez, June 13, 2019) on each item. On the response category scale of 1-7, an endorsement of 6 would be acceptable and indicate supervisor competency. Such a benchmark may inform future work on polytomous item response modeling, employing a Graded Response Model (GRM) instead of the Generalized Partial Credit Model (GPCM).

**Limitations**

The results of this study need to be considered in light of the context of the following limitations: small sample size, recruitment barriers, and the nature of instrument refinement. Statistical calculations and resultant conclusions about fitness to the model, are critical for the analysis of item-level performance. As such, the determination of the SE-SC's subscales to be "fitted" or "misfitted" to the model are beholden to the sensitive nature of absolute goodness-of-fit statistics (RMSEA and S-$\chi^2$ ; Sharma et al., 2005). This is critical to note as estimation parameters are impacted in their precision by sample size. This study did not seek to conduct a

large-scale calibration study, as may be appropriate for future research with larger samples, so the precision of the resulting estimations is a notable limitation due to the small *n*.

Subscales of the revised scale, such as Normative, suffer from a lack of adherence to traditional/classical test development practices, such as a latent variable-item minimum of three. As item response theory and classical test theory are complimentary and work in concert, any future revisions of the SE-SC require empirical support using both analytic methods. These item fit statistical limitations inhibit the generalizability of the study findings, but do contribute to a larger body of evidence for future research to consider in refining the SE-SC. As the SE-SC is a relatively new instrument in evaluating supervisor competence, the findings presented herein will contribute to the ongoing development and refinement of the instrument in due course.

Representativeness is critical for demonstrating data credibility and coverage of a population (Chow, 2002; Ramsey & Hewitt, 2005). As we did not obtain sample representativeness to the master's student population in CACREP-accredited programs (CACREP, 2018), the generalizability of the results of this study to the larger population of students in CACREP-accredited programs is limited. This is not surprising, per se, within online survey research (Vicente & Reis, 2007) but does require consideration as to how the lack of representativeness may impact data outcomes and data variance observed (e.g. McGlashan et al., 1988). Indeed, this an important limitation as it impacts the potential utility of these findings to CACREP-accredited programs. Nonetheless, it is worth noting that the sample included CITs from all five ACES regions in the country.

Research efforts were designed to recruit a population representative sample of CITs from CACREP-accredited programs. In execution, however, multiple challenges contributed to the small sample size and, in theory, the data collected. First, recruitment limitations included the

roundabout method of contacting participants, namely through program based CACREP liaisons. Using multiple email contacts to encourage the passing along of recruitment material is an indirect sampling method. Further, the advent of a novel coronavirus, COVID-19 (Zhou et al., 2020) that reached global pandemic status during the recruitment outreach phase might have negatively impacted recruitment. During the recruitment phase, national anxiety (Wang et al., 2020) was heightened and clinical training across the country was disrupted in counseling programs (CACREP, 2020). It is plausible that the resultant in daily life and training interruptions had dampened CITs' interest in participating in research. In short, despite the limitations presented above, this study contributes to a larger body of evidence suggesting a cautious use of the SE-SC within CACREP-accredited programs and a need to further examine the SE-SC's psychometric properties across populations and settings as well as refinement.

## Implications and Recommendations

Findings from this study have implications for (a) counselor education programs, (b) advancing a common measurement approach for supervision scholarship, and (c) constructing multiculturally responsive instruments.

### *Counselor Education Programs*

The ongoing evaluation of supervisor competency is of concern for CACREP-accredited counseling programs. Given the concerning reports from supervisees of harmful and inadequate supervision that exist in the field (Cook, 2019; Ellis et al., 2015), program administrators and faculty would be served well by implementing systematic mechanisms for supervisor evaluation, alongside supervisee and supervision evaluation. With few psychometrically sound instruments to select from to measure supervisor competency, professional counselors and counselor educators require precise and theory-driven tools. As an instrument that was constructed to

specifically assess supervisor competence, the SE-SC is well suited for counselor education programs. However, as it was built internationally and for psychologists, it requires empirical scrutiny – beyond this study – for use with CITs and professional counselors. The revised SE-SC, while theoretically in-tact within the item-response model applied, still possesses significant concerns that require addressing. Namely, advanced item-response theory techniques and classical test analysis are necessary to perform in order to assess the psychometric properties of the SE-SC, and its revised scale, before use with a US-based CIT sample.

At best, the revised SE-SC may be useful for counselor education program to assess the minimal level of competency of their supervisory staff, but only after further research and testing of the psychometric properties. The ongoing monitoring of supervisor competence is equal parts a clinical training concern, a program quality/accreditation concern, and a supervisee welfare concern. The revised SE-SC represents one possible time-saving assessment tool that requires ongoing research and development to suggest its widespread use in counselor education programs.

### *Advancing a Common Measurement Approach for Supervision Research*

This study contributes to the supervision literature in furnishing data to suggest a need for continued refinement of the SE-SC instrument as a measure to evaluate supervisor competency. The ongoing international effort to foster a common measurement approach in supervision research is critical to the advancement of evidence-based supervision. Instruments, such as the SE-SC, employed in methodologically diverse supervision research require especial scrutiny if scholarship is to advance. The findings of this study contribute to the supervision research literature by casting doubt on the item-level performance of the SE-SC for a sample of CITs

enrolled in CACREP-accredited programs in the US, though evidences of score reliability and concurrent validity at the instrument level were found.

As supervision is a multifaceted interdisciplinary intervention and practiced globally, a shared measurement of supervisor competence, like the revised SE-SC, represents a priority for multiple international helping professions. This research is the first of its kind to assess the psychometric properties of an internationally developed instrument with a US-based population. In order to continue to construct and develop robust-enough instruments to assess supervisor competence, further data, analysis, and scrutiny of the SE-SC is required (Ziegler et al., 2014).

### *Building Multiculturally Responsive Supervision Instruments*

In order to utilize an instrument for research purposes, a full accounting of an instrument's psychometric properties is essential. If an instrument or its items perform differently per participant based on demographics or cultural considerations, then further examination is necessary of the instrument's cross-cultural utility. As indicated in the findings of our study, we caution the use of instruments for use in supervision or research without further scrutiny with item response theory methods of analysis or classical test theory methods of analysis.

Future research using item-response theory with the SE-SC will need to explore differential item functioning (DIF) across participants. As a near analogy, differential item function is to item response theory as measurement invariance is to classical test theory (Andrich & Marais, 2019). DIF would assist in determining the cross-cultural utility of the revised, or original, SE-SC by analyzing participant-based differences in performance on an item. While not within the scope of the research question(s) of this study, DIF evaluation of items is critical for item response theory applications of the SE-SC.

## Conclusion

Using a GPCM of polytomous item response theory, we examined the item-level fitness to the model for the SE-SC. Findings indicate the need for ongoing development of the SE-SC's item response categories and revision, or deletion, of misfitting items. Based on the available parameters and fitness estimates from the data, we proposed an abbreviated 11-item version of the SE-SC. As instrument development is an ongoing research praxis and critical for supervision scholars interested in doing advanced multivariate work, the furnishing of evidence to suggest the SE-SC's refinement and potential adoption in CACREP-accredited programs is a first step. This study adds support towards these ends and has contributed significantly toward the development of a common measurement approach within supervision scholarship.

# References

Allan, R., McLuckie, A., & Hoffecker, L. (2016). Clinical supervision of psychotherapists: A systematic review. Retrieved from https://campbellcollaboration.org/media/k2/attachments/Allan_Clinical_Supervision_Title.pdf

Althubaiti, A. (2016). Information bias in health research: Definition, pitfalls, and adjustment methods. *Journal of Multidisciplinary Healthcare*, *9*(1), 211-217. https://doi.org/10.2147/JMDH.S104807

American Counseling Association (2014). ACA code of ethics. Alexandria, VA: Author.

Andrich, D., & Marais, I. (2019). *A course in Rasch measurement theory: Measuring in the educational, social and health sciences*. Springer.

Baker, F.B. (2001). *The basics of item response theory*. College Park, MD: ERIC Clearinghouse on Assessment and Evaluation.

Bambling, M. (2014). Creating positive outcomes in clinical supervision. n C.E. Watkins, Jr. & D.L. Milne (Eds.), *Wiley international handbook of clinical supervision* (pp. 445-457). Oxford, United Kingdom: Wiley.

Bernard, J.M., & Goodyear, R.K. (2014). *Fundamentals of clinical supervision* (5th edition). Boston, MA: Merrill.

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health, 6*(1). https://doi.org/10.3389/fpubh.2018.00149

Bond, T. G. & Fox, C. (2015). *Applying the Rasch Model: Fundamental Measurement in the*

*Human Sciences* (3rd ed.). New York, NY: Routledge.

Borders, L. D. (2014). Best practices in clinical supervision: Another step in delineating effective supervision practice. *American Journal of Psychotherapy*, *68*(2), 151-162. https://doi.org/10.1176/appi.psychotherapy.2014.68.2.151

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York, NY: Guilford Press.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological methods & research*, *21*(2), 230-258. https://doi.org/10.1177/0049124192021002005

Chalmers, R., P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software, 48*(6), 1-29. https://doi.org/10.18637/jss.v048.i06

Chon, K. H., Lee, W. C., & Dunbar, S. B. (2010). A comparison of item fit statistics for mixed IRT models. *Journal of Educational Measurement*, *47*(3), 318-338. https://doi.org/10.1111/j.1745-3984.2010.00116.x

Chow, S. L. (2002). Issues in statistical inference. *History and Philosophy of Psychology Bulletin*, *14*(1), 30-41.

Council for Accreditation of Counseling and Related Educational Programs. (2018). CACREP vital statistics 2017: Results from a national survey of accredited programs. Alexandria, VA: Author.

Council for Accreditation of Counseling and Related Educational Programs [CACREP]. (2020). CACREP Response to COVID-19. https://www.cacrep.org/news/cacrep-statement-on-covid-19/

Copeland, P., Dean, R. G., & Wladkowski, S. P. (2011). The power dynamics of supervision:

Ethical dilemmas. *Smith College Studies in Social Work*, *81*(1), 26-40. https://doi.org/10.1080/00377317.2011.543041

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of applied psychology*, *78*(1), 98. https://doi.org/10.1037/0021-9010.78.1.98

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334. https://doi.org/10.1007/BF02310555

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of consulting psychology*, *24*(4), 349-354. https://doi.org/10.1037/h0047358

Dawson, M., Phillips, B., & Leggat, S. (2013). Clinical supervision for allied health professionals: A systematic review. *Journal of Allied Health*, *42*(2), 65-73. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/23752232

DeVellis, R. F. (2017). *Scale development: Theory and applications* (5th ed). Los Angeles: Sage Publications.

Dziak, J. J., Coffman, D. L., Lanza, S. T., Li, R., & Jermiin, L. S. (2020). Sensitivity and specificity of information criteria. *Briefings in bioinformatics*, *21*(2), 553-565. https://doi.org/10.1093/bib/bbz016

Efstation, J. F., Patton, M. J., & Kardash, C. M. (1990). Measuring the working alliance in counselor supervision. *Journal of Counseling Psychology, 37*(3), 322–329. https://doi.org/10.1037/0022-0167.37.3.322

Ellis, M. V., D'Iuso, N., & Ladany, N. (2008). State of the art in the assessment, measurement,

and evaluation of clinical supervision . In A. K. Hess, K. D. Hess, & T. H. Hess, (Eds.), *Psychotherapy supervision: Theory, research, and practice* , (2nd ed., pp. 473 – 499 ). New York: Wiley.

Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. Mahwah, NJ: Lawrence Erlbaum.

Falender, C. A. (2014). Clinical supervision in a competency-based era. *South African Journal of Psychology*, *44*(1), 6-17. https://doi.org/10.1177/0081246313516260

Friedlander, M. L., & Ward, L. G. (1984). Development and validation of the Supervisory Styles Inventory. *Journal of Counseling Psychology*, *31*(4), 541-557. https://doi.org/10.1037/0022-0167.31.4.541

Gonsalvez, C. J., & Calvert, F. L. (2014). Competency-based models of supervision: Principles and applications, promises and challenges. *Australian Psychologist*, *49*(4), 200-208. https://doi.org/10.1111/ap.12055

Gonsalvez, C. J., & McLeod, H. J. (2008). Toward the science-informed practice of clinical supervision: The Australian context. *Australian Psychologist*, *43*(2), 79-87. https://doi.org/10.1080/00050060802054869

Gonsalvez, C. J., Hamid, G., Savage, N. M., & Livni, D. (2017). The supervision evaluation and supervisory competence scale: Psychometric validation. *Australian Psychologist*, *52*(2), 94-103. https://doi.org/10.1111/ap.12269

Kilminster, S. M., & Jolly, B. C. (2000). Effective supervision in clinical practice settings: a literature review. *Medical Education*, *34*(10), 827-840. https://doi.org/10.1046/j.1365-2923.2000.00758.x

Ladany, N., & Muse-Burke, J.L. (2001). Understanding and conducting supervision research. In

L. J. Bradley, & N. Ladany (Eds.), *Counselor supervision* (3rd ed., pp. 304–329). Philadelphia: Brunner-Routledge.

Lambie, G. W., Mullen, P. R., Swank, J. M., & Blount, A. (2018) The Counseling Competencies Scale: Validation and Refinement. *Measurement and Evaluation in Counseling and Development, 51*(1), 1-15. https://doi.org/10.1080/07481756.2017.1358964

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association, 83*, 1198–1202. https://doi.org/10.1080/01621459.1988.10478722

Loo, R., & Thorpe, K. (2000). Confirmatory factor analyses of the full and short versions of the Marlowe-Crowne Social Desirability Scale. *The Journal of Social Psychology*, *140*(5), 628-635. https://doi.org/10.1080/00224540009600503

Luke, M. (2019). Supervision in the counselor education context. In J.E.A. Okech & D.J. Rubel (Eds.), *Counselor education in the 21st century.* (pp. 35-52). Alexandria, VA: American Counseling Association.

McGlashan, T. H., Carpenter Jr, W. T., & Bartko, J. J. (1988). Issues of design and methodology in long-term followup studies. *Schizophrenia Bulletin*, *14*(4), 569-574. https://doi.org/10.1093/schbul/14.4.569

McKibben, W. B., & Silvia, P. J. (2016). Inattentive and socially desirable responding: Addressing subtle threats to validity in quantitative counseling research. *Counseling Outcome Research and Evaluation*, *7*(1), 53-64. https://doi.org/10.1177/2150137815613135

Mertler, C. A., & Vannatta, R. A. (2017). *Advanced and multivariate statistical methods: Practical application and interpretation* (5th ed.). Glendale, CA: Pyrczak.

Milne, D., & Reiser, R. P. (2012). A rationale for evidence-based clinical supervision. *Journal of Contemporary Psychotherapy, 42*(3). https://doi.org/10.1007/s10879-011-9199-8

Milne, D. L., Sheikh, A. I., Pattison, S., & Wilkinson, A. (2011). Evidence-based training for clinical supervisors: A systematic review of 11 controlled studies. *The Clinical Supervisor*, *30*(1), 53-71. https://doi.org/10.1080/07325223.2011.564955

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied psychological measurement, 16*(2), 159-176. https://doi.org/10.1002/j.2333-8504.1992.tb01436.x

Muraki, E. (1993). Information functions of the generalized partial credit model. *Applied Psychological Measurement, 17*(4), 351-363. https://doi.org/10.1002/j.2333-8504.1993.tb01538.x

Mvududu, N. H., & Sink, C. A. (2013). Factor analysis in counseling research and practice. *Counseling Outcome Research and Evaluation*, *4*(2), 75-98. https://doi.org/10.1177/2150137813494766

Olds, K., & Hawkins, R. (2014). Precursors to measuring outcomes in clinical supervision: A thematic analysis. *Training and Education in Professional Psychology*, *8*(3), 158. https://doi.org/10.1037/tep0000034

Proctor, B. (2011). Training for the supervision alliance. In J. R. Cutcliffe, K. Hyrkas & J. Fowler (Eds.), *Routledge Handbook of Clinical Supervision: Fundamental International Themes*. London: Routledge.

R Core Team. (2019). *R: A language and Environment for Statistical Computing.* R Foundation for Statistical Computing. https://www.R-project.org

Ramsey, C. A., & Hewitt, A. D. (2005). A methodology for assessing sample representativeness. *Environmental Forensics*, *6*(1), 71-75. https://doi.org/10.1080/15275920590913877

Revelle, W. (2019). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois. R package version 1.9.12

Revilla, M., & Ochoa, C. (2017). Ideal and maximum length for a web survey. *International Journal of Market Research*, *59*(5), 557-565. https://doi.org/10.2501/IJMR-2017-039

Rizopoulos, D. (2006). ltm: An R package for Latent Variable Modelling and Item Response Theory Analyses. *Journal of Statistical Software, 17*(5), 1-25. http://www.jstatsoft.org/v17/i05/

Rønnestad, M. H., Orlinsky, D. E., Schröder, T. A., Skovholt, T. M., & Willutzki, U. (2019). The professional development of counsellors and psychotherapists: Implications of empirical studies for supervision, training and practice. *Counselling and Psychotherapy Research, 19*(3), 214-230. https://doi.org/10.1002/capr.12198

Reynolds, W. M. (1982). Development of reliable and valid short forms of the Marlowe-Crowne Social Desirability Scale. *Journal of Clinical Psychology*, *38*(1), 119-125. https://doi.org/10.1002/1097-4679(198201)38:1<119::AID-JCLP2270380118>3.0.CO;2-I

Rubin, N. J., Bebeau, M., Leigh, I. W., Lichtenberg, J. W., Nelson, P. D., Portnoy, S., ... & Kaslow, N. J. (2007). The competency movement within psychology: An historical perspective. *Professional Psychology: Research and Practice*, *38*(5), 452-462. https://doi.org/10.1037/0735-7028.38.5.452

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art.

