2016

Development and Validation of the Client Ratings of Counselor Competence: Applying the Rasch Measurement Model

Hang Jo
University of Central Florida

DEVELOPMENT AND VALIDATION OF THE CLIENT RATINGS OF COUNSELOR COMPETENCE
: APPLYING THE RASCH MEASUREMENT MODEL

by

HANG JO
B.A. Hankuk University for Foreign Studies, 2005
M.A. Seoul National University, 2009

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the Department of Child, Family, and Community Sciences
in the College of Education and Human Performance
at the University of Central Florida
Orlando, Florida

Fall Term
2016

Major Professor: K. Dayle Jones

ABSTRACT

An important part of becoming a counselor is developing strong counselor competence, particularly for counselors-in-training. Thus, the main goal in counselor education is to develop students' competence to be capable to practice as a professional counselor. Assessing the competence of counselors-in-training remains the primary focus in counselor education and supervision (Bernard & Goodyear, 2014; McAuliffe & Eriksen, 2011; Swank & Lambie, 2012). There have been various attempts to measure the true construct of counselor competence (e.g., Hughes, 2014; Swank, Lambie, & Witta, 2012; Urbani, Smith et al., 2002). Those attempts tried to involve diverse voices around counselor competence in more comprehensive ways. Although there are numerous measures assessing supervisor ratings of counselor competence, there is still a lack of clients' voice in assessing counselor competence and performance in counselor education literature. In particular, there has been a deficit of direct measures to assess counselor competence *by clients* (Tate et al., 2014). Therefore, a new client-rated scale of counselor competence is required to provide invaluable information for enhancing a counselor's own professional competence as well as the quality of counselor preparation programs. The purpose of this study is to assess the psychometric properties using a Rasch model on a newly developed client-rated scale of counselor competence, named *Client Ratings of Counselor Competence* (CRCC).

For this purpose of this study, the CRCC was developed, following the procedures for a scale development that the Rasch measurement model proposed. The development process consisted of (a) defining hierarchical attributes of what to measure, (b) generating a pool of items

corresponding to the defined attributes, (c) determining the scale-type of measurement, (d) expert reviewing, (f) conducting a field test to a research sample, (g) evaluating the items using Rasch analysis, and (h) determining the final scale. Specifically, the initial pool of 85 items was generated and reduced to 36 items through expert review and a pilot test. The participants in this study were 84 adult clients who received counseling service from counselor trainees in a community counseling center.

This study investigated diverse aspects of validity in the 36-item CRCC using the Rasch model, following the guideline by Wolfe and Smith (2007). In specific, content evidence, substantive evidence, structural evidence, generalizability, and interpretability evidence were investigated with the results of the Rasch analysis.

The result showed that negatively worded items were commonly misfitted to the model. The rating scale analysis result showed that a 3-point rating scale format could be more appropriate than the current 4-point scale. In addition, the investigation of item difficulty hierarchy perceived by clients were mostly consistent with the assumed hierarchical structure in the test specification, empirically supporting *microskills hierarchy* (Ivey et al., 2013). The dimensionality analysis result showed the presence of possible additional dimension in the current CRCC. The reliability level of CRCC was acceptable as well as some bad items functioning differently across gender were detected with the DIF analysis. Additionally, the practicum level counselors-in-training in this study showed higher level of competence above the level that the current CRCC items could measure.

Lastly, implications of the study, limitations, and future research were discussed. Some implications of the findings include: (a) the use of the Rasch model to assess the psychometric

properties of the CRCC scale can make the developing instrument more valid and reliable, overcoming the major weakness of the classical test theory; (b) item difficulty level in the Rasch analysis can be a useful tool to empirically demonstrate whether a theoretical concept or model, especially with hierarchical or developmental structure, exists with real data; (c) the item-person map in the Rasch model can provide useful information for evaluating the instruments as well as interpreting the test scores; and (d) after more revisions and further validation studies, the CRCC could be utilized as additional assessment when counselor educators want to assess whether the trainees develop the competence above the expected level, especially from clients' perspective.

ACKNOWLEDGMENTS

First of all, I would like to thank God for His grace and love with me. He is being with me and filling in what I've needed. I am so grateful to my lovely wife, Misun. I could not start this challenging journey and accomplish it without your constant support and dedication. This is all for you. You are my best friend and partner, ever. I love you!! Next thank is for my lovely princesses, Haewon and Haeri. You are the wonderful gift that God gave me. I am so happy and get some energy to step forward whenever I see your smiling, dancing, even crying. I love you all so much. To my parents, you all are great. Thanks to your praying for me, I was able to accomplish my doctoral degree. I will never forget your unconditional love and support. My mother-in-law, I love you. Thank you for giving me such a wonderful wife to me. To all my friends and relatives, thank you for praying for me and my family.

Thank you Dr. Dayle Jones, you saved my life here. Thank God for meeting you at UCF and having you as my adviser and dissertation chair. I appreciate you for mentoring and advising me with this dissertation as well as teaching, supervision, and everything. I am lucky having such a great mentor as you. Thank you for worrying about my family like your own family during the hurricane Matthew coming. I hope you keep being healthy and happily teaching our UCF students so that they can learn a lot from your passion and mentorship. I also am grateful to my other committee members, Dr. Edward Robinson, Dr. Gulnora Hundley and Dr. Haiyan Bai. Dr. Robinson, I know how much you helped me, sometimes invisibly. Maybe, you should be one of key helpers for letting me be here. Thank you for putting me into this awesome program and supporting me. Dr. Hundley, thank you for your emotional and practical

supports. I stopped by your office whenever I met some struggles in this study. With your advice, I could pass well through them. Thank you Dr. Bai, I could not start developing a new instrument in my dissertation, even my second language if you would not be my committee.

I would like to give my special thanks to Dr. Mark E. Young, Dr. Allen E. Ivey, and Dr. Glenn W. Lambie for participating as the expert reviewers in my research. I feel lucky working with the masters like you all in developing my new instrument. Thank you so much. In addition, I am grateful to Dr. Stacy Van Horn and Dr. Columbus Brand. Thank you for mentoring me in my teaching and supervision practicum. With your special cares, I was able to complete all my practicum hours. Thank you Ms. Joyce Goodman, Judith Montilla, and Lillian Ramos, for your amazing support.