*Psychological Methods*, *7*(2), 147-177. https://doi.org/10.1037/1082-989X.7.2.147

Schmitt, T. A. (2011). Current methodological considerations in exploratory and confirmatory factor analysis. *Journal of Psychoeducational Assessment*, *29*(4), 304-321. https://doi.org/10.1177/0734282911406653

Schutt, M. A. (2012). *Replication and extension of Ellis, Ladany, Krengel, and Shult (1996); Clinical supervision and research from 1981 to 1993: A methodological critique*. (Doctoral dissertation). Retrieved from https://preserve.lehigh.edu/cgi/viewcontent.cgi?article=2297&context=etd

Sharma, S., Mukherjee, S., Kumar, A., & Dillon, W. R. (2005). A simulation study to investigate the use of cutoff values for assessing model fit in covariance structure models. *Journal of Business Research*, *58*(7), 935-943. https://doi.org/10.1016/j.jbusres.2003.10.007

Snowdon, D. A., Millard, G., & Taylor, N. F. (2016). Effectiveness of clinical supervision of allied health professionals: A survey. *Journal of Allied Health*, *45*(2), 113-121.

Spence, S. H., Wilson, J., Kavanagh, D., Strong, J., & Worrall, L. (2001). Clinical supervision in four mental health professions: A review of the evidence. *Behaviour Change, 18*(3), 135–155. http://doi.org/10.1375/bech.18.3.135

Tabachnick, B., & Fidell, L. (2019). *Using multivariate statistics* (7th ed.). Boston, MA: Pearson.

Toland, M. D. (2014). Practical guide to conducting an item response theory analysis. *The Journal of Early Adolescence*, *34*(1), 120-151. https://doi.org/10.1177/0272431613511332

Vicente, P., & Reis, E. (2007, October). Methodological issues in online surveys. In *Proceedings of the IADIS International Conference on WWW/Internet* (Vol. 2, pp. 173-176).

Watkins Jr, C. E. (2012a). Educating psychotherapy supervisors. *American Journal of Psychotherapy*, *66*(3), 279-307. https://doi.org/10.1176/appi.psychotherapy.2012.66.3.279

Watkins Jr, C. E. (2012b). Psychotherapy supervision in the new millennium: Competency-based, evidence-based, particularized, and energized. *Journal of Contemporary Psychotherapy*, *42*(3), 193-203. https://doi.org/10.1007/s10879-011-9202-4

Watkins Jr., C. E. (2014). The supervisory alliance: A half century of theory, practice, and research in critical perspective. *American Journal of Psychotherapy, 68*(1), 19–55. https://doi.org/10.1176/appi.psychotherapy.2014.68.1.19

Watkins Jr., C. E. (2018). The generic model of psychotherapy supervision: An analogized research-informing meta-theory. *Journal of Psychotherapy Integration*, *28*(4), 521-536. https://doi.org/10.1037/int0000114

Wang, C., Pan, R., Wan, X., Tan, Y., Xu, L., Ho, C. S., & Ho, R. C. (2020). Immediate psychological responses and associated factors during the initial stage of the 2019 coronavirus disease (COVID-19) epidemic among the general population in China. *International Journal of Environmental Research and Public Health*, *17*(5), 1729. https://doi.org/10.3390/ijerph17051729

Wheeler, C. E., & Barkham, D. L. (2014). A Core Evaluation Battery for Supervision. In C.E. Watkins, Jr. & D.L. Milne (Eds.), *Wiley international handbook of clinical supervision* (pp. 367-385). Oxford, United Kingdom: Wiley.

Wheeler, S., & Richards, K. (2007). The impact of clinical supervision on counsellors and

therapists, their practice and their clients. A systematic review of the literature. *Counselling and Psychotherapy Research*, *7*(1), 54-65. https://doi.org/10.1080/14733140601185274

Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist, 34*(6), 806-838. https://doi.org/10.1177/0011000006288127

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square-fit values. *Rasch Measurement Transactions, 8,* 370.

Zhou, F., Yu, T., Du, R., Fan, G., Liu, Y., Liu, Z., ... & Guan, L. (2020). Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *The Lancet*. https://doi.org/10.1016/S0140-6736(20)30566-3

Ziegler, M., Kemper, C. J., & Kruyen, P. (2014). Short Scales–Five Misunderstandings and Ways to Overcome Them. *Journal of Individual Differences*, *35*(4), 185-189. https://doi.org/10.1027/1614-0001/a000148

Ziegler, M., & Hagemann, D. (2015). Testing the Unidimensionality of Items. *European Journal of Psychological Assessment*, *31*(4), 231-237. https://doi.org/10.1027/1015-5759/a000309

# Chapter 4: General Conclusions

This chapter synthesizes the findings and implications of the two studies presented in Chapters 2 and 3. Each study focused on evaluating the psychometric properties of a supervision instrument using a polytomous item response model, amongst other validity and reliability parameters. As supervision effectiveness (Chapter 2) and supervisor competence (Chapter 3) remain complex and dynamic latent constructs for supervision researchers to evaluate, each study meaningfully contributes to the body of evidence to suggest the revision and refinement of two—already empirically robust—supervision instruments. Participants in both studies were master's-level counselors-in-training (CIT) enrolled at CACREP-accredited counseling programs. Both studies drew from the same sample of 135 participants, with 86 participants furnishing usable data. In both studies, based on data from the study sample, we sought to address whether the two selected supervision evaluation instruments had rigorous psychometric evidence. The specific research questions for each study are listed below. After summarizing each study, we offer recommendations for the future of this line of research.

## Summary of Manuscript I

The first study, Manuscript I (Chapter 2), explored the Manchester Clinical Supervision Scale – 26 (MCSS-26; Winstanley & White, 2014) as an evaluation instrument of supervision effectiveness. In March of 2020, we conducted one sampling of three rounds of email outreach to all CACREP-accredited master's-level CACREP liaisons and core faculty of specialty programs (clinical mental health counseling, clinical rehabilitation counseling, addictions counseling, and marriage and family counseling) across the country. The research questions for this study were:

1. Does the MCSS-26 and its subscales possess evidence of internal consistency?
2. Does the MCSS-26 possess item-level fitness?

3. When compared with a measure of training environment, does the MCSS-26 possess evidence of concurrent validity?
4. Does social desirability present a significant threat to the validity of the MSCSS-26?

By exploring item-level fitness to a generalized partial credit model (GPCM), we examined the psychometric properties of the MCSS-26 related to validity, reliability, item response model assumptions, item-level difficulty, and item-level discrimination. In examining model fitness to the data, multiple items did not fit the model as expected. Only a few subscales' items fit the model with any coherency. The remaining items were organized into a revised version of the MCSS-26, a 9-item instrument (Appendix F). As the results suggest a revised version of the MCSS-26, the most immediate concerns include item revision for the identified mis-fitting items and item-level response category calibration for those satisfying model assumptions. Such response category calibration work will require a large dataset to suggest substantive revision of the MCSS-26. While the data provided evidence to support the classical instrument psychometrics, such as reliability ($\alpha$ = .92), discriminant validity (CTES: $r$ = .18, p < .10) statistics, the item-level psychometrics cast doubt on the utility of the MCSS-26 in its current form for CITs in U.S.-based accredited training programs.

**Summary of Manuscript II**

The second study, Manuscript II (Chapter 3), explored the Supervision Evaluation and Supervisor Competency Scale (SE-SC; Gonsalvez et al., 2016) as an evaluation instrument of supervisor competence. Following the one-sampling method as described in Manuscript I and II, we reached out to every CACREP-accredited specialty program's (clinical mental health, clinical rehabilitation counseling, addiction counseling, and marriage and family counseling) CACREP liaison and core faculty for assistance with the invitation to participate. Across three rounds of

emails over the course of nine weeks, we engaged in study recruitment. The research questions for this study were:

1. Does the SE-SC possess internal consistency?
2. Does the SE-SC possess item-level fitness?
3. When compared with a measure of supervisory relationship (SWAI-T), does the SE-SC possess concurrent validity?
4. Does social desirability present a significant threat to validity?

In order to systematically explore the psychometric properties of the SE-SC, we analyzed reliability, validity, item response model assumptions, item-level difficulty, and item-level discrimination estimates. To conduct the item-level performance we employed a generalized partial credit model (GPCM), which is a polytomous item-response theory derived statistical model of analysis. The results of this study indicate that a significant number of items of the SE-SC did not fit the GPCM, thus warranting further development of the instrument to validate item-level precision and indicate use in practice. As the subscale-focus of inter-item relationships was the main theoretical assumption of model analysis, the revised SE-SC (Appendix I) is presented with limitations and caution. In order to continue the revision and precision development of the SE-SC item misfit and response categories require calibration with a significantly larger dataset ($n > 500$). Though there was evidence at the instrument-level to indicate reliability ($\alpha = .97$ [original]; .94 [11-item]; .96 [15-item]) and concurrent validity (SWAI-T: $r = .77$, $p < .0001$) correlation, item-level psychometrics from the study sample cast doubt on the use and dissemination of the SE-SC with U.S.-based CIT.

## Limitations

A primary limitation of both studies was the small sample size. Indicated previously, statistical analysis of goodness-of-fit and modeling were limited in their precision. Generally speaking, as *n* increased in item response models so, too, does precision of parameters and fitness estimates. Findings from both studies should thus be presented in light of this important context, though existing literature indicates that the sample size similar to what was in the current studies was acceptable, with a minimum $n = 30$ (Bond & Fox, 2015; Chen et al., 2014; Linacre, 1994).

Another key limitation for consideration with both studies was the representativeness of the sample of the larger population of CACREP-accredited enrollees. The study sample was under-representative of Black/African participants and over-representative of Caucasian/White participants, multiracial identity participants, and female participants (cf. CACREP, 2018). Finally, it seems reasonable to insert that study recruitment was also affected by the global pandemic of COVID-19 (Zhou et al., 2020).

Based on sampling limitations, findings from both studies may not generalize to other CITs in clinically based CACREP-accredited programs and non-CACREP-accredited programs in the U.S. context. Hence, educators, supervisors, and researchers would benefit from using caution in decision-making based on the results.

## Implications and Recommendations

Notwithstanding the above-mentioned limitations, a number of implications from the findings exist for (a) supervision scholarship and (b) counselor education. Each study supports the articulated need for instrument refinement, precision calibration, and construct clarity of what is collectively defined as “effectiveness” and “competency” in supervision scholarship

(Goodyear et al., 2016). While each study examined the psychometric properties of a supervision evaluation instrument, it is perhaps not surprising that the item-level performance of each instrument was not entirely conforming to item response theory assumptions. Supervision is a complex, dynamic, and challenging to quantify phenomena as it is interpersonally delivered, intrapersonally experienced (for the supervisee), and, necessarily, externally evaluated (Goodyear et al., 2016). In order for supervision research to advance and begin implementation of complex research designs and statistical modeling practices – so that we may ultimately begin to engage in comparative analysis – the field requires psychometrically precise, valid, and reliable instruments; such as those considered above.

Instrument selection in research design and execution is simply one strategy to foster methodological rigor. Item response theory (IRT) is one side dimension of facilitating the development of supervision instruments. I recommend future research on the two supervision instruments examined in this dissertation project to employ classical test theory (e.g. multiple group factor analysis) and IRT (e.g. differential item functioning) in concert with each other with a large, representative sample of CITs in the U.S. and other social milieus to verify findings in this project as well as the instruments' psychometric properties in populations that have yet to be verified. Such research may leader to constructing a robust and precise measurement model for supervision instruments, such as the MCSS-26 and SE-SC, for large-scale use in multivariate quantitative research designs and cross-cultural comparative studies. I further recommend further research on the studied instruments' utility in other U.S.-based clinicians such as clinical and counseling psychologists, clinical social workers, and marriage and family therapists.

The practice of supervising CITs remains an integral part of counselor education across the country. So long as the profession continues to grow and rise to the occasion of building a

national workforce that can attend to the country's mental health, rehabilitative, and addiction-related needs, so, too, will counseling programs increasingly rely on supervision to be the primary tool of experiential learning and growth of their graduate students. Counselor education programs invested in CACREP accreditation require tools with empirical support to facilitate learning and satisfy accreditation standards. Thus, supervision evaluation instruments that are theory-driven, brief, and grounded in psychometrically relevant evidence are critical to the task of providing counselors-in-training effective supervision from competent supervisors. The added benefit of incorporating supervision evaluation instruments, such as the MCSS-26 and the SE-SC, into counselor education means that administrators may be more able in monitoring occurrences of harmful or inadequate supervision experiences. However, findings in this project call for further psychometric investigations of these two instruments before an evidence-based decision can be made of the utility of instruments in U.S.-based counseling training program evaluation.

## Conclusion

These two studies examined psychometric properties of an instrument to assess supervision effectiveness and an instrument to assess supervisor competency. In both studies, we found evidence that supported the revision of each instrument according to a polytomous item-response theory model, the generalized partial credit model (GPCM), and contribute to the ongoing refinement of each instrument for eventual use with a US-based trainee population. This original research addressed the main research question, "Do existing supervision evaluation instruments maintain rigorous psychometric evidence for a sample of CITs from CACREP-accredited programs?" by presenting evidence to suggest that existing supervision evaluation

instruments do not possess robust item-level properties, but may be revised according to a polytomous item-response model to build greater theoretical coherency.

As shorter instruments, the recommended 9-item version of the MCSS-26 and the 15-item version of the SE-SC possess utility for supervision research and counselor education due to their brevity and theoretical parsimony (Ziegler et al., 2014), but require further scrutiny before deployment in the field or for conclusive research designs with US-based populations. Psychometrically precise, valid, and reliable instruments are the backbone of any well-designed stud, and the bedrock for any advanced research methodology, such as large-scale online surveys and longitudinal (Sandy et al., 2014). Already quite strong instruments for supervision evaluation with non-U.S.-based populations, our examination of the psychometric properties of the MCSS-26 and the SE-SC resulted in contrary evidence to suggest a pause to their discontinued use with U.S.-based populations until further supportive data of the extant versions, or robust refinement of the revised version from larger and representative samples.

**Bibliography**

Association for Counselor Education and Supervision. (1990). Standards for counseling supervisors. *Journal of Counseling & Development, 69*(1), 30–32. https://doi.org/10.1002/j.1556-6676.1990.tb01450.x

Bond, T. G. & Fox, C. (2015). Applying the Rasch Model: Fundamental Measurement in the Human Sciences (3rd ed.). New York, NY: Routledge.

Boone, W. J. (2016). Rasch analysis for instrument development: why, when, and how?. *CBE—Life Sciences Education*, *15*(4), pp. 1-7. https://doi.org/10.1187/cbe.16-04-0148

Buus, N., & Gonge, H. (2009). Empirical studies of clinical supervision in psychiatric nursing: A systematic literature review and methodological critique. *International Journal of Mental Health Nursing, 18*(4), 250-264. https://doi.org/10.1111/j.1447-0349.2009.00612.x

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York, NY: Guilford Press.

Chen, W. H., Lenderking, W., Jin, Y., Wyrwich, K. W., Gelhorn, H., & Revicki, D. A. (2014). Is Rasch model analysis applicable in small sample size pilot studies for assessing item characteristics? An example using PROMIS pain behavior item bank data. *Quality of life research*, *23*(2), 485-493.

Cook, R. (October, 2019). *Incidents of Inadequate and Harmful Clinical Supervision Experienced by Post-Master's Counselors*. Presentation at the Association for Counselor Education and Supervision at the Sheraton in Seattle, WA.

DeVellis, R. F. (2017). *Scale development: Theory and applications* (5th ed). Los Angeles: Sage Publications.

Dye, H. A., & Borders, L. D. (1990). Counseling supervisors: Standards for preparation and

practice. *Journal of Counseling & Development, 69*(1), 27-29. https://doi.org/10.1002/j.1556-6676.1990.tb01449.x

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of life research*, *16*(1), 5. https://doi.org/10.1007/s11136-007-9198-0

Ellis, M. V., Creaner, M., Hutman, H., & Timulak, L. (2015). A comparative study of clinical supervision in the Republic of Ireland and the United States. *Journal of Counseling Psychology*, *62*(4), 621-631. https://doi.org/10.1037/cou0000110

Ellis, M. V., D'Iuso, N. A. D. I. A., & Ladany, N. (2008). State of the art in the assessment, measurement, and evaluation of clinical supervision. In A. K. Hess, K. D. Hess, & T. H. Hess (Eds), *Psychotherapy supervision: Theory, research, and practice* (pp. 473-499). Hoboken, NJ, US: John Wiley & Sons Inc.

Ellis, M.V., Ladany, N., Krengel, M., & Schult, D. (1996). Clinical supervision research from 1981 to 1993: A methodological critique. *Journal of Counseling Psychology, 43*(1), 35 - 50. https://doi.org/10.1037/0022-0167.43.1.35

Ellis, M., Singh, N. N., Dennin, M. K., & Tosado, M. (2014). Anticipatory supervisee anxiety scale. Unpublished instrument. In J. M. Bernard & R. K. Goodyear (Eds.), *Fundamentals of clinical supervision*, (5th ed., pp. 330–331). Boston, MA: Pearson.

Falender, C.A., Cornish, J.A.E., Goodyear, R.K., Hatcher, R., Kaslow, N.J., Leventhal, G., . . . & Grus, C. (2004). Defining competencies in psychology supervision: A consensus statement. Journal of Clinical Psychology, 60, 771–785.

Falender, C.A., & Shafranske, E.P. (2004). *Clinical supervision: A competency-based approach.* Washington, DC: American Psychological Association.

Fonseca, J. R. S. (2013). Clustering in the field of social sciences: That's your choice, *International Journal of Social Research Methodology, 16*(5), 403-428. https://doi.org/10.1080/13645579.2012.716973

Heppner, P. P., Wampold, B. E., Owen, J., Thompson, M.N., & Wang, K. T. (2016). *Research design in counseling* (4th ed.). Boston, MA: Cengage.

Jones, L. V., & Thissen, D. (2007). A history and overview of psychometrics. In C.R. Rao and S. Sinharay (Eds). *Handbook of Statistics, 26: Psychometrics* (pp.1-27). Amsterdam: North Holland.

Kadushin, A. (1985). *Supervision in social work.* London, UK: Columbia University Press.

Ladany, N., Mori, Y., & Mehr, K. E. (2013). Effective and ineffective supervision. *The Counseling Psychologist*, *41*(1), 28-47. https://doi.org/10.1177/0011000012442648

Lambert, M. J., & Ogles, B. M. (1997). The effectiveness of psychotherapy supervision. In C. E. Watkins, Jr. (Ed.), *Handbook of psychotherapy supervision* (pp. 421-446). Hoboken, NJ, US: John Wiley & Sons Inc.

Lehrman-Waterman, D., & Ladany, N. (2001). Development and validation of the evaluation process within supervision inventory. *Journal of Counseling Psychology, 48*(2), 168–177. https://doi.org/10.1037/0022-0167.48.2.168

Linacre, J. M. (1994). Sample size and item calibration stability, *Rasch Measurement Transactions, 7*(4), 328.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power Analysis and Determination of Sample Size for Covariance Structure Modeling. Psychological Methods, 1, 130-149. https://doi.org/10.1037/1082-989X.1.2.130

MacDonald, J., & Ellis, P. M. (2012). Supervision in psychiatry: terra incognita? *Current*

*Opinion in Psychiatry, 25*(4), 322-326. https://doi.org/10.1097/YCO.0b013e3283541ecc

Milne, D. L., & Watkins, E. Jr. (2014). Defining and Understanding Clinical Supervision: A Functional Approach. In C.E. Watkins, Jr. & D.L. Milne (Eds.), *Wiley international handbook of clinical supervision* (pp. 3-19). Oxford, United Kingdom: Wiley.

Mvududu, N. H., & Sink, C. A. (2013). Factor analysis in counseling research and practice. *Counseling Outcome Research and Evaluation*, *4*(2), 75-98. https://doi.org/10.1177/2150137813494766

Proctor, B. (2011). Training for the supervision alliance. In J. R. Cutcliffe, K. Hyrkas & J. Fowler (Eds.), *Routledge Handbook of Clinical Supervision: Fundamental International Themes*. London: Routledge.

Psychology Board of Australia (2018). *Guidelines for supervisors*. Retrieved November 5, 2019 from https://www.psychologyboard.gov.au/Standards-and-Guidelines/Codes-Guidelines-Policies.aspx

Roth, A., & Pilling, S. (2015). *A competence framework for the supervision of psychological therapies.* Retrieved November 5, 2019 from https://www.ucl.ac.uk/pals/sites/pals/files/background_document_supervision_competences_july_2015.pdf

Schutt, M. A. (2012). *Replication and extension of Ellis, Ladany, Krengel, and Shult (1996); Clinical supervision and research from 1981 to 1993: A methodological critique.* (Doctoral dissertation). Retrieved from https://preserve.lehigh.edu/cgi/viewcontent.cgi?article=2297&context=etd

Ventimiglia, M., & MacDonald, D. A. (2012). An examination of the factorial dimensionality of

the Marlowe Crowne Social Desirability Scale. *Personality and Individual Differences*, *52*(4), 487-491. https://doi.org/10.1016/j.paid.2011.11.016

Watkins Jr, C. E. (2014). The supervisory alliance: A half century of theory, practice, and research in critical perspective. *American Journal of Psychotherapy*, *68*(1), 19-55.

Watkins Jr, C. E. (2015). Extrapolating Gelso's tripartite model of the psychotherapy relationship to the psychotherapy supervision relationship: A potential common factors perspective. *Journal of Psychotherapy Integration*, *25*(2), 143-157. https://doi.org/10.1037/a0038882

Watkins Jr, C. E. (2017). How does psychotherapy supervision work? Contributions of connection, conception, allegiance, alignment, and action. *Journal of Psychotherapy Integration*, *27*(2), 201-216. https://doi.org/10.1037/int0000058

Watkins, C. E., & Milne, D. L. (2014). Clinical supervision at the international crossroads: current status and future directions. In C. E. Watkins, Jr. & D. L. Milne (Eds.), *Wiley international handbook of clinical supervision* (pp. 673-696). Oxford, United Kingdom: Wiley.

Wong, P. T. P., & Wong, L. C. J. (2014). Multicultural Supervision Competencies Questionnaire. In J. M. Bernard & R. K. Goodyear (Eds.), *Fundamentals of clinical supervision* (5th ed., pp. 334–337). Boston, MA: Pearson.

# Appendix A

## IRB Approval Documents

Oregon State University
Research Office

Human Research Protection Program
& Institutional Review Board
B308 Kerr Administration Bldg, Corvallis OR 97331
(541) 737-8008
IRB@oregonstate.edu
http://research.oregonstate.edu/irb

| | | | |
|---|---|---|---|
| Date of Notification | February 13, 2020 | | |
| Notification Type | Approval Notice | | |
| Submission Type | Initial Application | Study Number | IRB-2020-0494 |
| Principal Investigator | Kok-Mun Ng | | |
| Study Team Members | Field, Thomas A; Litherland, Gideon R; Muzacz, Arien K; | | |
| Study Title | Cross-Validation of Two Supervision Instruments with Counseling Trainees from CACREP-accredited Programs | | |
| Review Level | FLEX | | |
| Waiver(s) | Documentation of Informed Consent | | |
| Risk Level for Adults | Minimal Risk | | |
| Risk Level for Children | Study does not involve children | | |
| Funding Source | None | Cayuse Number | N/A |

**APPROVAL DATE:** 02/12/2020 **EXPIRATION DATE:** 02/11/2025

A new application will be required in order to extend the study beyond this expiration date.

**Comments**: Waiver of documentation of Informed Consent under Institutional Policy.

The above referenced study was approved by the OSU Institutional Review Board (IRB). The IRB has determined that the protocol meets the minimum criteria for approval under the applicable regulations pertaining to human research protections. The Principal Investigator is responsible for ensuring compliance with any additional applicable laws, University or site-specific policies, and sponsor requirements.

Study design and scientific merit have been evaluated to the extent required to determine that the regulatory criteria for approval have been met [45CFR46.111(a)(1)(i), 45CFR46.111(a)(2)].

**Adding any of the following elements will invalidate the FLEX determination and require the submission of a project revision:**

- Increase in risk
- Federal funding or a plan for future federal sponsorship (e.g., proof of concept studies for federal RFPs, pilot studies intended to support a federal grant application, training and program project grants, no-cost extensions)
- Research funded or otherwise regulated by a federal agency that has signed on to the Common Rule, including all agencies within the Department of Health and Human Services
- FDA-regulated research
- NIH-issued or pending Certificate of Confidentiality
- Prisoners or parolees as subjects
- Contractual obligations or restrictions that require the application of the Common Rule or which require annual review by an IRB
- Classified research
- Clinical interventions

Oregon State University
Research Office

Human Research Protection Program
& Institutional Review Board
B308 Kerr Administration Bldg, Corvallis OR 97331
(541) 737-8008
IRB@oregonstate.edu
http://research.oregonstate.edu/irb

**Principal Investigator responsibilities:**

- Keep study team members informed of the status of the research.
- Obtain IRB approval for project revisions prior to implementing changes as required by section 8.6 of the Policy Manual.
- Report all unanticipated problems involving risks to participants or others within three calendar days.
- Use only approved consent document(s).

Oregon State University
Research Office

Human Research Protection Program
& Institutional Review Board
B308 Kerr Administration Bldg, Corvallis OR 97331
(541) 737-8008
IRB@oregonstate.edu
http://research.oregonstate.edu/irb

| | | | |
|---|---|---|---|
| Date of Notification | March 24, 2020 | | |
| Notification Type | Approval Notice | | |
| Submission Type | Project Revision | Study Number | IRB-2020-0494 |
| Principal Investigator | Kok-Mun Ng | | |
| Study Team Members | Field, Thomas A; Litherland, Gideon R; Muzacz, Arien K; | | |
| Study Title | Cross-Validation of Two Supervision Instruments with Counseling Trainees from CACREP-accredited Programs | | |
| Review Level | FLEX | | |
| Waiver(s) | Documentation of Informed Consent | | |
| Risk Level for Adults | Minimal Risk | | |
| Risk Level for Children | Study does not involve children | | |
| Funding Source | Western Association for Counselor Education and Supervision | Cayuse Number | N/A |

**APPROVAL DATE:** 03/24/2020 **EXPIRATION DATE:** 02/11/2025

A new application will be required in order to extend the study beyond this expiration date.

**Comments**: Added funding from Western Association for Counselor Education and Supervision. Please see the HRPP website for COVID-19 guidance and updates.

The above referenced study was approved by the OSU Institutional Review Board (IRB). The IRB has determined that the protocol meets the minimum criteria for approval under the applicable regulations pertaining to human research protections. The Principal Investigator is responsible for ensuring compliance with any additional applicable laws, University or site-specific policies, and sponsor requirements.

Study design and scientific merit have been evaluated to the extent required to determine that the regulatory criteria for approval have been met [45CFR46.111(a)(1)(i), 45CFR46.111(a)(2)].

## Adding any of the following elements will invalidate the FLEX determination and require the submission of a project revision:

- Increase in risk
- Federal funding or a plan for future federal sponsorship (e.g., proof of concept studies for federal RFPs, pilot studies intended to support a federal grant application, training and program project grants, no-cost extensions)
- Research funded or otherwise regulated by a federal agency that has signed on to the Common Rule, including all agencies within the Department of Health and Human Services
- FDA-regulated research
- NIH-issued or pending Certificate of Confidentiality
- Prisoners or parolees as subjects
- Contractual obligations or restrictions that require the application of the Common Rule or which require annual review by an IRB

Oregon State University
Research Office

Human Research Protection Program
& Institutional Review Board
B308 Kerr Administration Bldg, Corvallis OR 97331
(541) 737-8008
IRB@oregonstate.edu
http://research.oregonstate.edu/irb

- Classified research
- Clinical interventions

**Principal Investigator responsibilities:**

- Keep study team members informed of the status of the research.
- Obtain IRB approval for project revisions prior to implementing changes as required by section 8.6 of the Policy Manual.
- Report all unanticipated problems involving risks to participants or others within three calendar days.
- Use only approved consent document(s).

## Appendix B

Research Participation Email and Survey

Dear Drs. [LName1], [LName2], [LName3], [LName4], [LName5],

Greetings! My name is Gideon Litherland and I am a doctoral candidate in the PhD Counseling program at Oregon State University. I'm reaching out to you all in your capacity as the CACREP Liaison or as core faculty at your institution. My doctoral dissertation research focuses on counselors-in-training who are engaged in supervision. I am primarily focused on the cross-validation of two different supervision instruments. I am working hard to recruit at least 300 participants. To get a clearer understanding of this research, please review the invitation letter to participants below.

I am requesting your support in the recruitment of potential participants. You can support this project by forwarding the email below to students in your program who are (1) presently engaged in clinical supervision, and (2) enrolled in the clinical mental health counseling, clinical rehabilitation counseling, addictions counseling, or marriage and family counseling tracks.

I am appreciative of your time and attention as you've read this email. Thank you for your consideration and get in touch with any questions you may have.

Kind regards,

Gideon Litherland
litherlg@oregonstate.edu

---

Dear Counselors-in-Training,

My name is Gideon Litherland and I am a doctoral candidate at Oregon State University. I appreciate your time and thank you for considering participating in this research, which is voluntary and anonymous. As a counseling student in a CACREP-accredited university engaged in clinical supervision, I need your assistance with this study.

You are eligible to participate if you are (a) currently engaged in clinical supervision, (b) enrolled within a CACREP-accredited specialty track in clinical mental health counseling, clinical rehabilitation counseling, addictions counseling, or marriage and family counseling.

The purpose of this study is to validate multiple instruments for use in supervision practice and research with counselors-in-training enrolled in CACREP-accredited programs. If you choose to participate your responses will be recorded securely and remain confidential. Your participation and responses will not be reported to your supervisor, program, or school. If you agree to participate in this study you will complete the following:

- Demographic questionnaire
- Manchester Clinical Supervision Scale (MCSS-26)
- Counseling Training Environment Scale (CTES)
- Supervision Evaluation and Supervisor Competence (SE-SC)
- Supervision Working Alliance Inventory-Trainee Version (SWAI-T)

- Marlowe-Crowne Social Desirability Scale Short (MCSD-A)
- Supervision Outcomes Scale (SOS)

Please note that you may discontinue participation in this study at any time, as your participation is voluntary.

At the end of this letter is a link to the study site. If you decide to engage in the research as a participant, we will request your informed consent then proceed to the survey that will take approximately 15-20 minutes to complete. After completing the survey in its entirety, you will be eligible to enter a drawing for one of eight $20 Starbucks gift cards. Your email address for this drawing will not be linked to your survey responses to maintain your confidentiality. The results of this survey will be analyzed and the data will be included in my dissertation and any subsequent publications.

If you choose to participate, or have questions about participating, or have questions about the study itself, then do not hesitate to reach out. My contact information is below along with the primary investigator and co-investigator.

Thank you for considering participating in this study.

https://oregonstate.qualtrics.com/jfe/form/SV_2n8zVGNkV2nSWG1

Sincerely,

Gideon Litherland MA, LCPC (IL), CCMHC, BC-TMH, Doctoral Candidate
litherlg@oregonstate.edu

Dr. Kok-Mun Ng, Principal Investigator
Kokmun.ng@oregonstate.edu

Dr. Thom Field, Co-Investigator
fieldth@oregonstate.edu

Study Title: Cross-Validation of Two Supervision Instruments with Counseling Trainees from CACREP-accredited Programs

**Appendix B (cont.)**

Follow Up Thank You Email for Liaisons and Core Faculty

Dear Drs. [LName1], [LName2], [LName3], [LName4], [LName5],

Hello again! I hope this message finds you all well. I am happy to share that this research is under way. While I am not knowledgeable if you have or have not forwarded this email to your counselors-in-training, I simply wanted to follow up on this thread to express my thanks!

If you haven't yet, you can support this project by forwarding the email below to students in your program who are (1) presently engaged in clinical supervision, and (2) enrolled in the clinical mental health counseling, clinical rehabilitation counseling, addictions counseling, or marriage and family counseling tracks.

I am appreciative of your time and attention!

Kind regards,

Gideon Litherland
litherlg@oregonstate.edu

**Appendix B (cont.)**

Participant Informed Consent

**Thank you for your interest in our research and considering participating!**

Study Title: Cross-Validation of Two Supervision Instruments with Counseling Trainees from CACREP-accredited Programs

We are interested in understanding supervision effectiveness and supervisor competency. You will be presented with multiple questions related to your clinical supervision experience and asked to respond to all questions. Please be assured that your responses will be kept completely confidential and research data will be securely stored.

The study should take you around 15-20 minutes to complete. You may complete the survey in your own space and on your own time. Though there is no compensation for participating, you will be invited to enter a drawing for one of eight $20 Starbucks gift cards upon completion of the survey. Your participation in this research is voluntary. You have the right to withdraw at any point during the study, for any reason, and without any prejudice. If you do withdraw at any point during the study, your collected data will be destroyed and will not be used by researchers. If you would like to contact the Principal Investigator in the study to discuss this research, please e-mail kokmun.ng@oregonstate.edu.

Please review the following information so that you may make a decision about whether you would like to participate. Also, consider reviewing this research participant education worksheet to consider other important questions that may not be addressed below (https://www.hhs.gov/ohrp/sites/default/files/questions_full_list_v5-remediated_12222016.pdf). After this section, you may decide to participate and complete the Informed Consent.

**What is the purpose of this research?** We are seeking to evaluate the psychometric properties of multiple supervision instruments that have not previously been used with US-based counselors-in-training. You, as a counselor-in-training currently engaged in clinical supervision in a CACREP-accredited program, provide a valuable perspective. This study is done as part of the requirements for the PhD in Counseling degree by Gideon Litherland, under the direction of Dr. Kok-Mun Ng, Professor of Counselor Education at Oregon State University.

**What would I be asked to do as a participant?** You would complete a demographic questionnaire, the Manchester Clinical Supervision Scale, the Counseling Training Environment Scale, the Supervision Evaluation and Supervisory Competence Scale, the Supervision Outcomes Scale, the Supervision Working Alliance Inventory-Trainee version, and the Marlowe-Crowne Social Desirability Scale short form A. While the demographic questions are optional for you to complete, you will not be allowed to skip any instrument-based questions. As such, this survey will not allow you to skip any questions that require a response. If you do not wish to answer any

instrument-based questions, you always have the option to simply exit the study, but will not be eligible for the raffle. In total, you will be asked to respond to 130 questions comprising demographic (18) and instrument questions (111). Your responses will only be used by researchers if you complete the whole survey.

**What are the participation criteria?** In order to participate in this research, you must be: (a) Age 18 and above.
(b) Enrolled in CACREP-accredited master's level specialty tracks.

> (i) Clinical Mental Health Counseling. (ii) Clinical Rehabilitation counseling. (iii) Addiction Counseling. (iv) Couple, Marriage, and Family Counseling.

(c) Currently enrolled in field placements (i.e., practicum or internship).

(d) Currently engaged in regular supervision with program supervisor (e.g., faculty and/or doctoral supervisor) and/or counseling site supervisor.

**What are possible risks to participants?** There are no professional or educational risks to participating in this study due to the anonymity of participating. You are able to exit the study any time by simply closing their internet browser. You may discontinue participation at any time without penalty or loss of benefits to which you are otherwise entitled. Your withdrawal from participation will not have any negative impact on your academic performance in this institution.

As with all online activity, some risk is incurred by participants in sharing even de-identified personal data. The security and confidentiality of information collected from participants online cannot be guaranteed. To reduce the risk of a data breach and attempt to ensure participant confidentiality, all information will be entered and stored on an approved, secure, encrypted platform, Qualtrics, with no accompanying identifying information collected. Only authorized research team members will have access to the study data. Participants can discontinue their participation at any time by closing the browser on their computer.