To my friend, Hannah from Ghana, I am missing you. I always thank my Lord for being friend with you. I do not know how well I could express my appreciation to you regarding what you've done for me. You are the great person, friend, and colleague for me. I learned a lot from you when I saw how you treated poor persons and overcame your struggles only with your belief. To my cohort, the UN, you all are awesome. Thank you for being with this journey. I was able to go through all courses thanks to your support and cares. I wish you the best peace and happiness in your new journey. Thank you Scholarly Survivors and Stupendous Seven Cohorts for your warmth. I wish best luck with the rest of your journey. Thank you Joseph. You are a perfect mentor for me.

To all the UCF master's and doctoral students who have helped me with collecting my data, thank you. Thanks to my supervisee students for being with me in supervisor practicum. Special thanks to God for all you did, again.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE: INTRODUCTION

Counseling is defined as "a profession that empowers diverse individuals, families, and groups to accomplish mental health, wellness, education, and career goals" (Kaplan, Tarvydas, & Gladding, 2014, p. 368). Counseling, as a profession, requires diverse professional competencies to counsel multicultural clients with a variety of issues in different settings. Developing and maintaining counselor competence has been a pivotal issue in counselor education and supervision since counseling was considered to be a profession (Swank & Lambie, 2012; Tate, Bloom, Tassara, & Caperton, 2014). The need to demonstrate that counselors are capable of providing professional, competent service has also increased as counseling services have become more related to third party funding agencies (McLeod, 1996). Thus, identifying the construct of counselor competence and assessing counselor competence in a systematic and comprehensive way has been an urgent call within the counseling field for being accountable to the public.

Counselor competence is widely defined as an ability to offer effective counseling services to various types of clients in an ethical and professional manner (Fairburn & Cooper, 2011; Hill & Thompson, 2005; Swank, 2010). Building counseling competence requires (a) the possession of unique knowledge and skills, (b) understanding of ethical standards and its application, and (c) the integration of professional knowledge, skills, and affiliations into a professional identity (Parsons & Zhang, 2014). Likewise, the construct of counselor competence includes a variety of aspects to assess; thus, measuring the precise competence of an individual counselor is a complicated process. Additionally, because the evaluation of counselor competence can vary depending on the evaluator (e.g., supervisors, peers, clients), it is difficult

work to obtain consensus within diverse perspectives of evaluators regarding how to assess counselor competence (Wheeler, 2003).

The literature (e.g., Swank & Lambie, 2012; Tate et al., 2014; Wheeler, 2003) in assessing counselor competence stated that comprehensive measurement must require a specific definition of the construct of competence and the inclusion of diverse voices relevant to counselor competence in the evaluation. Examples of "diverse voices" may include (a) a counselor's self-assessment, (b) the clinical supervisor's view on the counselor, (c) peer evaluation of the counselor, and (d) the client's ratings of the counselor. Firstly, regarding counselor self-evaluation, counselor's self-rated inventories have been the most widely used for assessing counselor competence. According to a review of counselor competence scales (Tate et al., 2014), over 60% of 41 reviewed instruments were using self-reported formats. Among the instruments using self-report format, measuring the self-efficacy of counselors is the most common (e.g., Johnson, Baker, Kopala, Kiselica, & Thompson, 1989; Larson et al., 1992; Lent, Hill, & Hoffman, 2003). With self-rated evaluation, we can measure wider range of counselor competence and evaluate various abilities or characteristics that counselors possess because counselors have a great deal of information about themselves.

Secondly, the evaluations by the clinical supervisors or peers are mostly utilized when assessing a counselor's performance. As video or audio technology developed, the use of audio or video-recordings is more common to assess how well counselors are able to utilize their competence in sessions. Reviewing video-recorded sessions are time consuming; however, it provides invaluable benefits by gaining a vivid picture of the capabilities of the counselor (Wheeler, 2003). For instance, the Counseling Skills Scale (CSS; Eriksen & McAuliffe, 2003)

and the Counseling Competencies Scale (CCS; Swank et al., 2012) are rated by experts like supervisors or peer counselors based on the observation of live or recorded performance.

Lastly, diverse methods have been used to include clients' voice in evaluating counselor competence (Thompson & Hill, 1993). One way to obtain clients' opinion is through client satisfaction surveys on their treatment, which have widely been used. Another way is to use counseling outcomes. Under the trend addressing evidence-based practice, many investigations of change in clients' symptoms have been conducted through client outcome studies. For instance, both client feedback measurements - the Partners for Change Outcome Management System (Miller, Duncan, Sorrell, & Brown, 2005) and the Outcome Questionnaire (Lambert et al., 1996) were used to measure the client progress across sessions. With these client outcome measurements, the extent of counselor competence can be weighed by assessing the change in client outcomes; however, client outcomes are easily influenced by different variables (e.g., clinic environment, client contribution) other than the competence of counselors. Thus, using client outcomes is not a main resource of true counselor competence, but just a supplemental method of assessing counselor competence.

Consequently, although client feedback has historically been a key factor for evaluating counselor performance or effectiveness (Reese, Usher et al., 2009), the aforementioned methods of using clients' feedback were *indirect* ways to measure counselor competence, rather than *directly* assessing it. Therefore, we need a psychometrically sound client-rated measure to assess counselor competence in an appropriately direct way. The present study aimed to develop a new client-rated scale of counselor competence through systematic scale development procedures. This study employed the Rasch model measurement theory in order to develop a new reliable

and valid client-rated instrument with a sample of clients who received counseling from counselors-in-training. In specific, this study generated an item pool to measure clients' perception of the counselor competence as well as investigated the psychometric qualities of the developed client-version counselor competence scale by applying Rasch model.

Statement of the Problem

Obtaining professional competence in counseling and demonstrating it to the public has been one of the most important tasks in counseling in order to help advance the counseling profession. As such, assessing the counselor competence has been a primary focus in counselor education and supervision (Bernard & Goodyear, 2014; McAuliffe & Eriksen, 2011; Swank & Lambie, 2012). There have been various attempts to measure the true construct of counselor competence (e.g., Hughes, 2014; Swank, Lambie, & Witta, 2012; Urbani, Smith et al., 2002). Those attempts also tried to involve diverse perspectives around counselor competence in more comprehensive ways. However, a recent review of counselor competence (Tate et al., 2014) highlighted a lack of clients' view in assessing counselor competence or performance, with most instruments being completed by counseling experts like supervisors, peer counselors, and instructors. In fact, Tate and his colleagues (2014) reported that there were only two inventories that partially included the clients' voice among the 41 scales reviewed. One of the two instruments (i.e., Conceptualization of Group Dynamics Inventory; Tate et al., 2013) measured a client's perception of group dynamics, and the other instrument (i.e., Cross-Cultural Counseling Inventory–Revised; LaFromboise et al., 1991) was related to multicultural competence. Both inventories include clients' evaluation as only a part of the instruments, as well as they do not

assess the core constructs of counselor competence such as counseling skills. Considering that the primary concern of counseling is client welfare, this result must be very surprising, and a clear gap in the counselor education and clinical supervision research. In sum, there has been a deficit of direct client-rated measures to assess a key construct of counselor competence. Excluding clients' opinion could result in constructing an inaccurate concept of counselor competence and its assessment, driven only by counseling experts.