There is no known or perceived physical, psychological, social, or economic risks involved in participating in this study. Of course, unforeseeable risks are challenging to account for so we may not know about all of the risks of being involved in this study. While eligible participants are above the age of 18, possible anticipated risks may include emotional discomfort or reactivity of participants as they respond to questions about their supervision context. If you find yourself stressed or concerned by any of the questions, we provide contact information for warmlines that are accessible 365 days a year:

- Warmlines: David Romprey Oregon Warmline (1.800.698.2392)
- SAMHSA’s National Helpline (1.800.662.4357)
- Text “START” to 741741

**What are possible benefits to participants?** There are no anticipated direct benefits to individuals. Indirect, or aspirational, benefits to participants and society are possible within publication of data analysis and determination of implications for the field.

**Who will see the information I share and what will be done with it (confidentiality)? I**n order to minimize the chances of a breach of confidentiality, no names or email addresses will be collected from participants. Participant demographics will be collected, but only reported in aggregate form. Further, a approved platform, Qualtrics, is being used for survey administration to reduce participant exposure.

Collected data will be stored on an approved platform that is sponsored by Oregon State University, Qualtrics. Data will be stored for a minimum of seven (7) years post-study termination in electronic format on the same OSU-sponsored platform. In addition to OSU cloud storage, data may be downloaded and stored in password-protected electronic devices. Participant data will be reported in aggregate in future publications with no identifying information shared. Aggregate data may also be shared with original instrument developers per licensing or written agreement.

Names and emails as entries for the raffle will be collected separately from the actual study data. Study data will not be linked to any identifiers. Study data is collected anonymously and not linked to raffle names/emails.

**Who do I contact if I have questions?** If you have any questions, as a prospective participant or as a participant, please contact Dr. Kok- Mun Ng at kokmun.ng@oregonstate.edu or 1.541.737.3741, or Gideon Litherland at litherlg@oregonstate.edu or 1.630.212.1128. If you have questions about your rights or welfare as a participant, please contact the Oregon State University Human Research Protection Program (HRPP) office, at 1.541.737.8008 or by email at IRB@oregonstate.edu

Please print the page and/or take a screenshot if you would like to retain a copy of this informed consent for your records.

---

Please answer these two questions before you can access the research materials.

1. I have read the information provided, any questions have been answered, and I agree to participate.
    - Yes, I have read the information provided, my questions have been answered, and I agree to participate.
    - No, I do not want to participate.

2. Informed Consent Agreement - By clicking the button below, you acknowledge (a) that your participation in the study is voluntary, (b) you are 18 years of age and meet the participation criteria, (c) you have reviewed the provided materials describing the scope of this research and the voluntary nature of your participation, and (d) that you are aware that you may choose to terminate your participation in the study at any time and for any reason, without consequence.

**Your participation in this research is voluntary. You have the right to withdraw at any point during the study, for any reason, and without any prejudice.**

**Please note that this survey will be best displayed on a laptop or desktop computer. Some features may be less compatible for use on a mobile device.

Do you consent to participate in this study?

- I do
- I do not
- I am not yet ready to participate. I have further questions that I will follow up with the researcher.

## Appendix B (cont.)

Demographic Questionnaire

Starts on following page.

**Demographics**

Please indicate your stage of clinical training within your current program

- ○ I am not currently completing hours at a practicum or internship placement.
- ○ I am currently completing my practicum or internship hours at a site placement.

Based on the response you selected, you are not eligible to participate in this survey at this time.

Thank you for taking time to consider participating. Feel free to follow up with us with any questions you may have:

Kok-Mun Ng
kokmun.ng@oregonstate.edu

Gideon Litherland
litherlg@oregonstate.edu

How do you self-identify in regards to gender?

- ○ Female
- ○ Male
- ○ Gender non-binary

Do you self-identify as a sexual minority?

- ○ Yes.
- ○ No.
- ○ Prefer not to disclose.

Please indicate your age, in years.

Please indicate your race/ethnicity.

- ☐ Caucasian/European/White
- ☐ Black/African
- ☐ American Indian or Alaska Native
- ☐ Asian
- ☐ Native Hawaiian or Pacific Islander
- ☐ Latinx/Hispanic/Spanish
- ☐ Multiracial
- ☐ Other
- ☐ Prefer to not disclose

Are you an international student?

- ○ I am an international student.
- ○ I am not an international student.
- ○ I prefer to not disclose.

Please indicate your program's region within the US.

- ○ North Atlantic (Connecticut, Delaware, District of Columbia, Maine, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, Vermont)
- ○ North Central (Illinois, Indiana, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, Oklahoma, South Dakota, Wisconsin)
- ○ Rocky Mountain (Colorado, Idaho Montana, New Mexico, Utah, Wyoming)
- ○ Southern (Alabama, Arkansas, Florida, Georgia, Kentucky, Louisiana, Maryland, Mississippi, North Carolina, South Carolina, Tennessee, Texas, Virginia, West Virginia)
- ○ Western (Alaska, Arizona, California, Hawaii, Nevada, Oregon, Washington)
- ○ Outside of regional boundaries

Please indicate the instructional environment - learning delivery - of your program

- ○ Hybrid (Definition: 30-50% of your coursework is delivered via the internet with some face to face, in-person interactions with peers and faculty)
- ○ Fully Online (Definition: 50% or more of your coursework is delivered via the internet)
- ○ Traditional, In-Person, On Ground (Definition: Less than 30% of coursework delivered via the internet)

How many practicum and internship hours have you **approximately** completed in total, at the time of taking this survey? (If in practicum, just practicum hours. If in internship, add practicum and internship hours thus far.)

Instructions for all proceeding questions:

To meaningfully respond to the following questions, think about your clinical supervision and the **current supervisor** with whom you are working. Depending on your program, clinical supervision may be provided off-site by non-University affiliated staff or, even, on-site by your faculty within a University clinic. In either case, the following questions are about your current supervision experience.

Please answer the following set of questions about you and your **current (or most recent) individual supervisor**. Again, remember if you have MORE THAN ONE individual supervisor in your current (or most recent) clinical training setting, please choose one and use the same one throughout the survey.

*Friendly reminder:* Your participation is confidential and your responses are not shared with your school.

Based on the supervisor you selected, please indicate your supervisors' affiliation:

- ○ My supervisor is a faculty member.
- ○ My supervisor is a doctoral student.
- ○ My supervisor is at my site (site supervisor).
- ○ Other:

What setting does supervision take place with this supervisor?

- ○ University-affiliated clinic
- ○ Community mental health center/Agency
- ○ Private practice office
- ○ Group practice office
- ○ Telesupervision via videoconference

Please indicate the theoretical orientation(s) that **you** ascribe to or practice from

- ☐ Psychodynamic
- ☐ Cognitive-Behavioral
- ☐ Humanistic
- ☐ Systems

- [ ] Interpersonal
- [ ] Eclectic
- [ ] Other not listed: (please write-in)

Please indicate the theoretical orientation(s) that you think (or know) **your supervisor** ascribes to or practices from

- [ ] Psychodynamic
- [ ] Cognitive-Behavioral
- [ ] Humanistic
- [ ] Systems
- [ ] Interpersonal
- [ ] Eclectic
- [ ] Other not listed: (please write-in)

What type of supervision you receive with this supervisor?

- ○ Individual supervision
- ○ Triadic supervision
- ○ Group supervision
- ○ Mix of individual and group

How long have you been receiving clinical supervision with this supervisor?

Year(s)
Month(s)

On average, how often are your clinical supervision sessions?

- ○ Weekly
- ○ Every two (2) weeks
- ○ Monthly
- ○ 2-3 months
- ○ >3 months

On average, how long are your clinical supervision sessions?

- ○ <15 minutes
- ○ 15-30 minutes
- ○ 31-45 minuts
- ○ 46-60 minutes
- ○ >60 minutes

**Appendix B (cont.)**

Manchester Clinical Supervision Scale (MCSS-26; White & Winstanley, 2014)

You are invited to participate in this confidential survey, which aims to evaluate the effectiveness of Clinical Supervision (CS) provided to you at your workplace. There are two sections that will take about 10 minutes to complete. This investment of your time will provide unique and valuable insights, to help inform the future development of Clinical Supervision.

**Section A is designed to for individuals *currently* receiving Clinical Supervision (CS).**

Drawing on your current experience of receiving Clinical Supervision at your workplace, please indicate your level of agreement with the following 26 statements, by selecting the box which best represents your answer. Do not spend too long thinking about each question; your first response is probably the best one.

| Strongly Disagree | Disagree | Somewhat Disagree | Neither agree nor disagree | Agree | Strongly Agree |
|---|---|---|---|---|---|

1. Other work pressures interfere with CS sessions
2. It is difficult to find the time for CS sessions
3. CS sessions are not necessary/don't solve anything
4. Time spent on CS takes me away from my real work in the clinical area
5. Fitting CS sessions in can lead to more pressure at work
6. I find CS sessions time consuming
7. My supervisor gives me support and encouragement
8. CS sessions are intrusive
9. CS gives me time to 'reflect'
10. Work problems can be tackled constructively during CS sessions
11. CS sessions facilitate reflective practice
12. My supervisor offers an 'unbiased' opinion

13. I can discuss sensitive issues encountered during my clinical casework with my supervisor
14. My CS sessions are an important part of my work routine
15. I learn from my supervisor's experiences
16. It is important to make time for CS sessions
17. My supervisor provides me with valuable advice
18. My supervisor is very open with me
19. Sessions with my supervisor widen my clinical knowledge base
20. CS is unnecessary for experienced/established staff
21. My supervisor acts in a superior manner during our sessions
22. Clinical supervision makes me a better practitioner
23. CS sessions motivate staff
24. I can widen my skill base during my CS sessions
25. My supervisor offers me guidance with patient/client care
26. I think receiving clinical supervision improves the quality of care I give

**Appendix B (cont.)**

Counseling Training Environment Scale (CTES; Lau, Ng, & Vallett, 2019)

The purpose of the CTES is to assess your perceptions and experiences of the counseling training environment in the counseling and related mental health training program you are attending right now. Please note that due to the nature of some of the items, you must be at least in your second clinical placement of your training.

The items will assess your perceptions about what your current training environment is actually like. Please read each item and using the 5-point Likert-type scale (1 = Strongly Disagree [SD]; 2= Disagree [D]; 3 = Agree [A]; 4 = Strongly Agree [SA]; 5 = Not Applicable [NA]), rate your level of agreement with each item by selecting the appropriate number.

| Strongly Disagree | Disagree | Agree | Strongly Agree | Not Applicable |
|---|---|---|---|---|
| (1) | (2) | (3) | (4) | (5) |

*In my counseling training program...*

1. Questions from students are welcomed in all my classes
2. I get regular feedback from my professors
3. My clinical site supervisor treats me with respect
4. My clinical site supervisor creates a safe environment for the discussion of difficult topics
5. Students have access to University/college resources to facilitate learning and training (e.g., writing labs)
6. Our program has a good relationship with the local community
7. Skills and knowledge gained in my classes are relevant to the work I am doing at my clinical field placement
8. University/college procedures and department procedures for addressing student grievances are consistent
9. Program faculty are active in addressing issues that arise at my clinical field experience site
10. Students are kept abreast of the mental health needs of the community
11. Students are made aware of opportunities to volunteer in community activities
12. My training helps me become cognizant of the impact that my background and life experiences have on my clients and how these may affect my clients
13. Faculty incorporate their clinical experiences into the classroom training
14. Faculty are well-connected within the profession
15. My clinical site supervisor shares clinical resources with me
16. Training curricula meets state standards for professional licensure and/or certification
17. An emphasis is placed on adhering to the ethical codes set forth by the profession
18. We are taught to recognize both within-group and between-group differences
19. My knowledge, awareness, and skills in multicultural counseling has been challenged
20. The program has helped me become mindful of my personal development through time
21. The program is intentional in facilitating students' growth and development
22. My training curricula reflects the current trends of the profession
23. My training is current and reflective of the issues impacting our society today

**Appendix B (cont.)**

Supervision Evaluation and Supervisory Competence Scale (SE-SC; Gonsalvez et al., 2017)

Use the following Likert scale to evaluate the supervision you received by your primary supervisor (individual and group) at the placement you just completed.

| Not at all, Strongly disagree | | | Moderately, Neutral | | | Very much so, Strongly agree |
|---|---|---|---|---|---|---|
| (1) | (2) | (3) | (4) | (5) | (6) | (7) |

1. Overall, my expectations of supervision were matched or exceeded
2. Overall, I would gladly recommend this supervisor to others
3. Overall, supervision significantly enhanced my competence as a practitioner and professional
4. Overall, supervision significantly contributed to my achieving better outcomes for my clients
5. In day-to-day dealings, I got along well with the supervisor
6. The supervisor was understanding and open to a sharing of ideas
7. The supervisor was accepting of my mistakes and inadequacies
8. The supervisor was caring and supportive
9. The supervisor was approachable and interested in my personal and professional development
10. The supervisor impressed me as a skilled therapist
11. The supervisor was knowledgeable and could communicate theoretical concepts clearly
12. The supervision plan appropriately reflected important clinical competencies
13. Supervision objectives were in accordance with my level of professional development
14. The supervisor organized and managed supervision efficiently
15. Supervision methods were varied to match supervision objectives
16. Supervision objectives (goals) were negotiated and clearly articulated
17. Supervision sessions were structured and supervision activities were goal driven
18. I felt comfortable discussing my professional inadequacies in supervision
19. The supervisor was sensitive to my emotional and self-care needs
20. Supervision facilitated emotional ventilation and support as appropriate
21. The supervisor enhanced my abilities to reflect on my clinical work
22. The supervision sessions enhanced my self awareness as a person
23. The supervision furthered my understanding of my own positive and negative interaction patterns with clients
24. The supervisor helped me gain an understanding of my emotional reactions within therapy
25. The supervisor helped inspire me to remain excited about my clinical work and professional responsibilities
26. The supervision advanced my therapist-client relationship skills

**Appendix B (cont.)**

Supervision Outcomes Scale (Tsong & Goodyear, 2014)

Please answer the following set of questions about you and your **current (or most recent) individual supervisor**. Again, remember if you have MORE THAN ONE individual supervisor in your current (or most recent) clinical training setting, please choose one and use the same one throughout the survey.

Please describe the degree to which supervision with **your current (or most recent) individual supervisor** has contributed to the ***IMPROVEMENT*** of the following:

| Not helpful at all (1) | Helpful, but very little (2) | Somewhat helpful (3) | Very helpful (4) | Extremely helpful (5) |
|---|---|---|---|---|

1. Client symptoms (decrease in symptoms)
2. Your relationship with clients
3. Your counseling skills
4. Your case conceptualization ability
5. Your multicultural counseling skills (e.g., skills that are culturally appropriate in working with diverse clients)
6. Your multicultural beliefs/attitudes/awareness (e.g., awareness of your own worldviews)
7. Your multicultural knowledge (e.g., knowledge of worldviews of culturally different clients)

**Appendix B (cont.)**

Supervision Working Alliance Inventory-Trainee Version
(SWAI-T; Efstation, Patton, & Kadash, 1990)

Please indicate the frequency with which the behavior described in each of the following items seems characteristic of your work with your supervisor. After each item, check the space over the number corresponding to the appropriate point of the following 7-point scale: 1 = Almost Never; 7 = Almost Always.

1. I feel comfortable working with my supervisor.
2. My supervisor welcomes my explanations about the client's behavior.
3. My supervisor makes the effort to understand me.
4. My supervisor encourages me to talk about my work with clients in ways that are comfortable for me.
5. My supervisor is tactful when commenting about my performance.
6. My supervisor encourages me to formulate my own interventions with the client.
7. My supervisor helps me talk freely in our sessions.
8. My supervisor stays in tune with me during supervision.
9. I understand client behavior and treatment technique similar to the way my supervisor does.
10. I feel free to mention to my supervisor any troublesome feelings I might have about him/her.
11. My supervisor treats me like a colleague in our supervisory sessions.
12. In supervision, I am more curious than anxious when discussing my difficulties with clients.
13. In supervision, my supervisor places a high priority on our understanding the client's perspective.
14. My supervisor encourages me to take time to understand what the client is saying and doing.
15. My supervisor's style is to carefully and systematically consider the material I bring to supervision.
16. When correcting my errors with a client, my supervisor offers alternative ways of intervening with that client.
17. My supervisor helps me work within a specific treatment plan with my clients.
18. My supervisor helps me stay on track during our meetings.
19. I work with my supervisor on specific goals in the supervisory session.

& Kardash, C. M. (1990). Measuring the working alliance in counselor supervision. Journal of Counseling Psychology, 37(3), 322–329.
https://doi.org/10.1037/0022-0167.37.3.322

**Appendix B (cont.)**

Marlowe-Crowne Social Desirability Scale Short (MCSD-A; Reynolds, 1982)

*(items from original MCSD = 3, 6, 13, 15, 16, 19, 21, 26, 28, 30, 33)*

Listed below are a number of statement concerning personal attitudes and traits. Read each item and decide how it pertains to you.

Please respond either TRUE (**T**) or FALSE (**F**) to each item. Indicate your response by selecting the appropriate letter next to the item. Be sure to answer all items.

1. It is sometimes hard for me to go on with my work if I am not encouraged.
2. I sometimes feel resentful when I don't get my way.
3. No matter who I'm talking to, I'm always a good listener.
4. There have been occasions when I took advantage of someone.
5. I'm always willing to admit it when I make a mistake.
6. I sometimes try to get even rather than forgive and forget.
7. I am always courteous, even to people who are disagreeable.
8. I have never been irked when people expressed ideas very different from my own.
9. There have been times when I was quite jealous of the good fortune of others.
10. I am sometimes irritated by people who ask favors of me.
11. I have never deliberately said something that hurt someone's feelings.

**Appendix B (cont.)**

Qualtrics Survey Preview

Starts on following page.

**Welcome Page**

**Thank you for your interest in our research and considering participating!**

Study Title: Cross-Validation of Two Supervision Instruments with Counseling Trainees from CACREP-accredited Programs

We are interested in understanding supervision effectiveness and supervisor competency. You will be presented with multiple questions related to your clinical supervision experience and asked to respond to all questions. Please be assured that your responses will be kept completely confidential and research data will be securely stored.

The study should take you around 15-20 minutes to complete. You may complete the survey in your own space and on your own time. Though there is no compensation for participating, you will be invited to enter a drawing for one of eight $20 Starbucks gift cards upon completion of the survey. Your participation in this research is voluntary. You have the right to withdraw at any point during the study, for any reason, and without any prejudice. If you do withdraw at any point during the study, your collected data will be destroyed and will not be used by researchers. If you would like to contact the Principal Investigator in the study to discuss this research, please e-mail kokmun.ng@oregonstate.edu.
Please review the following information so that you may make a decision about whether you would like to participate. Also, consider reviewing this research participant education worksheet to consider other important questions that may not be addressed below (https://www.hhs.gov/ohrp/sites/default/files/questions_full_list_v5-remediated_12222016.pdf). After this section, you may decide to participate and complete the Informed Consent.

**What is the purpose of this research?** We are seeking to evaluate the psychometric properties of multiple supervision instruments that have not previously been used with US-based counselors-in-training. You, as a counselor-in-training currently engaged in clinical supervision in a CACREP-accredited program, provide a valuable perspective. This study is done as part of the requirements for the PhD in Counseling degree by Gideon Litherland, under the direction of Dr. Kok-Mun Ng, Professor of Counselor Education at Oregon State University.

**What would I be asked to do as a participant?** You would complete a demographic questionnaire, the Manchester Clinical Supervision Scale, the Counseling Training Environment Scale, the Supervision Evaluation and Supervisory Competence Scale, the Supervision Outcomes Scale, the Supervision Working Alliance Inventory-Trainee version, and the Marlowe-Crowne Social Desirability Scale short form A. While the demographic questions are optional for you to complete, you will not be allowed to skip any instrument-based questions. As such, this survey will not allow you to skip any questions that require a response. If you do not wish to answer any instrument-based questions, you always have the option to simply exit the study, but will not be eligible for the raffle. In total, you will be asked to respond to 130 questions comprising demographic (18) and instrument questions (111). Your responses will only be used by researchers if you complete the whole survey.

**What are the participation criteria?** In order to participate in this research, you must be:
(a) Age 18 and above.
(b) Enrolled in CACREP-accredited master's level specialty tracks.
(i) Clinical Mental Health Counseling. (ii) Clinical Rehabilitation counseling. (iii) Addiction Counseling. (iv) Couple, Marriage, and Family Counseling.
(c) Currently enrolled in field placements (i.e., practicum or internship).
(d) Currently engaged in regular supervision with program supervisor (e.g., faculty and/or doctoral supervisor) and/or counseling site supervisor.

**What are possible risks to participants?** There are no professional or educational risks to participating in this study due to the anonymity of participating.You are able to exit the study any time by simply closing their internet browser. You may discontinue participation at any time without penalty or loss of benefits to which you are

otherwise entitled. Your withdrawal from participation will not have any negative impact on your academic performance in this institution.

As with all online activity, some risk is incurred by participants in sharing even de-identified personal data. The security and confidentiality of information collected from participants online cannot be guaranteed. To reduce the risk of a data breach and attempt to ensure participant confidentiality, all information will be entered and stored on an approved, secure, encrypted platform, Qualtrics, with no accompanying identifying information collected. Only authorized research team members will have access to the study data. Participants can discontinue their participation at any time by closing the browser on their computer.

There is no known or perceived physical, psychological, social, or economic risks involved in participating in this study. Of course, unforeseeable risks are challenging to account for so we may not know about all of the risks of being involved in this study. While eligible participants are above the age of 18, possible anticipated risks may include emotional discomfort or reactivity of participants as they respond to questions about their supervision context. If you find yourself stressed or concerned by any of the questions, we provide contact information for warmlines that are accessible 365 days a year:

- Warmlines: David Romprey Oregon Warmline (1.800.698.2392)
- SAMHSA's National Helpline (1.800.662.4357)
- Text "START" to 741741

**What are possible benefits to participants?** There are no anticipated direct benefits to individuals. Indirect, or aspirational, benefits to participants and society are possible within publication of data analysis and determination of implications for the field.

**Who will see the information I share and what will be done with it (confidentiality)?** In order to minimize the chances of a breach of confidentiality, no names or email addresses will be collected from participants. Participant demographics will be collected, but only reported in aggregate form. Further, a approved platform, Qualtrics, is being used for survey administration to reduce participant exposure.

Collected data will be stored on an approved platform that is sponsored by Oregon State University, Qualtrics. Data will be stored for a minimum of seven (7) years post-study termination in electronic format on the same OSU-sponsored platform. In addition to OSU cloud storage, data may be downloaded and stored in password-protected electronic devices. Participant data will be reported in aggregate in future publications with no identifying information shared. Aggregate data may also be shared with original instrument developers per licensing or written agreement.

Names and emails as entries for the raffle will be collected separately from the actual study data. Study data will not be linked to any identifiers. Study data is collected anonymously and not linked to raffle names/emails.

**Who do I contact if I have questions?** If you have any questions, as a prospective participant or as a participant, please contact Dr. Kok- Mun Ng at kokmun.ng@oregonstate.edu or 1.541.737.3741, or Gideon Litherland at litherlg@oregonstate.edu or 1.630.212.1128. If you have questions about your rights or welfare as a participant, please contact the Oregon State University Human Research Protection Program (HRPP) office, at 1.541.737.8008 or by emails at IRB@oregonstate.edu

Please print the page and/or take a screenshot if you would like to retain a copy of this informed consent for your records.

---

Please answer these two questions before you can access the research materials.

I have read the information provided, any questions have been answered, and I agree to participate.

- ○ Yes, I have read the information provided, my questions have been answered, and I agree to participate.
- ○ No, I do not want to participate.

**Informed Consent Agreement**

By clicking the button below, you acknowledge (a) that your participation in the study is voluntary, (b) you are 18 years of age and meet the participation criteria, (c) you have reviewed the provided materials describing the scope of this research and the voluntary nature of your participation, and (d) that you are aware that you may choose to terminate your participation in the study at any time and for any reason, without consequence.

**Your participation in this research is voluntary. You have the right to withdraw at any point during the study, for any reason, and without any prejudice.**

**Please note that this survey will be best displayed on a laptop or desktop computer. Some features may be less compatible for use on a mobile device.

Do you consent to participate in this study?

- ○ I do.
- ○ I do not.
- ○ I am not yet ready to participate. I have further questions that I will follow up with the researcher.

Which of the following CACREP-accredited program tracks are you currently enrolled in?

- ○ Addictions Counseling
- ○ Clinical Mental Health Counseling
- ○ Clinical Rehabilitation Counseling
- ○ Marriage, Couple, and Family Counseling
- ○ I am not currently enrolled in one of those tracks

Based on the response you selected, you are not eligible to participate in this survey at this time.

Thank you for taking time to consider participating. Feel free to follow up with us with any questions you may have:

Kok-Mun Ng
kokmun.ng@oregonstate.edu

Gideon Litherland
litherlg@oregonstate.edu

**Demographics**

Please indicate your stage of clinical training within your current program

- ○ I am not currently completing hours at a practicum or internship placement.
- ○ I am currently completing my practicum or internship hours at a site placement.

Based on the response you selected, you are not eligible to participate in this survey at this time.

Thank you for taking time to consider participating. Feel free to follow up with us with any questions you may have:

Kok-Mun Ng
kokmun.ng@oregonstate.edu

Gideon Litherland
litherlg@oregonstate.edu

How do you self-identify in regards to gender?

- ○ Female
- ○ Male
- ○ Gender non-binary

Do you self-identify as a sexual minority?

- ○ Yes.
- ○ No.
- ○ Prefer not to disclose.

Please indicate your age, in years.

Please indicate your race/ethnicity.

- ☐ Caucasian/European/White
- ☐ Black/African
- ☐ American Indian or Alaska Native
- ☐ Asian
- ☐ Native Hawaiian or Pacific Islander
- ☐ Latinx/Hispanic/Spanish
- ☐ Multiracial
- ☐ Other
- ☐ Prefer to not disclose

Are you an international student?

- ○ I am an international student.
- ○ I am not an international student.
- ○ I prefer to not disclose.

Please indicate your program's region within the US.

- ○

North Atlantic (Connecticut, Delaware, District of Columbia, Maine, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, Vermont)

- North Central (Illinois, Indiana, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, Oklahoma, South Dakota, Wisconsin)
- Rocky Mountain (Colorado, Idaho Montana, New Mexico, Utah, Wyoming)
- Southern (Alabama, Arkansas, Florida, Georgia, Kentucky, Louisiana, Maryland, Mississippi, North Carolina, South Carolina, Tennessee, Texas, Virginia, West Virginia)
- Western (Alaska, Arizona, California, Hawaii, Nevada, Oregon, Washington)
- Outside of regional boundaries

Please indicate the instructional environment - learning delivery - of your program

- Hybrid (Definition: 30-50% of your coursework is delivered via the internet with some face to face, in-person interactions with peers and faculty)
- Fully Online (Definition: 50% or more of your coursework is delivered via the internet)
- Traditional, In-Person, On Ground (Definition: Less than 30% of coursework delivered via the internet)

How many practicum and internship hours have you **approximately** completed in total, at the time of taking this survey? (If in practicum, just practicum hours. If in internship, add practicum and internship hours thus far.)

Instructions for all proceeding questions:

To meaningfully respond to the following questions, think about your clinical supervision and the **current supervisor** with whom you are working. Depending on your program, clinical supervision may be provided off-site by non-University affiliated staff or, even, on-site by your faculty within a University clinic. In either case, the following questions are about your current supervision experience.

Please answer the following set of questions about you and your **current (or most recent) individual supervisor**. Again, remember if you have MORE THAN ONE individual supervisor in your current (or most recent) clinical training setting, please choose one and use the same one throughout the survey.

*Friendly reminder:* Your participation is confidential and your responses are not shared with your school.

Based on the supervisor you selected, please indicate your supervisors' affiliation:

- My supervisor is a faculty member.
- My supervisor is a doctoral student.
- My supervisor is at my site (site supervisor).
- Other:

What setting does supervision take place with this supervisor?

- ○ University-affiliated clinic
- ○ Community mental health center/Agency
- ○ Private practice office
- ○ Group practice office
- ○ Telesupervision via videoconference

Please indicate the theoretical orientation(s) that **you** ascribe to or practice from

- ☐ Psychodynamic
- ☐ Cognitive-Behavioral
- ☐ Humanistic
- ☐ Systems
- ☐ Interpersonal
- ☐ Eclectic
- ☐ Other not listed: (please write-in)

Please indicate the theoretical orientation(s) that you think (or know) **your supervisor** ascribes to or practices from

- ☐ Psychodynamic
- ☐ Cognitive-Behavioral
- ☐ Humanistic
- ☐ Systems
- ☐ Interpersonal
- ☐ Eclectic
- ☐ Other not listed: (please write-in)

What type of supervision you receive with this supervisor?

- ○ Individual supervision
- ○ Triadic supervision
- ○ Group supervision
- ○ Mix of individual and group

How long have you been receiving clinical supervision with this supervisor?

Year(s)

Month(s)

On average, how often are your clinical supervision sessions?

- ○ Weekly
- ○ Every two (2) weeks
- ○ Monthly
- ○ 2-3 months
- ○ >3 months

On average, how long are your clinical supervision sessions?

- ○ <15 minutes
- ○ 15-30 minutes
- ○ 31-45 minuts
- ○ 46-60 minutes
- ○ >60 minutes

**Research Instruments**

Instructions:

Please answer the following set of questions about you and your **current (or most recent) individual supervisor**. Again, remember if you have MORE THAN ONE individual supervisor in your current (or most recent) clinical training setting, please choose one and use the same one throughout the survey.

Please describe the degree to which supervision with **your current (or most recent) individual supervisor** has contributed to the *IMPROVEMENT* of the following:

| | Not helpful at all (1) | Helpful, but very little (2) | Somewhat helpful (3) | Very helpful (4) | Extremely helpful (5) |
|---|---|---|---|---|---|
| Client symptoms (decrease in symptoms) | ○ | ○ | ○ | ○ | ○ |
| Your relationship with clients | ○ | ○ | ○ | ○ | ○ |
| Your counseling skills | ○ | ○ | ○ | ○ | ○ |
| Your case conceptualization ability | ○ | ○ | ○ | ○ | ○ |
| Your multicultural counseling skills (e.g., skills that are culturally appropriate in working with diverse clients) | ○ | ○ | ○ | ○ | ○ |
| Your multicultural beliefs/attitudes/awareness (e.g., awareness of your own worldviews) | ○ | ○ | ○ | ○ | ○ |
| Your multicultural knowledge (e.g., knowledge of worldviews of culturally different clients) | ○ | ○ | ○ | ○ | ○ |

Instructions:

Please answer the following set of questions about you and your **current (or most recent) individual supervisor**.

Drawing on your **current experience of receiving Clinical Supervision (CS)** at your workplace, indicate your level of agreement with the following 26 statements by ticking the box which best represents your answer. 0 means you strongly disagree, 1 means you disagree, 2 means you have no opinion, 3 means you agree, 4 means you strongly agree.

Do not spend too long thinking about each question; your first response is probably the best one.

| | Strongly disagree (0) | Disagree (1) | No opinion (2) | Agree (3) | Strongly agree (4) |
|---|---|---|---|---|---|
| Other work pressures interfere with CS sessions | ○ | ○ | ○ | ○ | ○ |
| It is difficult to find the time for CS sessions | ○ | ○ | ○ | ○ | ○ |
| CS sessions are not necessary/don't solve anything | ○ | ○ | ○ | ○ | ○ |
| Time spent on CS takes me away from my real work in the clinical area | ○ | ○ | ○ | ○ | ○ |
| Fitting CS sessions in can lead to more pressure at work | ○ | ○ | ○ | ○ | ○ |
| I find CS sessions time consuming | ○ | ○ | ○ | ○ | ○ |
| My supervisor gives me support and encouragement | ○ | ○ | ○ | ○ | ○ |
| CS sessions are intrusive | ○ | ○ | ○ | ○ | ○ |
| CS gives me time to 'reflect' | ○ | ○ | ○ | ○ | ○ |
| Work problems can be tackled constructively during CS sessions | ○ | ○ | ○ | ○ | ○ |
| CS sessions facilitate reflective practice | ○ | ○ | ○ | ○ | ○ |
| My supervisor offers an 'unbiased' opinion | ○ | ○ | ○ | ○ | ○ |
| I can discuss sensitive issues encountered during my clinical casework with my supervisor | ○ | ○ | ○ | ○ | ○ |
| My CS sessions are an important part of my work routine | ○ | ○ | ○ | ○ | ○ |
| I learn from my supervisor's experiences | ○ | ○ | ○ | ○ | ○ |
| It is important to make time for CS sessions | ○ | ○ | ○ | ○ | ○ |
| My supervisor provides me with valuable advice | ○ | ○ | ○ | ○ | ○ |
| My supervisor is very open with me | ○ | ○ | ○ | ○ | ○ |
| Sessions with my supervisor widen my clinical knowledge base | ○ | ○ | ○ | ○ | ○ |
| CS is unnecessary for experienced/established staff | ○ | ○ | ○ | ○ | ○ |
| My supervisor acts in a superior manner during our sessions | ○ | ○ | ○ | ○ | ○ |
| Clinical supervision makes me a better practitioner | ○ | ○ | ○ | ○ | ○ |
| CS sessions motivate staff | ○ | ○ | ○ | ○ | ○ |
| I can widen my skill base during my CS sessions | ○ | ○ | ○ | ○ | ○ |
| My supervisor offers me guidance with patient/client care | ○ | ○ | ○ | ○ | ○ |
| I think receiving clinical supervision improves the quality of care I give | ○ | ○ | ○ | ○ | ○ |

Instructions:

Please answer the following set of questions about you and your **current (or most recent) individual supervisor**.

Use the following Likert scale to evaluate the supervision you received by your primary supervisor (individual and group) at the placement you just completed (or are currently in).

| | Not at all, Strongly disagree (1) | (2) | (3) | Moderately, Neutral (4) | (5) | (6) | Very much so, Strongly agree (7) |
|---|---|---|---|---|---|---|---|
| Overall, my expectations of supervision were matched or exceeded | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Overall, I would gladly recommend this supervisor to others | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Overall, supervision significantly enhanced my competence as a practitioner and professional | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Overall, supervision significantly contributed to my achieving better outcomes for my clients | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| In day-to-day dealings, I got along well with the supervisor | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The supervisor was understanding and open to a sharing of ideas | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The supervisor was accepting of my mistakes and inadequacies | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The supervisor was caring and supportive | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The supervisor was approachable and interested in my personal and professional development | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The supervisor impressed me as a skilled therapist | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The supervisor was knowledgeable and could communicate theoretical concepts clearly | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The supervision plan appropriately reflected important clinical competencies | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Supervision objectives were in accordance with my level of professional development | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The supervisor organized and managed supervision efficiently | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Supervision methods were varied to match supervision objectives | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Supervision objectives (goals) were negotiated and clearly articulated | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Supervision sessions were structured and supervision activities were goal driven | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| I felt comfortable discussing my professional inadequacies in supervision | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The supervisor was sensitive to my emotional and self-care needs | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

| | Not at all, Strongly disagree (1) | (2) | (3) | Moderately, Neutral (4) | (5) | (6) | Very much so, Strongly agree (7) |
|---|---|---|---|---|---|---|---|
| Supervision facilitated emotional ventilation and support as appropriate | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The supervisor enhanced my abilities to reflect on my clinical work | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The supervision sessions enhanced my self awareness as a person | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The supervision furthered my understanding of my own positive and negative interaction patterns with clients | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The supervisor helped me gain an understanding of my emotional reactions within therapy | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The supervisor helped inspire me to remain excited about my clinical work and professional responsibilities | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The supervision advanced my therapist-client relationship skills | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

Instructions:

Please answer the following set of questions about you and your **current (or most recent) individual supervisor**.