Additionally, the literature shows that most previous instruments measuring counselor competence were developed in the theoretical context of the classical test theory (CTT). The CTT's key theoretical framework is based on that an observed score of an examinee is equal to the sum of the true score and the measurement error score (DeVellis, 2012). This basic concept has been a mainstream in measurement and assessment since CTT was introduced in the early 20th century (DeVellis, 2012). In the CTT's context, it was believed that more test items and more measurements enable us to measure the construct more precisely. Thus, item redundancy in the CTT context is necessary for precise measurement, since larger numbers of items are needed (DeVellis, 2012). This characteristic of the CTT let test developers in counseling rationally to develop instruments consisting of numerous items, like other CTT-based tests. Many items not only require more time for examinees to complete the test, but also easily trigger the test-tiredness of examinees, which could hinder precise assessment.

In addition to the item redundancy of CTT, Smith, Conrad, Chang, and Piazza (2002) indicated that CTT-based tests have the limitation called *circular dependency,* which means the sample dependency of item functions and the item dependency of person score (Fan, 1998). This limitation requires a large sample size when developing a test, and hinders the generalizability of

test results to other samples and directly compare the scores between different samples.

Lastly, many measurement experts (e.g., Fox & Jones, 1998; Liu, 2010) addressed that CTT has some possible statistical problems since the CTT pretends the ordinal scales to be interval in analyzing the data, especially inferential analysis. For example, the Likert scale, rating from strongly disagree to strongly agree, is a scale format widely used in psychological tests or instruments. Nevertheless, the Likert scale is obviously not a ratio or interval scale, but an ordinal scale. Liu (2010) highlighted that simply considering the ordinal raw scores as interval scores would result in reducing the statistical power to reject null hypotheses in inferential analysis since higher error variance could occur in raw scores. Therefore, the CTT may not be the best way, especially when developing a new instrument.

Significance of the Study

The literature (e.g., Bernard & Goodyear, 2014; McAuliffe & Eriksen, 2011; Swank et al., 2012; Tate et al., 2014) showed that counselor competence has been a principal focus in the counseling profession. Accordingly, many researchers have addressed the importance of defining and assessing counselor competence because the elements of the competence can enhance competence-oriented training and guide the performance evaluation of counselors or counseling trainees (Hughes, 2014; Wheeler, 2003). Nevertheless, there has been a lack of psychometrically sound scales to measure counselor competence with diverse perspectives (Swank et al., 2012). In particular, measurements directly by clients have been rare in the assessment of counselor competence (Tate et al., 2014). As previously addressed, the lack of clients' perspective results in assessing a counselor's competence only from counseling experts'

perspective, which is more likely to arise the evaluation irrelevant to the perception of clients, that is, the subjects served by counselors.

The newly developed client-rated scale in this study can reflect clients' perspectives in evaluating counseling performance. In other words, the developed scale can provide counseling trainees as well as professional counselors with opportunities to assess their counseling skills from clients' perspectives. As such, this new measure, as a useful tool for client feedback, could aid to improve the ability of a counselor. For instance, a new client-rated assessment measuring how counselors interact competently with clients may serve as a tool of providing clients' formative and summative feedback. In many studies (e.g., Miller et al., 2005; Lambert et al., 1996), it was demonstrated that clients' feedback has a significant influence on improving counselors' counseling ability. Likewise, using psychometrically sound ratings from clients could result in increased effectiveness of counselor education courses, such as practicum and internship where counselor-trainees meet actual clients. These advantages, ultimately, could result in the improvement of the quality of counselor preparation programs because a new scale measuring the clients' feedback can be used as a method of regularly tracking the improvement of counselors-in-training in counselor education programs.

In addition, this study could introduce the application of the Rasch model (Rasch, 1960) in developing a more valid and reliable instrument to measure competence in counselor education. The use of the Rasch model helps to overcome the limitations that most instruments developed in the CTT framework have, because theoretically the Rasch model overcomes the major weakness of CTT, which has circular dependency of item statistics (Bond & Fox, 2001; Engelhard, 2013). Specifically, Rasch analysis enables test developers to make a sample-free or

item-free measurement by estimating two latent variables (i.e., person ability and item difficulty), independent on the test sample or the items. In addition, the Rasch model provides reliability and validity evidence, such as item hierarchy, fit statistics, and differential item function that CTT does not provide. Further, Rasch analysis can handle the statistical problem of using ordinal data by transforming ordinal scores into interval ones. The specific procedure of Rasch analysis is presented in the next chapter.

Lastly, developing clients' ratings of counselor competence could promote continued research in assessing counselor competence. Previous research has focused mostly on assessing client satisfaction or client symptoms. This scale, which directly measures clients' perception on counselor competence, could help the researcher to investigate the true construct of counselor competence from clients' point of view. The new instrument in this study could enable future research to compare different perceptions of the competence across different evaluators such as clients, supervisors, peer counselors, or counselors themselves.

In sum, we need to pay more work and attention on measuring counselor competence. In particular, it is urgent in counselor education that this research develop a psychometrically sound, comprehensive, and practical scale to include the clients' view in evaluating counselor competence. Developing a new client-rated scale of counselor competence is needed to provide invaluable information for enhancing a counselor's own professional capability as well as the quality of counselor preparation programs. Therefore, this study employed the Rasch measurement model for developing a new client-rated instrument to measure counselor competence.

Purpose and Research Questions

The purpose of this study is to develop a new client-rated scale of counselor competence by adopting the Rasch model approach and to assess the psychometric properties of the newly developed scale, named *Client Ratings of Counselor Competence* (CRCC) under the Rasch context. The specific research question in the present study is the following:

Research Question: What are the psychometric properties of the CRCC using the Rasch model?