The purpose of the CTES is to assess your perceptions and experiences of the counseling training environment in the counseling and related mental health training program you are attending right now.

The items will assess your perceptions about what your current training environment is actually like. Please read each item and using the 5-point Likert-type scale (1 = Strongly Disagree [SD]; 2= Disagree [D]; 3 = Agree [A]; 4 = Strongly Agree [SA]; 5 = Not Applicable [NA]), rate your level of agreement with each item by selecting the appropriate number.

*In my counseling training program…*

| | Strongly disagree | Disagree | Agree | Strongly agree | Not Applicable |
|---|---|---|---|---|---|
| Questions from students are welcomed in all my classes | ○ | ○ | ○ | ○ | ○ |
| I get regular feedback from my professors | ○ | ○ | ○ | ○ | ○ |
| My clinical site supervisor treats me with respect | ○ | ○ | ○ | ○ | ○ |
| My clinical site supervisor creates a safe environment for the discussion of difficult topics | ○ | ○ | ○ | ○ | ○ |
| Students have access to University/college resources to facilitate learning and training (e.g., writing labs) | ○ | ○ | ○ | ○ | ○ |
| Our program has a good relationship with the local community | ○ | ○ | ○ | ○ | ○ |
| Skills and knowledge gained in my classes are relevant to the work I am doing at my clinical field placement | ○ | ○ | ○ | ○ | ○ |

| | Strongly disagree | Disagree | Agree | Strongly agree | Not Applicable |
|---|---|---|---|---|---|
| University/college procedures and department procedures for addressing student grievances are consistent | ○ | ○ | ○ | ○ | ○ |
| Program faculty are active in addressing issues that arise at my clinical field experience site | ○ | ○ | ○ | ○ | ○ |
| Students are kept abreast of the mental health needs of the community | ○ | ○ | ○ | ○ | ○ |
| Students are made aware of opportunities to volunteer in community activities | ○ | ○ | ○ | ○ | ○ |
| My training helps me become cognizant of the impact that my background and life experiences have on my clients and how these may affect my clients | ○ | ○ | ○ | ○ | ○ |
| Faculty incorporate their clinical experiences into the classroom training | ○ | ○ | ○ | ○ | ○ |
| Faculty are well-connected within the profession | ○ | ○ | ○ | ○ | ○ |
| My clinical site supervisor shares clinical resources with me | ○ | ○ | ○ | ○ | ○ |
| Training curricula meets state standards for professional licensure and/or certification | ○ | ○ | ○ | ○ | ○ |
| An emphasis is placed on adhering to the ethical codes set forth by the profession | ○ | ○ | ○ | ○ | ○ |
| We are taught to recognize both within-group and between-group differences | ○ | ○ | ○ | ○ | ○ |
| My knowledge, awareness, and skills in multicultural counseling has been challenged | ○ | ○ | ○ | ○ | ○ |
| The program has helped me become mindful of my personal development through time | ○ | ○ | ○ | ○ | ○ |
| The program is intentional in facilitating students' growth and development | ○ | ○ | ○ | ○ | ○ |
| My training curricula reflects the current trends of the profession | ○ | ○ | ○ | ○ | ○ |
| My training is current and reflective of the issues impacting our society today | ○ | ○ | ○ | ○ | ○ |

Instructions:

Listed below are a number of statement concerning personal attitudes and traits. Read each item and decide how it pertains to you.

Please respond either TRUE (**T**) or FALSE (**F**) to each item. Indicate your response by selecting the appropriate letter next to the item. Be sure to answer all items.

| | TRUE | FALSE |
|---|---|---|
| It is sometimes hard for me to go on with my work if I am not encouraged. | ○ | ○ |
| I sometimes feel resentful when I don't get my way. | ○ | ○ |
| There have been occasions when I took advantage of someone. | ○ | ○ |

| | TRUE | FALSE |
|---|---|---|
| I'm always willing to admit it when I make a mistake. | ○ | ○ |
| I sometimes try to get even rather than forgive and forget. | ○ | ○ |
| I am always courteous, even to people who are disagreeable. | ○ | ○ |
| I have never been irked when people expressed ideas very different from my own. | ○ | ○ |
| There have been times when I was quite jealous of the good fortune of others. | ○ | ○ |
| I am sometimes irritated by people who ask favors of me. | ○ | ○ |
| I have never deliberately said something that hurt someone's feelings. | ○ | ○ |

Instructions:
Please answer the following set of questions about you and your **current (or most recent) individual supervisor**.

Please indicate the frequency with which the behavior described in each of the following items seems characteristic of your work with your supervisor. After each item, check the space over the number corresponding to the appropriate point of the following 7-point scale: 1 = Almost Never; 7 = Almost Always.

| | Almost Never (1) | (2) | (3) | (4) | (5) | (6) | Almost Always (7) |
|---|---|---|---|---|---|---|---|
| I feel comfortable working with my supervisor. | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| My supervisor welcomes my explanations about the client's behavior. | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| My supervisor makes the effort to understand me. | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| My supervisor encourages me to talk about my work with clients in ways that are comfortable for me. | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| My supervisor is tactful when commenting about my performance. | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| My supervisor encourages me to formulate my own interventions with the client. | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| My supervisor helps me talk freely in our sessions. | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| My supervisor stays in tune with me during supervision. | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| I understand client behavior and treatment technique similar to the way my supervisor does. | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| I feel free to mention to my supervisor any troublesome feelings I might have about him/her. | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| My supervisor treats me like a colleague in our supervisory sessions. | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| In supervision, I am more curious than anxious when discussing my difficulties with clients. | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| In supervision, my supervisor places a high priority on our understanding the client's perspective. | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| My supervisor encourages me to take time to understand what the client is saying and doing. | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

| | Almost Never (1) | (2) | (3) | (4) | (5) | (6) | Almost Always (7) |
|---|---|---|---|---|---|---|---|
| My supervisor's style is to carefully and systematically consider the material I bring to supervision. | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| When correcting my errors with a client, my supervisor offers alternative ways of intervening with that client. | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| My supervisor helps me work within a specific treatment plan with my clients. | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| My supervisor helps me stay on track during our meetings. | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| I work with my supervisor on specific goals in the supervisory session. | ○ | ○ | ○ | ○ | ○ | ○ | ○ |



**Block 3**

Thank you for completing the survey. Would you like to enter a drawing for a $20 Starbucks gift card?

Participating in the drawing requires you to provide your name and email address. However, to maintain your confidentiality and the anonymity of your research participation, your responses to the research materials will not be matched to your name and email address. All participants who chose to enter the drawing will be informed electronically whether they are selected to receive a gift card.

○ Yes

○ No

Please follow this link to enter the drawing.

*Note*: You will be taken away from this secure site. Your responses to the survey questions will not be matched or linked to your drawing entry in order to maintain confidentiality. You will complete a GoogleForm to enter the drawing.

**https://forms.gle/A2AGxQBDhxHReV4TA**

## Appendix C

Sociodemographic Characteristics of Participants

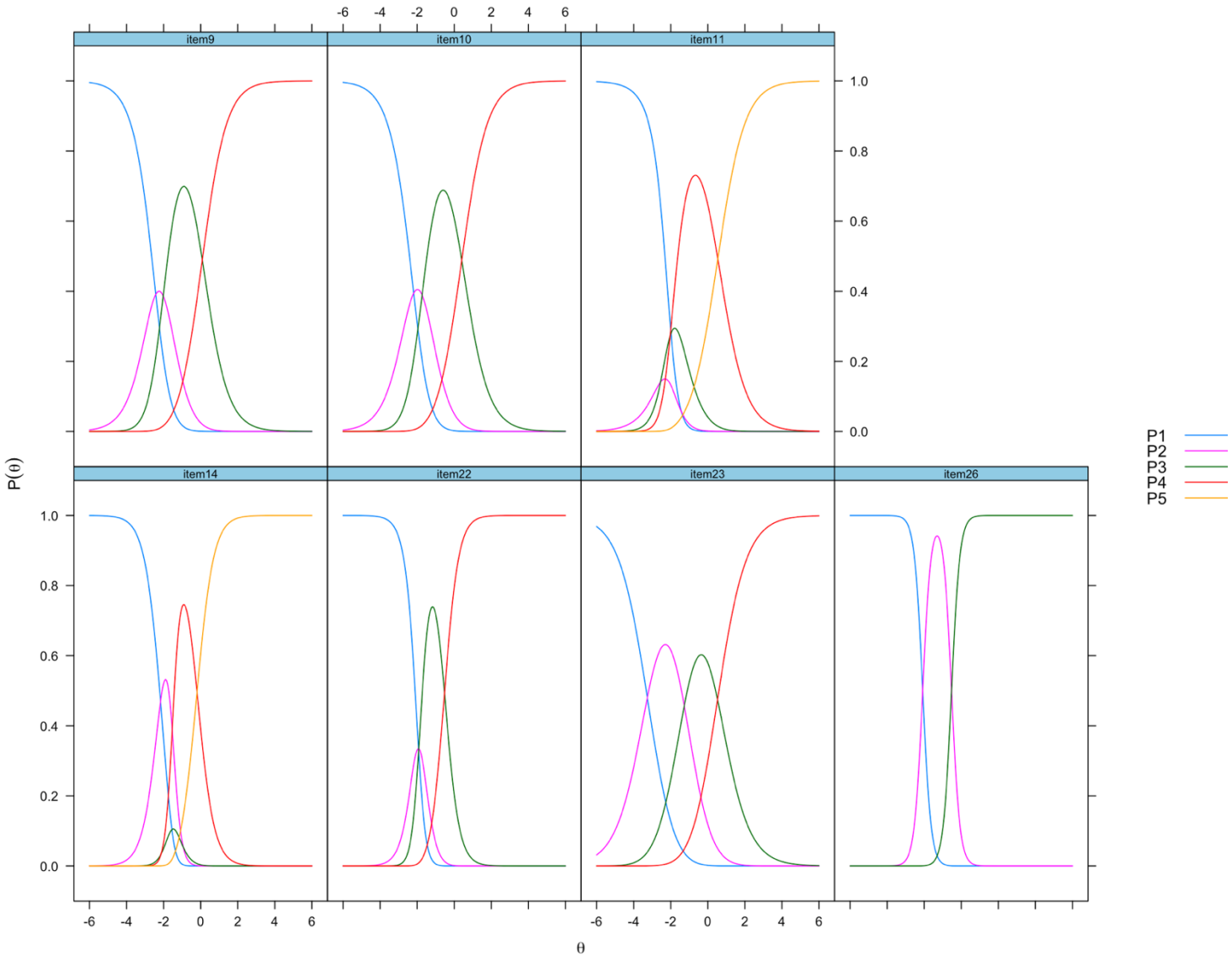
| Sample characteristic | *n* | % |
|---|---|---|
| Gender | | |
| Female | 71 | 88 |
| Male | 8 | 10 |
| Non-Binary | 2 | 2 |
| Race | | |
| Caucasian/European/White | 67 | 83 |
| Asian | 4 | 5 |
| Black/African | 3 | 4 |
| Latinx/Hispanic/Spanish | 3 | 4 |
| Multiracial | 3 | 4 |
| Age Range | | |
| 22-29 y.o. | 53 | 70 |
| 30-39 y.o. | 16 | 21 |
| 40-49 y.o. | 3 | 4 |
| 50-59 y.o. | 3 | 4 |
| 60-69 y.o. | 1 | 1 |
| Sexual Minority | | |
| Yes | 16 | 20 |
| No | 64 | 79 |
| International Student Status | | |
| Domestic | 76 | 94 |
| International | 5 | 6 |
| Program Delivery | | |
| Traditional/In-Person | 59 | 69 |
| Hybrid | 15 | 17 |
| Online | 12 | 15 |

*Note.* $n = 86$

# Appendix D

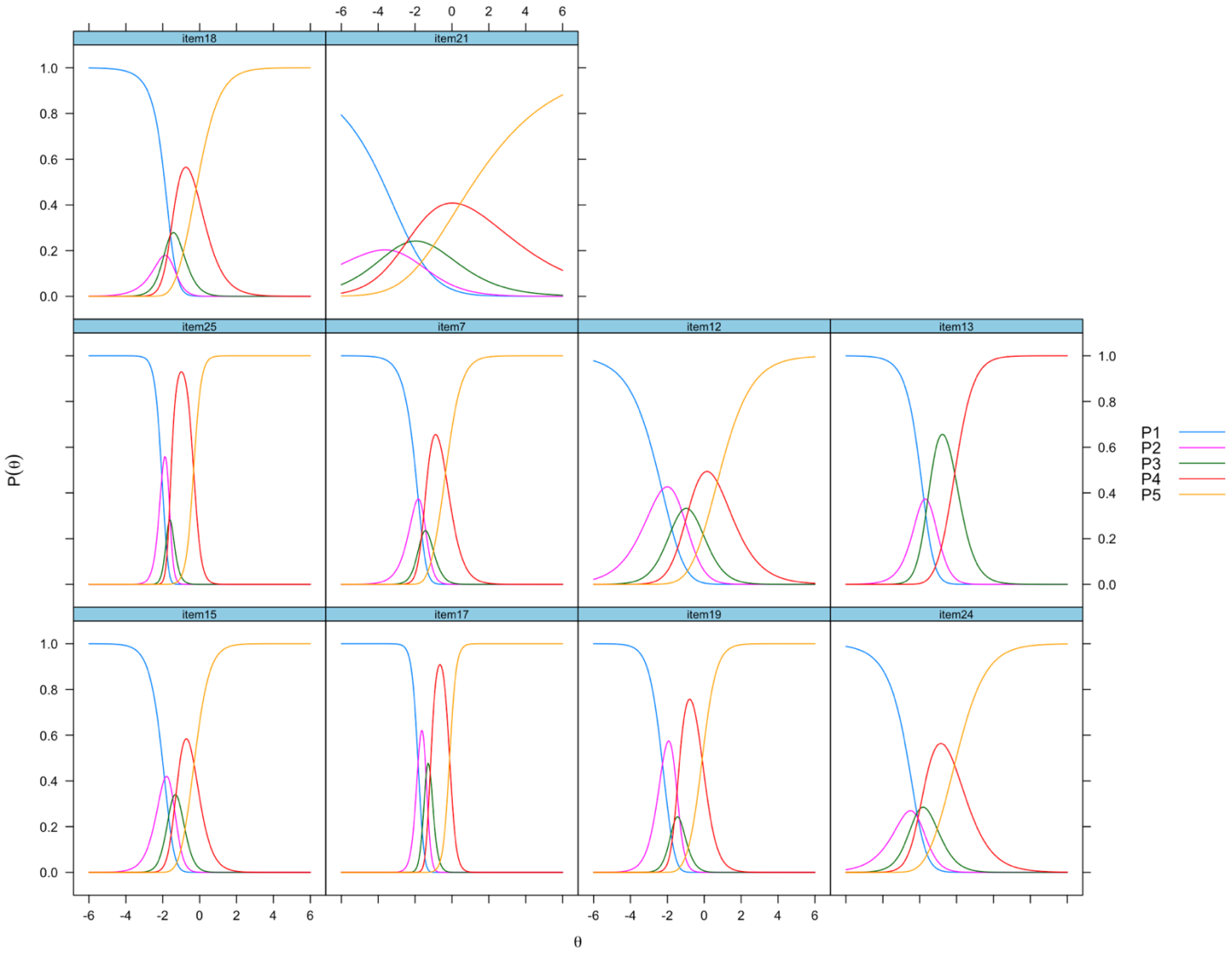
MCSS-26 Item Trace Lines

MCSS Subscale: Formative ($n$=7)



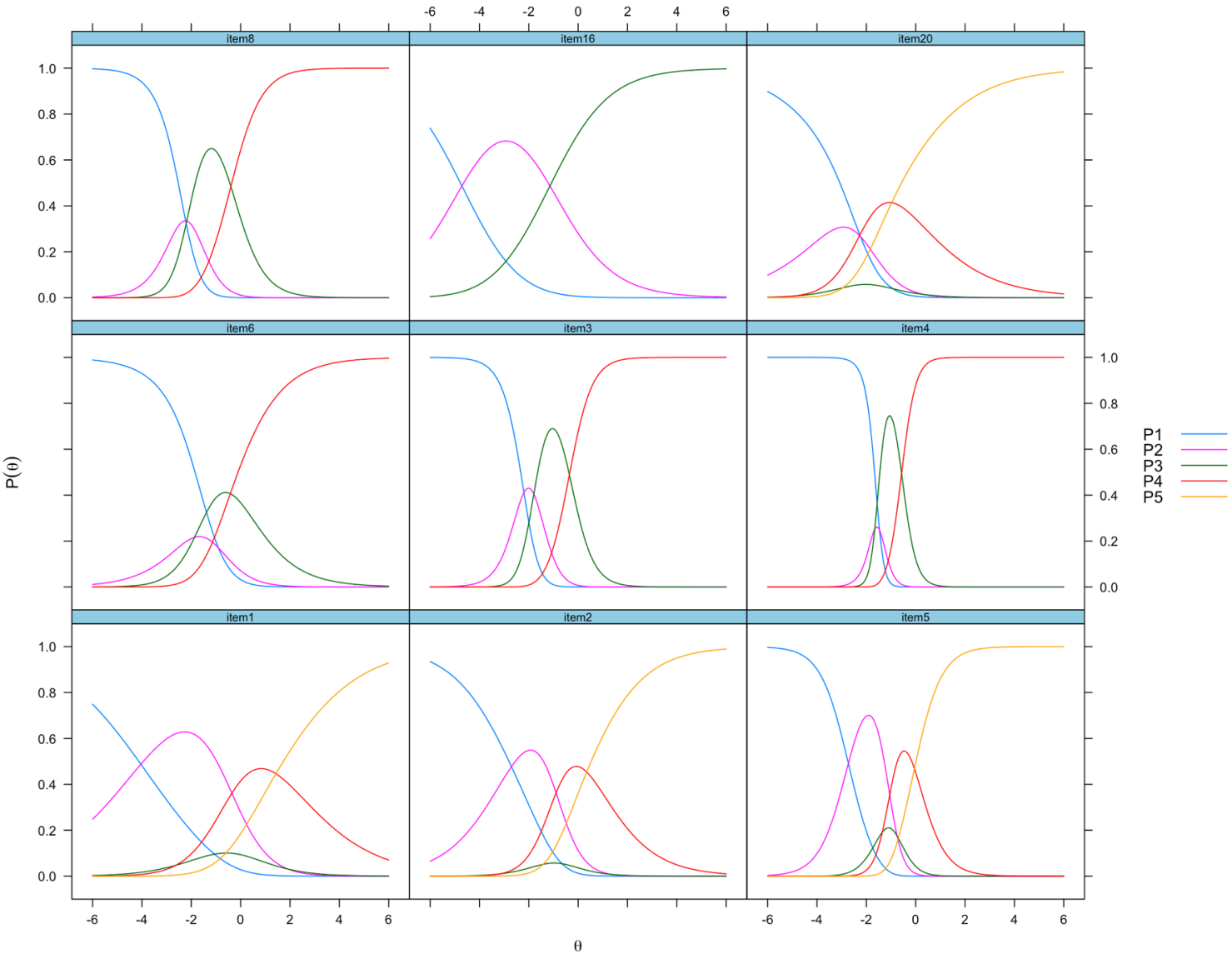
## Appendix D (cont.)

MCSS Subscale: Restorative ($n$=10)



## Appendix D (cont.)

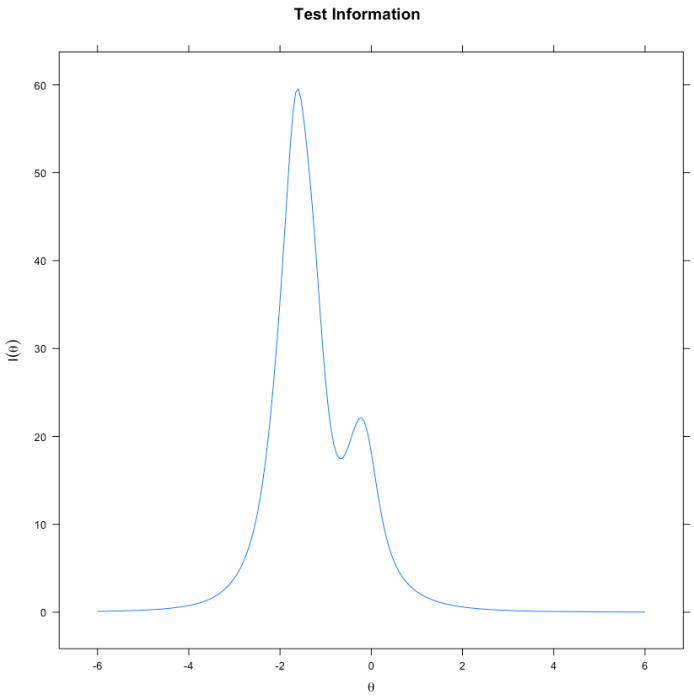
MCSS Subscale: Normative ($n$=9)



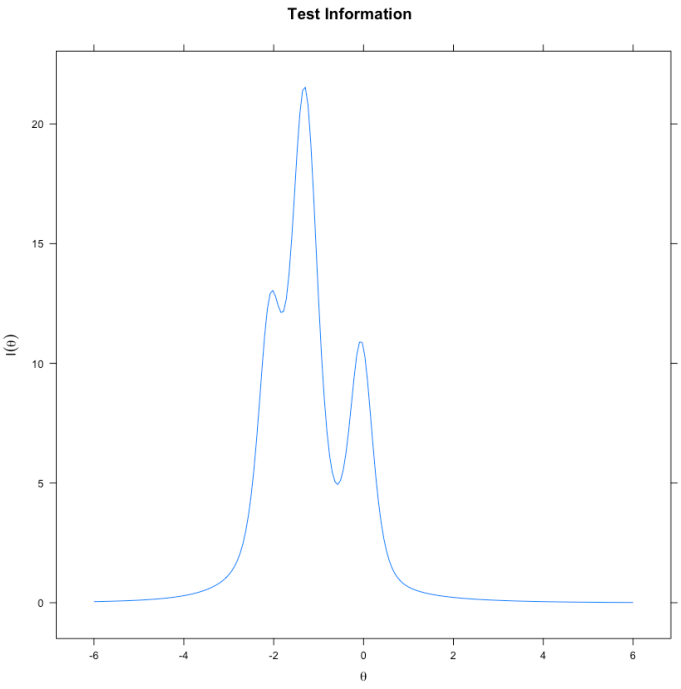
# Appendix E

## MCSS-26 Test Information Curves

### MCSS-26 Rest (Original) Test Information Curve

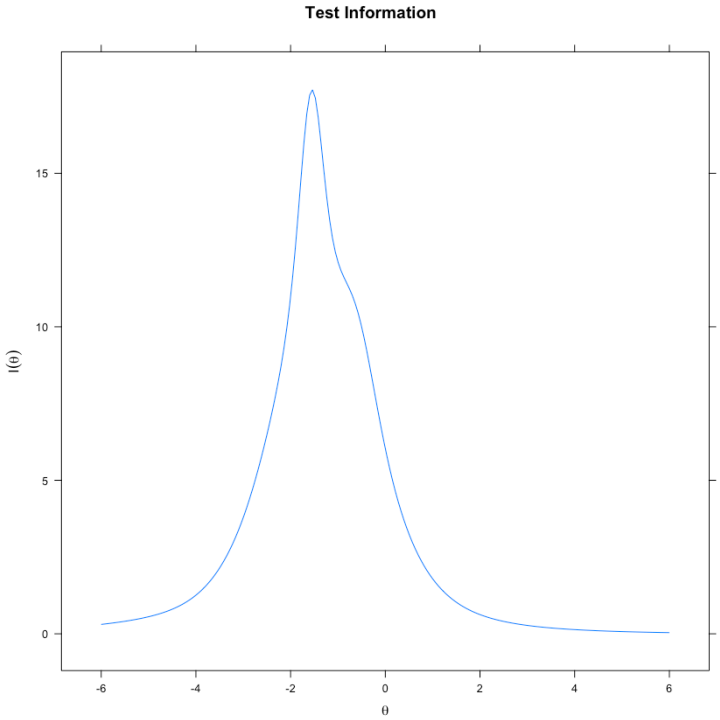



### MCSS-26 Rest (Revised) Test Information Curve

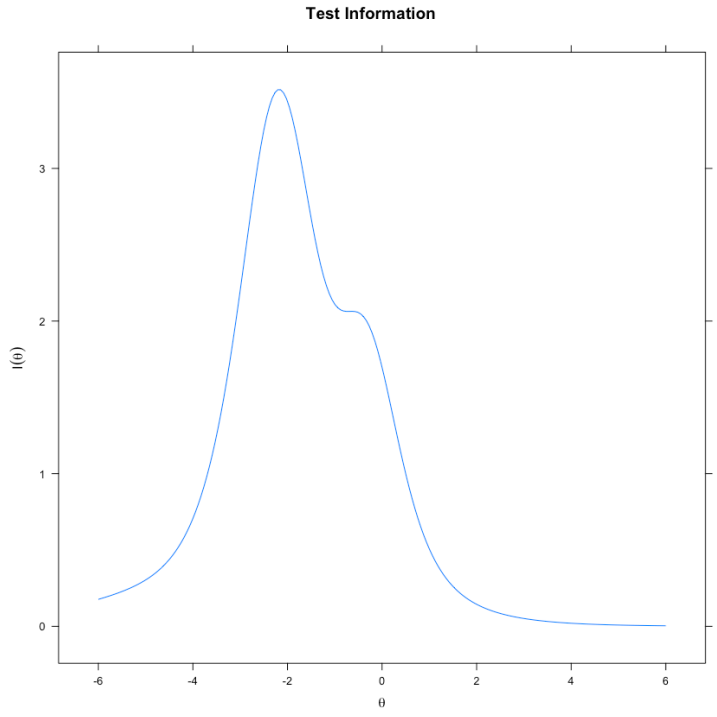


## Appendix E (cont.)

MCSS-26 Norm (Original) Test Information Curve

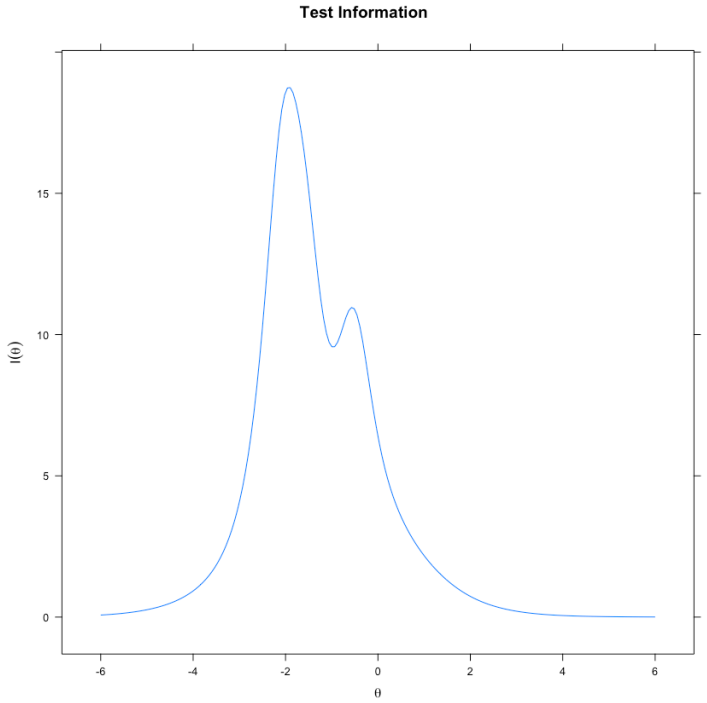



MCSS-26 Norm (Revised) Test Information Curve

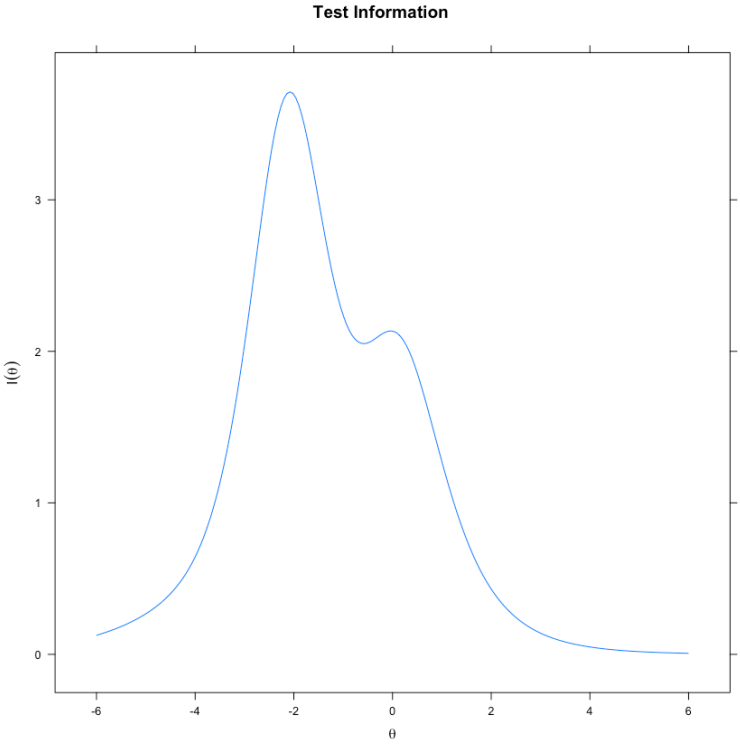


## Appendix E (cont.)

MCSS-26 Form (Original) Test Information Curve




MCSS-26 Form (Revised) Test Information Curve



## Appendix F

Revised MCSS-26, 9-item Instrument

| Strongly Disagree | Disagree | No Opinion | Agree | Strongly Agree |
|---|---|---|---|---|

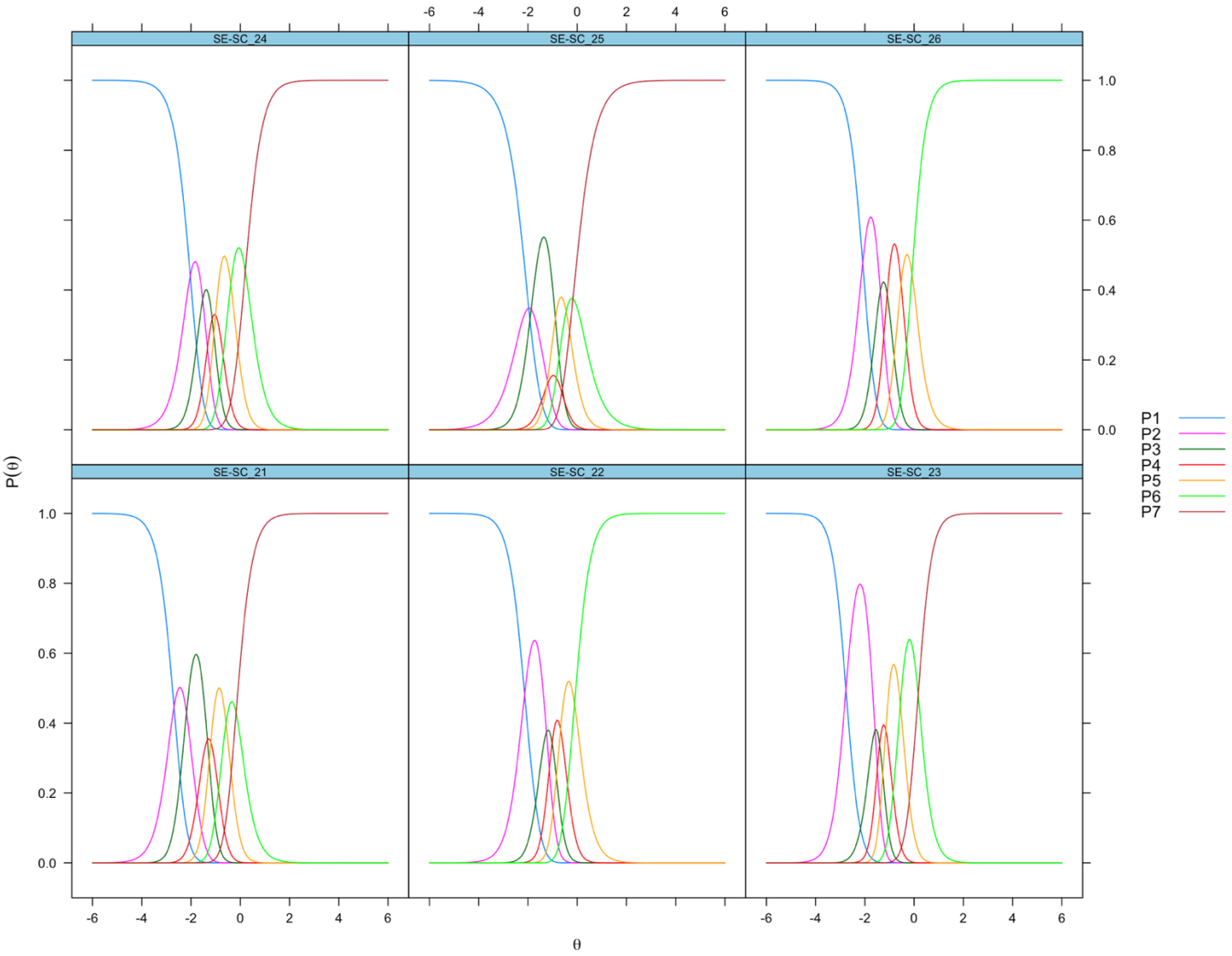
1. CS sessions are not necessary/don't solve anything
2. CS sessions are intrusive
3. CS gives me time to 'reflect'
4. Work problems can be tackled constructively during CS sessions
5. My supervisor offers an 'unbiased' opinion
6. I learn from my supervisor's experiences
7. It is important to make time for CS sessions
8. My supervisor provides me with valuable advice
9. CS sessions motivate staff