Q1. What is the content validity of the CRCC from the Rasch analysis?

Q2. What is the structural evidence of the CRCC using the Rasch?

Q3. What is the substantive validity of the CRCC within the Rasch model?

Q4. What is the generalizability of the CRCC in the Rasch analysis?

Q5. What is the interpretability of the CRCC from the Rasch model?

CHAPTER TWO: LITERATURE REVIEW

This chapter reviews the literature around the concept of counselor competence, its measurement, and the Rasch measurement model. First, this review presents how counselor competence relates to the counseling profession and counselor education, addressing its importance in counselor education. Second, the current literature review includes the definition of counselor competence and the core competencies required for the entry-level counselor. Third, this review describes the various ways to measure the counselor competence. Lastly, this chapter presents the characteristics, the types, and the analyzing procedures of Rasch modeling.

Why Counselor Competence is Important

Before defining counselor competence, it is necessary to have a notion of why counselor competence is important. As noted earlier, developing and maintaining counselor competence has been one of primary concerns in counseling. That is because being a competent counselor is highly connected with obtaining the accountability of the public for the counselor as a professional. Reflecting its importance, the American Counseling Association's (ACA) Code of Ethics (ACA, 2014) and the Council for Accreditation of Counseling and Related Educational Programs (CACREP) Standards (CACREP, 2016) for counselor education includes several statements regarding counselor competence. Thus, the following section describes how counselor competence is associated with client welfare, counselor education, and counselor program evaluation, with specific statements in the ACA's 2014 Code of Ethics as well as the CACREP's 2016 Standards.

Welfare of Clients

The importance of counselor competence relates to the welfare of clients, the principle goal of counseling. All counselors have the ethical responsibility to provide their clients with the best possible care or treatment (Fairburn & Cooper, 2011). The ACA Code of Ethics stated, "The primary responsibility of counselors is to respect the dignity and promote the welfare of clients" (ACA, 2014, Section A.1.a). To provide effective counseling services toward the clients' wellbeing, counselors need to demonstrate their professional competence to the public and maintain their competent ability through continuing education and consistent evaluation. Regarding this responsibility, the ACA Code of Ethics (ACA, 2014) contains the code that counselors have to "continually monitor their effectiveness as professionals" (Section C.2.d) and to "maintain their competence in the skills they use" (Section C.2.f). The framework of counselor competence can work as an indicator of effective, competent counselors; thus, it is necessary to define what counselor competence is and apply the standards to all professionals in counseling or counseling-related fields.

Guidelines for Counselor Training

Another reason for the importance of counselor competence relates to the requirements of counselor training. The theoretical framework of counselor competence can provide counselor preparation programs with structured guidelines about curriculum, practice, and other requirements for counselors-in-training. CACREP has designed and regularly revised standards for counselor training programs, which include a set of fundamental competencies required for counselors. The CACREP Standards have added or integrated new competencies (e.g., group

counseling, multicultural counseling competencies) reflecting the needs of society; the latest CACREP Standards (CACREP, 2016) posed eight core areas of counselor competence with detailed explanations as student learning outcomes. Thus, all counseling programs accredited by CACREP construct their training curriculum following the framework of counselor competence provided by the CACREP 2016 Standards. Likewise, the Counselor Preparation Comprehensive Examination (CPCE), a standardized exam for counselors, uses the CACREP's construct of counselor competence and evaluates counselor trainees' knowledge about the eight core competence areas. As such, defining and assessing counselor competence is remarkably important and relevant in counselor education; however, counselor competence, especially its assessment, has received little attention.

Program Evaluation

The third reason arises from concerns about the evaluation of counselor preparation program. As noted early, the definition of counselor competence shows the kind of competence counselor trainees should possess to achieve the desired outcomes. Thus, assessing the counselor competence of graduate counseling students can be an important resource for evaluating counseling programs. Urofsky and Bobby (2012) reported that assessment of student learning shifted from input-based to outcome-based approach in the CACREP 2001 Standards, and the 2009 Standards have finally consolidated competence-based student learning outcomes. More specifically, CACREP Standards (2016) state that all counselor preparation programs should "have a documented, empirically based plan for systematically evaluating the program objectives, including student learning." (p. 17). The elements of counselor competence are key

features to assess student-learning outcomes in counselor education. Therefore, what and how to measure regarding counselor competence is a primary consideration in the evaluation of counselor training programs.

Counselor Competence

The concept of counselor competence is not easy to be defined. After reviewing the definition of counselor competence, the key components of counselor competence are explained in the following part.

Key Definitions

In order to make a precise notion of counselor competence, it is necessary to distinguish the difference within meanings of competence, counselor competence, and competencies. McLeod (1992) refers to *competenc*e as "any qualities or abilities of the person which contribute to effective performance of a role or task" (p.360). This generic definition of competence is applicable across all professions and relates more to a set of competencies and micro-skills relevant required outcomes (Ridley, Mollen, & Kelly, 2011). Similar to the definition of competence, *counselor competence* can be viewed as the combination between two nouns-counselor and competence. That is, the meaning of counselor competence can be defined as abilities of an individual counselor, which contribute to effective therapeutic outcomes. This definition is consistent with the notion of the literature viewing counselor competence as an ability of an individual counselor to provide an effective and professionally ethical counseling service to diverse population (Fairburn & Cooper, 2011; Swank, 2014; Swank & Lambie, 2012). Lastly, competencies are viewed as identifiable elements of competence (Leigh, Smith, et al.,

2007; Ridley et al., 2011). This implies that competencies are sub-concepts under competence. In summary, the literature showed that many researchers addressed the relevant outcomes or effectiveness in the professions when defining competence. As such, the current study defines counselor competence as a set of competencies of counselors for providing positive therapeutic outcomes to diverse clients.

Core Competencies

Although key constructs of counselor competence are identified differently by theorists (e.g., Swank, Lambie, & Witta, 2012; Tate et al., 2014; Urbani, Smith et al., 2002), its constructs are generally considered to consist of (a) professional knowledge, (b) skills, (c) propositions, (d) multicultural counseling competence, and (e) ethical and legal competence. Therefore, the researcher reviews how the literature determines the core elements of counselor competence in the following section.

Knowledge

The 2016 CACREP Standards posited eight core knowledge areas that all counseling trainees should acquire during their master-level education. The eight fundamental knowledge parts involves (a) professional counseling orientation, (b) social and cultural diversity, (c) human growth and development, (d) career development, (e) helping relationship, (f) group work, (g) assessment, and (h) research and program evaluation (CACREP, 2016). These areas include a comprehensive knowledge that counseling trainees should learn in entry-level education; thus, the curriculum in CACREP-accredited programs also cover the fundamental knowledge. In addition, the Counselor Preparation Comprehensive Examination (CPCE), a standardized test for

the qualification of counselors, is designed to assess the level of knowledge of counselors-in-training in terms of those eight knowledge areas. Although the knowledge is a key component of competence, it is hard to say the construct that client can assess. In addition, standardized tests like the CPCE are more appropriate method to assess the knowledge areas of counselors. Thus, the developed measure in this study does not include items to evaluate the knowledge part.

Skills

Counseling skills has been a primary focus in counselor training as well as one of core learning goals required for counseling students. Since Ivey (1971) suggested a *microskills* hierarchy model for intentional interviewing, skills-oriented training has been more focused in counselor education. The microskills hierarchy (Ivey 1971; Ivey, Ivey & Zalaquett, 2013) has three parts, whose bottom part is called attending behaviors. The attending skills include appropriate nonverbal language such as eye contact, body gesture, and vocal tone. The middle part, called basic listening skills, involves reflection of content and feeling. The last skills in the hierarchy are advanced skills consisting of confrontation, focusing, reflection of meaning, and influencing skills. Similarly, another prominent contributor, Egan (2013) also addressed the acquisition of basic counseling skills. He identified eight groups of skills for effective counselors: (a) establishing working alliance, (b) basic and advanced communication skills, (c) challenging skills, (d) clarifying problems, (e) goal setting, (f) developing a treatment plan, (g) implementation, and (h) continual evaluation.

In addition, Gazda (1997) posited three skill clusters that counselors need at three phases of helping. In the facilitation stage, counselors need empathy, respect, and warmth in order to

promote clients' self-understanding and self-exploration. The next phase, transition phase includes concreteness, genuineness, and self-disclosure skills, which leads to clients' commitment to change. Lastly, skilled counselors in the action phase need to use confrontation and immediacy to encourage clients' action for intended change. As noted, Gazda's (1997) model emphasized counselor's attitude or proposition rather than basic counseling skills, compared to Egan or Ivey's skill framework.

Lastly, Young (2013) conceptualized six categories of the basic helping skills, called *therapeutic building blocks*. Like Gazda's framework, Young (2013) categorized the basic skills according to skills needed in each stage of the helping process (i.e., developing relationship, assessing, generating goal, intervening and taking action, evaluating and reviewing). More specifically, the six groups of building block skills consist of (a) invitational skills, (b) reflecting skills, (c) advanced reflecting skills, (d) challenging skills, (f) goal-setting skills, (g) change technique (Young, 2013). The invitational skills include nonverbal skills such as eye contact, body position, and appropriate physical distance as well as opening skills like questioning and communication encouragers. Reflecting skills involve paraphrasing (reflection of contents) and reflection of feelings. Advanced reflecting skills have summarizing and reflecting meaning. In addition, challenging skills include giving feedback and confrontation. The goal-setting category involves focusing skills and identifying the problem. The last category, called change techniques, contains giving advice and information, reframing, and brainstorming.

Propositions

The question, "What is the characteristic of an effective counselor?" has been a long

issue in counseling field. Carl Rogers (1967), a founder of person-centered therapy, provided three core therapeutic propositions that all counselors should develop, which is widely admitted within professionals. The first proposition is *congruence*, which the attitude to be genuine with other individuals. The next core characteristics are *positive regards* to others. It means to respect diverse values of individuals and understand other without any prejudice. Moreover, the third condition required for counselors is *empathy*. The empathy attitude is the ability to deeply understand others' feelings, values, and view of world (Young, 2013).

Multicultural Counseling Competence

Multicultural counseling is on the agenda of most counselor training programs in the USA as the result of the new counseling paradigm addressing multicultural perspectives. Likewise, assessing multicultural counseling competence is a hot issue in counselor education; in this atmosphere, many measures of multicultural counseling competence have been developed (Ponterotto, Rieger, Barrett, & Sparks, 1994) based on the definition of the competencies by Sue, Arredondo, and McDavis (1992). A recent review of counselor competence instrument (Tate et al., 2014) also presented the trend indicating that inventories assessing multicultural competence (n =13) were almost one-thirds of all instruments (n = 41) included in the study. Those scales include items on awareness of personal cultural attitudes, bias, and prejudice as well as knowledge of culturally diverse values (Wheeler, 2003). However, when looking close at items included the instruments about multicultural competence, many items seem to be overlapped with counselors' characteristic factors, particularly therapeutic relational ones. Thus, the items about multicultural competence are excluded in this study.

Ethical and Legal Competence

The last competence element concerns dealing with ethical and legal issues. The competence relevant to ethical and legal issue is one of the most difficult to define and measure in diverse aspects of counselor competence. However, the knowledge and ability to make appropriate decisions ethically and legally have been consistently emphasized in ethic codes and standards of ACA (ACA, 2014; CACREP, 2016). This emphasis on ethics and laws seems deserved in the nature of counseling as one of professions requiring professional judgments; however, the ethical and legal competence has received comparatively less attention (Mullen, Lambie, & Conley, 2014). There was only one instrument found in the literature: Ethical Legal Issues in Counseling Self-Efficacy Scale (ELICSES; Mullen et al., 2014). The ELICSES is a self-reported inventory to assess the ethical and legal knowledge of counselors and their self-esteem in dealing ethically and legal sensitive issues. This scale is consisted 23 items using the range from zero (cannot do at all) to 100 (highly certain can do). The ELICSES has three sub-scales labeled as (a) general ethical and legal issues in counseling self-efficacy, (b) suicide, violence, abuse and neglect self-efficacy, and (c) counselor development and wellness self-efficacy. As just noted, the ethical and legal counselor competence is an important construct; however, these ethical and legal competence-related variables are also not included in this study since the ethical and legal competence is thought to be hardly observed by clients in sessions.

Assessment of Counselor Competence

Assessing counselor competence is a complex and challenging process because there are many elements constructing the competence as latent variable as well as diverse perspectives

among the assessment (McLeod, 1992; Swank & Lambie, 2012). There is no single way sufficient for counselor competence assessment. In this section, what and how to measure counselor competence within the previous studies are reviewed. In other words, various methods of assessing competence are presented. Further, these methods include self-reports, standardized tests, performance assessments, and the use of client feedback.