Reverse scored = 1, 2

# Appendix G

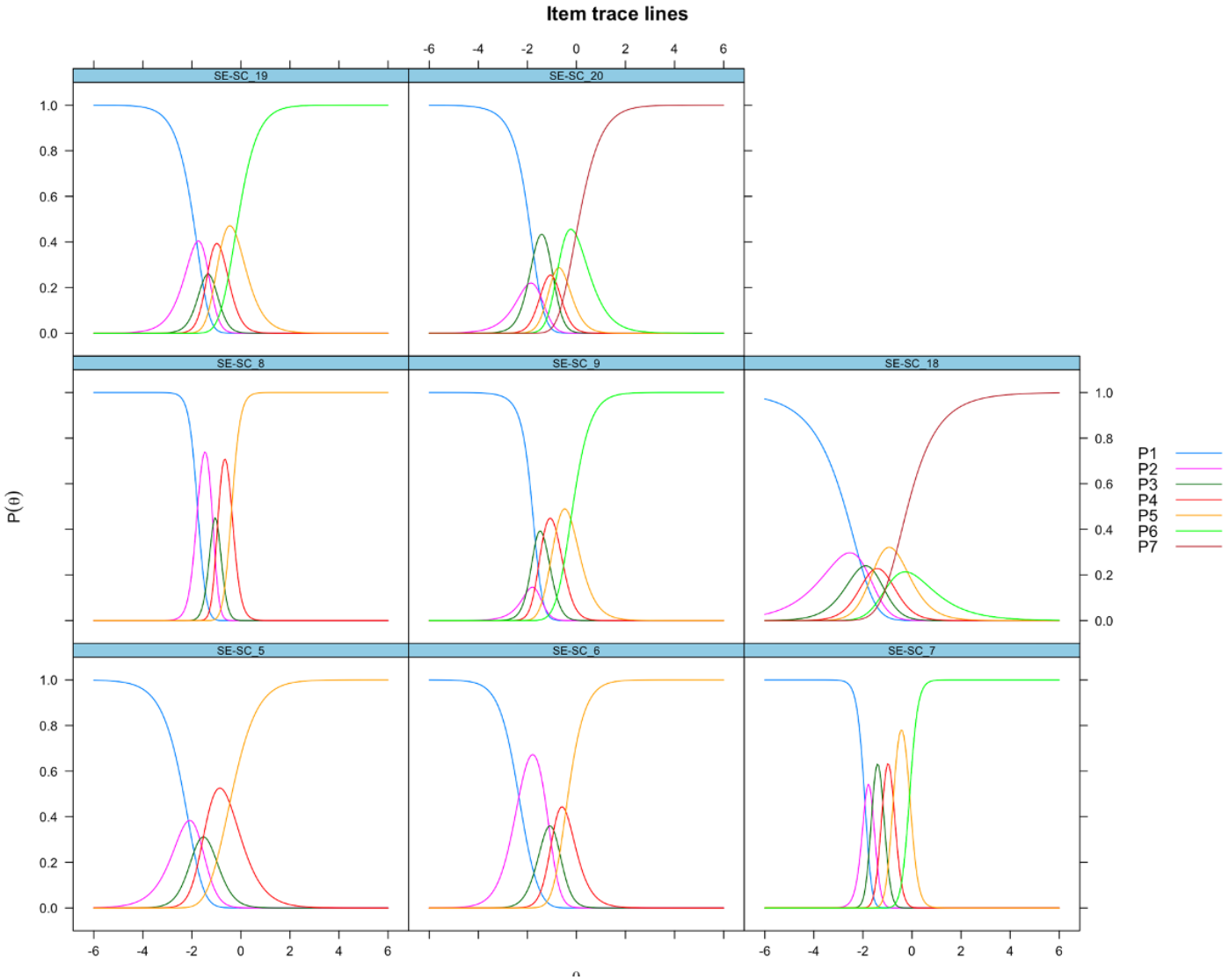
SE-SC Item Trace Lines

SE-SC Subscale: Formative (*n*=6)



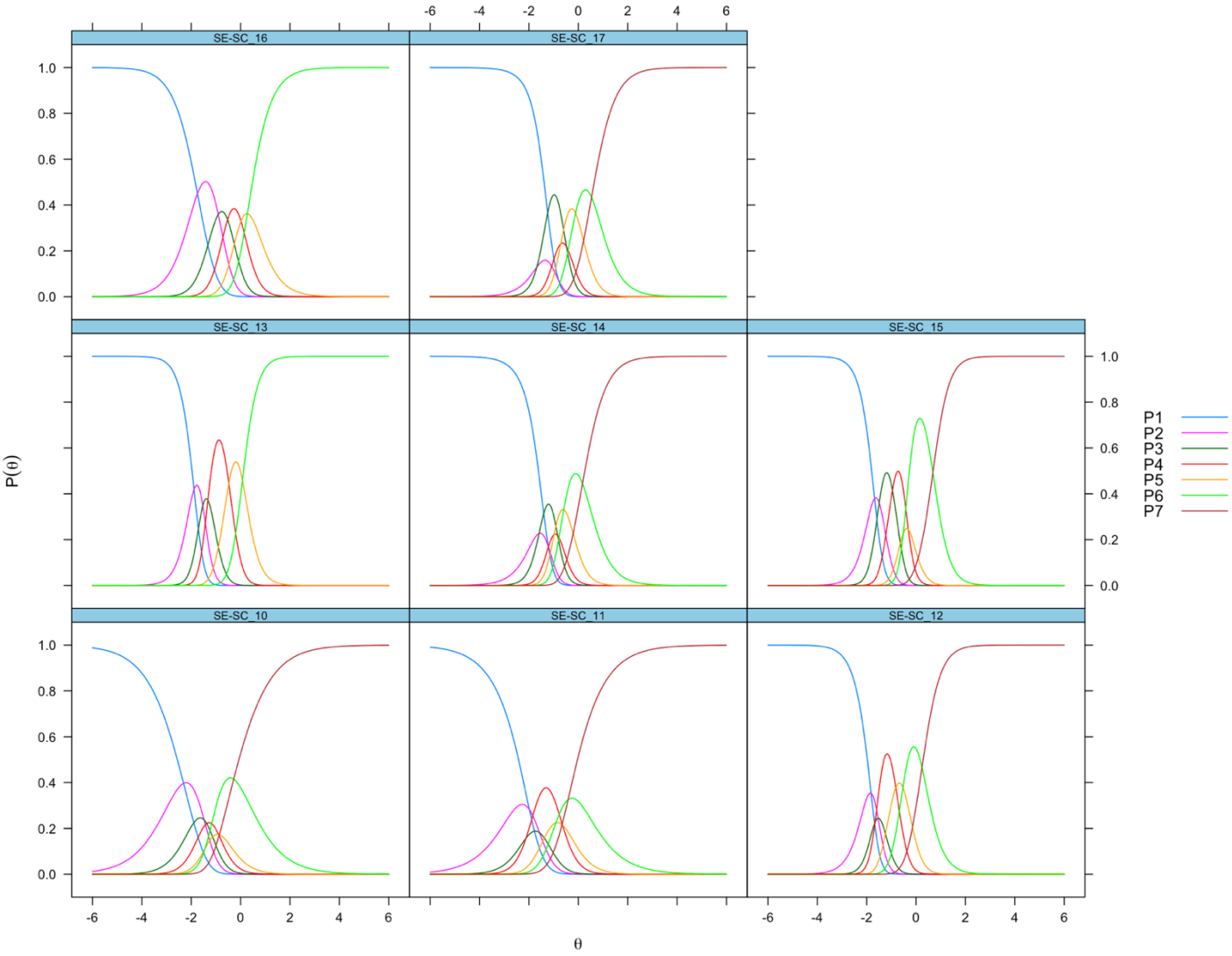
## Appendix G (cont.)

SE-SC Subscale: Restorative (*n*=8)



## Appendix G (cont.)

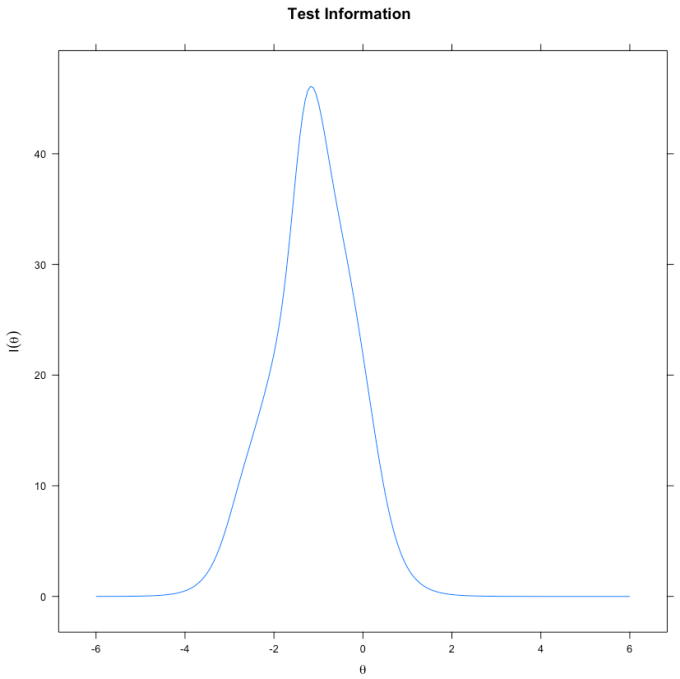
SE-SC Subscale: Normative (*n* =8)



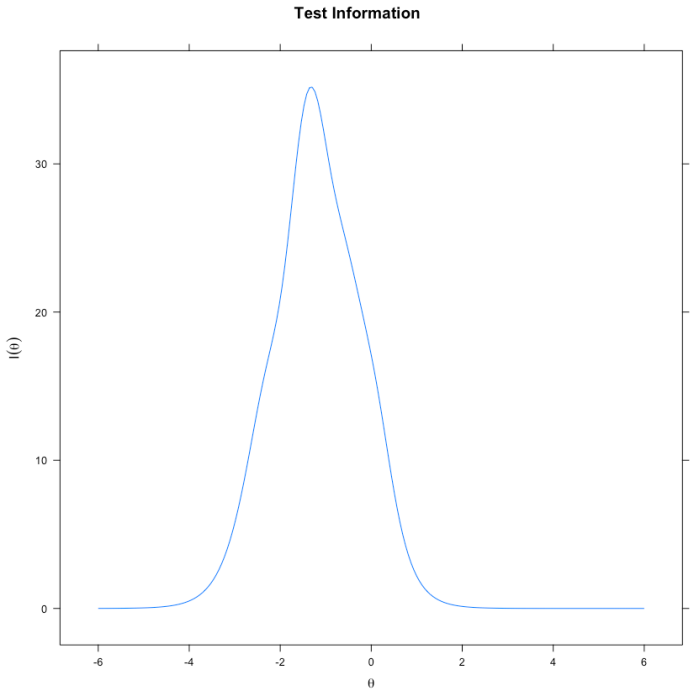
# Appendix H

## SE-SC Test Information Curves

### SE-SC Form (Original) Test Information Curve

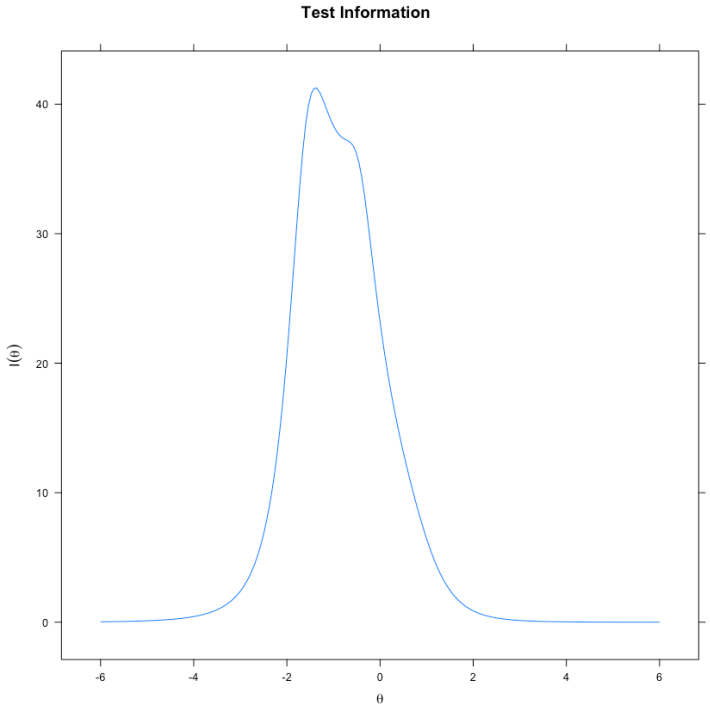



### SE-SC Form (Revised) Test Information Curve

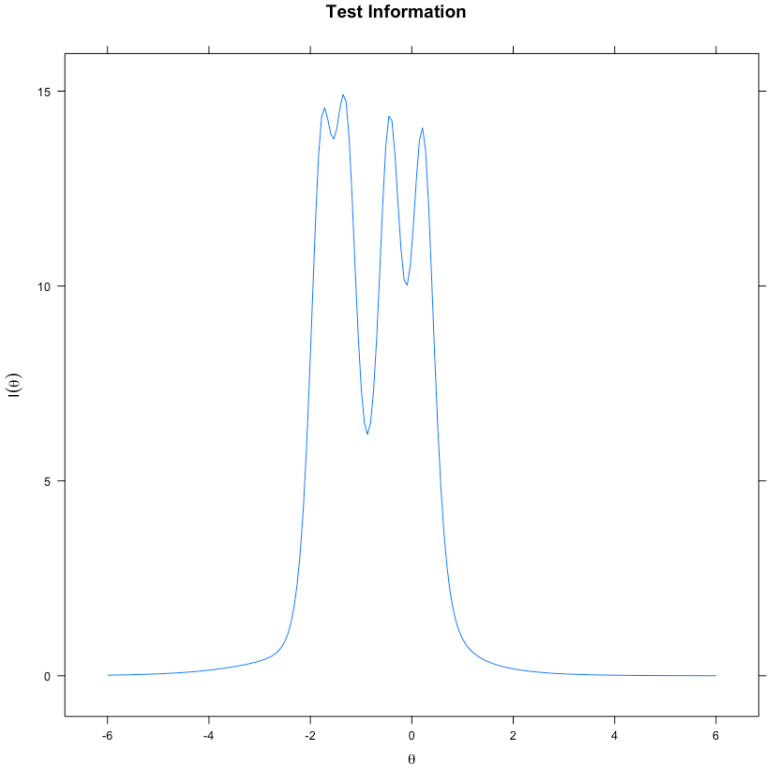


## Appendix H (cont.)

SE-SC Norm (Original) Test Information Curve

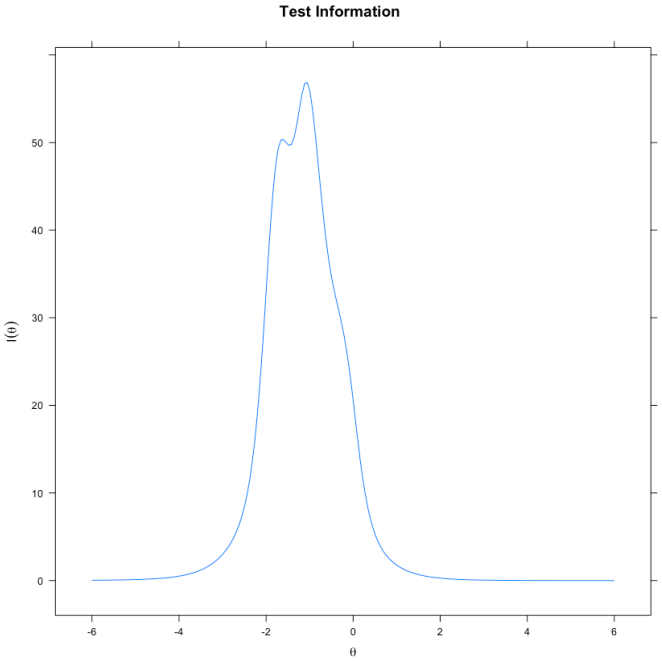



SE-SC Norm (Revised) Test Information Curve

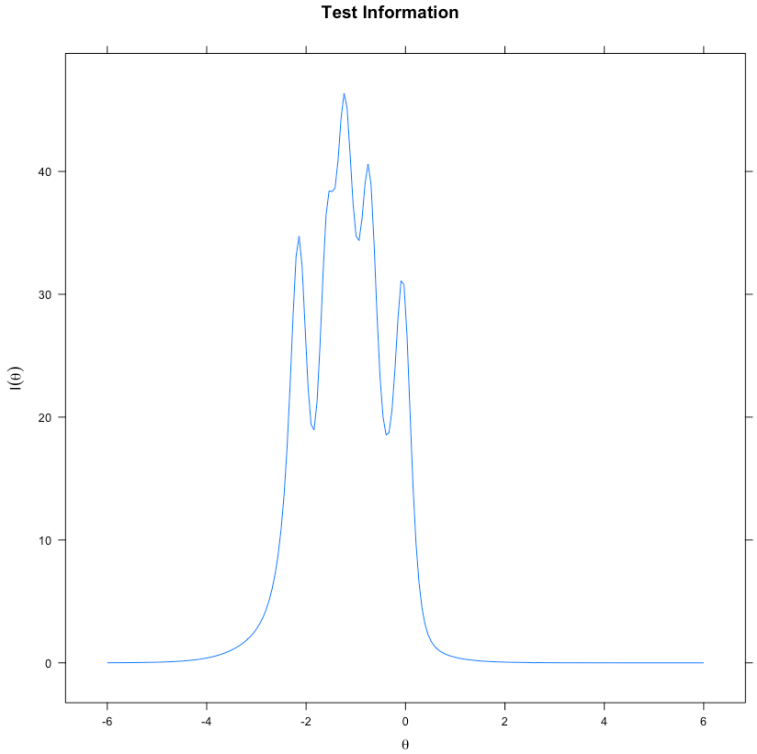


## Appendix H (cont.)

SE-SC Rest (Original) Test Information Curve




SE-SC Rest (Revised) Test Information Curve



## Appendix I

Revised SE-SC, 15-Item Instrument

Use the following Likert scale to evaluate the supervision you received by your primary supervisor (individual and group) at the placement you just completed.

| Not at all, Strongly disagree | | | Moderately, Neutral | | | Very much so, Strongly agree |
|---|---|---|---|---|---|---|
| (1) | (2) | (3) | (4) | (5) | (6) | (7) |

1. Overall, my expectations of supervision were matched or exceeded*
2. Overall, I would gladly recommend this supervisor to others*
3. Overall, supervision significantly enhanced my competence as a practitioner and professional*
4. Overall, supervision significantly contributed to my achieving better outcomes for my clients*
5. In day-to-day dealings, I got along well with the supervisor
6. The supervisor was understanding and open to a sharing of ideas
7. The supervisor was accepting of my mistakes and inadequacies
8. The supervisor was caring and supportive
9. Supervision objectives were in accordance with my level of professional development
10. Supervision objectives (goals) were negotiated and clearly articulated
11. The supervisor enhanced my abilities to reflect on my clinical work
12. The supervision sessions enhanced my self awareness as a person
13. The supervision furthered my understanding of my own positive and negative interaction patterns with clients
14. The supervisor helped me gain an understanding of my emotional reactions within therapy
15. The supervision advanced my therapist-client relationship skills

*Items left out of analysis/scope of research