Self-Assessment for Diverse Aspects of Counselor Competence

The simplest and easiest way to assessing counselors' competence is to evaluate their own competence and performance. With this reason, self-rated inventories are the most widely used for assessment various counselor competence in counseling. According to a systematic review about counselor competence inventories (Tate et al., 2014), almost two-thirds of reviewed 41 instruments use self-report format. Specifically, the self-rated instruments (n = 25) in the meta-review included general counselor competence like counseling skills (n = 6), multicultural competence (n = 11), group counseling competence (n = 2), school counseling competence (n = 2), career counseling competence (n = 1), addiction counseling competence (n = 1), and others (n = 2). Tate and colleagues (2014) found that most self-report instruments related to the concept of self-efficacy (Bandura, 1977). Since Albert Bandura (1977) coined self-efficacy as self-belief or self-perception to make a successful performance in a certain task, self-efficacy has been widely employed to measure individual's estimate ability in diverse tasks of different fields. Likewise, the assessments using self-efficacy is commonly utilized to measure self-esteem or confidence of counseling trainees in counselor education and supervision (e.g., Johnson, Baker, Kopala, Kiselica, & Thompson, 1989; Larson et al., 1992; Lent, Hill, & Hoffman, 2003). For

instance, the Counseling Self Estimate Inventory (COSE; Larson et al., 1992) measures the self-esteem of a counselor as to using counseling skills, attending procedure, dealing with difficulty client responses, multicultural competence, and self-awareness. The Counselor Activity Self-Efficacy Scales (CASES; Lent et al., 2003) also was designed to assess counselors' own competence, which reflects an individual counselor's self-confidence about using counseling skills, managing session, and dealing with challenging issues with clients.

Additionally, individual-oriented competence such as self-awareness, self-care, and multicultural perspective is more related to self-perception, and tricky to measure by others; thus, the use of self-rated format to measure those variables seems quite rational and logical. As such, many instruments to assess personal variables like awareness, self-wellness, and cultural sensitivity use self-rated assessing formats. Regarding this, Tate et al. (2014) in their meta-analysis reported that the instruments assessing multicultural counseling competence contained self-reported instruments (n = 11), indicating almost 85% of total instruments (n = 13). In addition, despite small number of instruments contained in the study, all inventories to measure specific competence required in school (n = 2), career (n = 1), addiction counseling (n =1) were self-reported (Tate et al., 2014). In short, although self-reported methods are the best way to contain counselors' own perspective in the assessment of counselor competence, the evaluation through only self-assessment cannot demonstrate the actual competence of a counselor. In other words, only self-assessment is not sufficient to measure a comprehensives level of development in counselor competence.

Standardized Test for Knowledge

When counseling students are close to their graduation, most of programs evaluate their comprehensive knowledge fundamental for providing professional practices. Some programs may employ standardized exams to evaluate broad knowledge of students as parts of counselor competence such as the Counselor Preparation Comprehensive Examination (CPCE) and the National Counselor Examination (NCE). Specifically, the CPCE includes the assessment of eight core CACREP areas: (a) professional identity, (b) social and cultural diversity, (c) human growth and development, (d) career development, (e) helping relationship, (f) group work, (g) assessment, and (h) research and program evaluation. The NCE, consisting of 200 multiple-choice items, contains not only the eight content areas like the CPCE, but also practical knowledge including (a) fundamental counseling issues, (b) counseling process, (c) diagnostic and assessment services, (d) professional practice, and (e) professional development, supervision and consultation. However, both CPCE and NCE assessments work just as a minimal criterion of requirements to acquire licensure and certification. Additionally, both tests mostly focus more on knowledge-related areas of competencies of counselors-in-training, rather than other key portions like counseling skills and professional attitude. In order to gain more reliable and valid measurement in counseling students' competence, supplemental methods of assessing the diverse aspects of competence are needed.

Performance Assessment for Counseling Skills

In addition to the acquisition of knowledge, the implementation of their knowledge and counseling skills in practice is another key assessment of counselor competence, probably the

most important area considering the nature of counseling with clients. Since the emerging age of counseling profession, evaluating counselor performance has been a primary concern in counselor training (Tate et al., 2014). For assessing the performance of counselors, written resources such as verbatim and case studies were widely used in the beginning generation of profession. As video or audio technology develops, the use of audio or video-recorded tapes is more common to assess how well counselor utilizes their competence in sessions. When performance raters, mostly experts like supervisors or instructors, assess other counselors' competence of using counseling skills, they fill out a structured rubric form while they review the whole or parts of a recorded session. The Interpersonal Process Recall (IPR) developed by Kagan (1963) is a useful tool of using tape recording to provide immediate feedback based on live assessments. Reviewing taped sessions are time consuming; however, it provides invaluable benefits gaining a vivid picture of the capabilities of the counselor (Wheeler, 2003). With this advantage of using recorded performance material, several instruments (e.g., Counseling Competencies Scale [CCS; Swank et al., 2012], Counseling Skills Scale [CSS; Eriksen & McAuliffe, 2003]) use expert-rated format based on the observation of live or recorded performance. More specifically, the manual for CCS (Swank et al., 2012) includes that all raters should review at least 15 minutes of the recorded clip and evaluate the level of counseling skills of counselors of counselors-in-training. Similarly, the CSS (Eriksen & McAuliffe, 2003) was designed be evaluated based on observations of actual in-session performance by experts.

This method of using a counselor's actual behaviors is an accountable way of assessing the counselor's competence, in terms of that it can directly assess the counselor's in-session performance at implementing counseling skills. However, it has several problems. First, when

using this way, it has often shown low inter-rater reliability due to difficulty to reach a consensus in defining the behaviors of performers (e.g., Swank & Lambie, 2012). In spite of many efforts developing a structured protocol, providing a detailed manual, or acquiring training for administration, the difference between raters still exists in the nature of social constructivism addressing the existence of individual perspective. The second problem relates to the selection of the session to be rated (Fairburn & Cooper, 2011). Specifically, when the evaluator chooses a session or its parts for assessing, an error of sampling is likely to occur, probably resulting to a biased assessment. The third problem concerns the relationship between a rater and the person to be rated (Fairburn & Cooper, 2011). In practice, counselor performance is commonly evaluated by the individuals like course instructors, supervisors, or peers who have already known the counselor. As a result, the relationship between them that already formed might affect the final result of assessing the counselor's competence. Lastly, it seems problematic to assess a counselor's performance not by the client receiving his or her treatment, but mostly by a third party like counseling experts who observe the session. Regarding this, many studies (e.g., Tate et al., 2014) addressed that lack of clients' voice in assessing counselor performance is likely to bring about the emphasis of the competence less relevant to clients' outcomes.

Using the Feedback of Clients in Assessment

One way of containing the voice of clients in assessing counselor competence is to use client outcomes. This is an indirect method to measure counselor competence. Lambert and Shimokawa (2011) in their meta-analysis study introduced two client outcomes systems: (a) the Partners for Change Outcome Management System (PCOMS; Miller et al., 2005) and (b) the

Outcome Questionnaire system (OQ; Lambert et al., 1996). The PCOMS uses two brief scales with only four items for each, which include the Outcome Rating Scale (ORS; Miller, Duncan, Brown, Sparks, & Claud, 2003) assessing the mental health functioning of clients and the Session Rating Scale (SRS; Duncan & Miller, 2008) measuring the therapeutic relationship with the counselor. Additionally, the OQ system employs the Outcome Questionnaire-45 (OQ-45; Lambert et al., 1996) to assess client progress during the treatment. The OQ-45 with a 45 self-reported items was designed to assess three aspects of client functioning: psychological symptoms, interpersonal relationship, and social role functioning (Lambert et al., 1996). Both measurements provide visual graphs presenting the change of client outcomes measured by their own scales. In sum, the extent to how competent a counselor is indirectly weighed by assessing the change in client outcomes. This is an attractive way to reflect clients' feedback in assessing counselor competence; however, client outcome measures should be used as a supplemental assessment because client outcomes are easily influenced by different variables (e.g., clinic environment, client contribution) other than the capacity of counselors.

Another way of using client perspectives is to have clients directly rate their counselor competence. Inevitably, this is the most trustworthy method to reflect clients' perspective on the ability of the helper counseling them. Many previous studies emphasized that there was a significant difference in perceptions between clients and counselors (e.g., Dill-Standiford, Stiles, & Rorer, 1988; Thompson & Hill, 1993). Despite this fact, lack attention has been paid to clients' rating of counselor competence. According to Tate et al. (2014), there were only two inventories containing clients' voice among 41 counselor competence instruments: Conceptualization of Group Dynamics Inventory (CGDI; Tate et al., 2013) and Cross-Cultural

Counseling Inventory-Revised (CCCI-R; LaFromboise et al., 1991). In the independent search for this study, two more instruments to measure clients' perceptions were additionally found. First, the Helping Skills Measure (HSM; Hill & Kellems, 2002) was developed to assess client perception of the counseling skills used by counseling trainees in sessions. The HSM has 13 items using a 5-Likert scale ranging from strongly disagree to strongly agree. The higher score in the HSM indicates that that the counselor is more competent in using counseling skills. This scale also consisted of three sub-scales: exploration, insight, and action, with each of subscale measures the capability of using the skills required in a developmental process of helping. The second inventory is the Multicultural Therapy Competency Inventory-Client Version (MTCI-CV; Cole, Piercy, Wolfe, & West, 2014). This client-rated scale with 32 items assesses counselors' multicultural competence from clients' perceptions. In order to assess the level of competence, the MTCI-CV use a three-point Likert scale indicating "Does this very well", "Does this adequately", and "Does this poorly". More specifically the MTCI-CV measures counselor's self-awareness of own cultural values and client's worldview, use of culturally acceptable interventions, multicultural attitude.

Measurement Theory

Measurement theories provide a theoretical foundation and specific procedures for developing a more valid and reliable measurement. This section briefly reviews the classical test theory, a dominant measurement theory in 20th century and its limitation, and then addresses the definition, characteristics, advantages, and functions of the Rasch model, a more modern alternative measurement theory, focusing on how the Rasch model can overcome the limitations

of the classical test theory.

Classical Test Theory

Classical test theory (CTT) has been the popular and dominant model for test development (DeVellis, 2012). This classical measurement theory was established on the fundamental concept of that the observed score of a subject is the sum of the subject's true score and the measurement error score (DeVellis, 2012). The CTT assumes a true score to be measured through infinite observations of what to be measured (Liu, 2010). This is represented by the following formula:

$$X = T + E,$$

where X represents a subject's observed score measured with a set of items, T represents the subject's estimated true score or level on the latent variable, and E represents a random measurement error component (Crocker & Algina, 2008; DeVellis, 2012).

The CTT model has three key assumptions (DeVellis, 2012). First, true scores on the latent variable are not correlated with each item's error scores. In the context of the CTT, the error term is viewed to be associated only with that particular item. The second assumption is that average error score in the population of examinees is zero. This means that the mean of the error scores associated with individual items reaches to zero when applying the items for a larger number of samples. Third, the CTT assumes that the error term of a single item is not correlated with other items' error scores.

Based on this theoretical assumption, there are several advantages of the CTT model. The major advantage of the CTT is relatively easy to understand and apply because the CTT does not

need complicated requirements. Thus, the results of the CTT's analysis are relatively easy to meet with real test data. Since the data analysis of CTT focuses mostly on group correlation scores, the CTT model is relatively simple at the item level. How a person responds to a single item is not examined. Under the CTT-driven analysis, the evaluation of items is successfully conducted by only demonstrating a modest relationship to the underlying variable being measured in the measure (DeVellis, 2012). Additional advantages of CTT are easy to use. As major statistical packages (i.e., SPSS, SAS, LISREL) basically provide analyzing functions for performing the analyses required for CTT (i.e., factor analyses, computing coefficient alpha, etc.), the CTT can be more available for researchers to use in developing a measure, without additional education and cost (Soska, 2012).

Limitations of Classical Test Theory

Several limitations of the CTT have been discussed in the literature. Regarding the limitations, Smith et al. (2002) summarized three major limitations of the CTT model: sample dependency of item indices and item dependency of person ability, inability of detecting how a person responds to any given item, and assuming ordinal scale to be interval.

When applying the CTT, the indices (e.g., point biserial correlations, reliability) to evaluate the quality of items are defendant on the tested sample. Likewise, the evaluation of a person's ability theoretically defends on the items used in the test. This characteristic of the CTT makes it difficult for test developers to develop sample-free or item-free tests. For instance, if a certain item is given to the sample group with higher level of ability, the proportion of individuals answering the item correctly would be higher. If a test consisting of more difficult

items is given to a certain person, the person would get a lower score in the test. Fan (1998) addressed that both dependencies are correlated circularly and he defined it as circular dependency. This circularly dependent relationship poses theoretical limitation of the CTT, not developing an invariant measurement (Engelhard, 2013).

Another major limitation of the CTT by Smith et al. (2002) is its inability to examine how an individual responds to a particular item. Specifically, the CTT cannot provide the information on how a person with a certain ability answers to an item or question with a certain item difficulty level. In evaluating a person's performance or an item's functioning, more precise investigation (e.g., detecting unexpected response pattern) required in person-level or item-level analysis is not possible for the CTT. This drawback of the CTT might cause the risk that validity evidence is established mostly by reliance on the correlation statistics between items (Sammet, 2012). The validity evidence dependent mostly on factor analysis and correlations between different tests is not enough to demonstrate whether a test or instrument is valid.

The third major limitation identified by Smith et al. (2002) is that the CTT assumes ordinal scales as interval scales. Many raw scores from instruments or surveys using the Likert-scale or similar are not interval, but ordinal (Bond & Fox, 2001). Due to pretending the ordinal scales to be interval in analyzing the data, the CTT could have some possible statistical problems. Liu (2010) pointed out the risk that considering the ordinal scores as interval scores could reduce the statistical power of rejecting null hypotheses in inferential analysis since higher error variance could occur in raw scores.

In addition to these limitations of the CTT, the item redundancy that the CTT typically creates a test with many items was a common limitation of the CTT discussed, due to its

theoretical basis viewing the redundancy as the root of reliability (DeVellis, 2012).

As the limitations of the CTT have been identified, several alternative measurement theories (e.g., item response theory, Rasch model) have been suggested to overcome these limitations of the classical measurement approach. Among them, Rasch measurement model can be used to address the limitations of CTT and provides a more robust method for constructing valid and reliable measures.

Rasch Model

Rasch model is a modern measurement theory to provide a strict guideline to identify, construct, and evaluate items to measure a distinct construct of interest. It was originally developed by the Danish mathematician Georg Rasch in 1960 and has been advanced and extended by several subsequent researchers such as Andrich (1978), Wright and Masters (1982), and Adams, Wilson, and Wang (1997). The Rasch model makes it possible to develop a scale with more reliability and validity by evaluating whether the data fits the requirements of the Rasch model rather than exploring a model to best fit the data (Bond & Fox, 2001). The model also provides an explanation of how a person responds to a specific item in his or her own way, which enables test developers to see each item's function, not focusing on group statistics (Liu, 2010). In the following part, the characteristics, functions, variations, and advantages of the Rasch model are described specifically.

Characteristics

There are several distinct features of the Rasch model. First, the Rasch model uses an interval scale as the unit of analysis, which enables researchers to conduct the item analysis with

less statistical errors, compared to using ordinal scales. For instance, assuming ordinal scores to be interval, not using true interval scales, could cause higher error in variance, resulting in reducing the statistical power in analysis (Liu, 2010). For solving this limited issue, the Rasch model uses a logarithmic transformation of ordinal scores to create truly interval scores (Bond & Fox, 2001). The specific steps of the transformation the scale will be explained later in this section.

Second, the Rasch model assumes that a scale measures a unidimensional construct of interest, which means that each item on a test or an instrument contributes to the measurement of a single attribute. The concept of unidimensionality is usually easy to understand when we use the measurement of size, height, and temperature, which are explicit attributes. We simply admit that these measurements of distinct values focus on only one attribute and can be assessed separately. However, when applying the unidimensional concept to social science, it is more complicated to determine whether a test can measure only one attribute that a researcher intends to assess. That is because most variables in social science are latent and their underlying constructs are difficult to be identified and measured in a clear way. For example, tests measuring an individual's intelligence like the Wechsler's scale were originally developed by combining several sub-tests to measure different abilities like reasoning and working memory; thus, these composite tests were not fundamentally unidimensional. In addition, many instruments or inventories using the classical test theory had two or more subscales, which were determined by exploratory factor analysis. The Career Thoughts Inventory (Sampson et al., 1996), for instance, had three subscales to measure career-related dysfunctional thoughts, called *decision-making confusion*, *commitment anxiety*, and *external conflict*. Bond and Fox (2001)

addressed the importance of measuring a single attribute at a time although most attributes in real life always are complex. Regarding this, Bond and Fox (2001) described "Although the complexity of what we are measuring appears to be lost, it is through measuring one attribute at a time that we can develop both useful and meaningful composite descriptions" (p.25). Consistent to this perspective, the Rasch model originally aims to develop a scale to measure only one attribute, which guides a whole process regarding a scale development including item creation, calibration, analysis, and evaluation.

The third feature of the Rasch model is local independence, which assumes that a scale should be invariant across the sample of respondents. In other words, a measurement should show the same performance no matter who takes the test. Liu (2010) explained this concept using a meter stick example. If a meter stick, being used to measure the height of student, was considered a good measurement, the meter stick was obviously invariant across persons and its measurement should be influenced by only the student's height, regardless of other characteristics of the student being measured. Like the meter stick measurement, Rasch model assumes that the function of each item should be maintained to all persons in the same manner. Specifically, if an item is easier, it means in Rasch model that all persons are more likely to answer the item correctly. This characteristic is very important for a scale development because a person's true score should not change across a sample of subjects taking a test.

Advantages

The Rasch model has several advantages over the classical test theory thanks to the Rasch model's features explained above. First, using the Rasch model can provide information about

validity that the classical test theory cannot in terms of item analysis. For instance, the results in the Rasch's analysis report information about the level of item difficulty, person reliability, separation statistics, and unidimensionality. Second, the Rasch model helps determine which items are most useful in measuring the variable of interest. The Rasch model provides the fit statistics of each item, which shows the individual item's location and function, whereas the traditional test model gives only the information about the extent to which of items contribute to the variance among the measured variable. Third, when applying the Rasch model into scale development, the researcher can reduce item redundancy through deleting items with similar difficulty level. This advantage can reduce total number of items on an instrument, resulting in decreasing the total time needed for administering and scoring.

How to function

Like other models within item response theory, the Rasch model uses a logistic transformation of proportion scores for items and persons, known as log-odds units, simply logits. The logits represent the Rasch model's scale units similar to the centimeters in rulers. The logistic-transformed score for items is called *item difficulty* in the Rasch model, while the logistic-transformed score of persons is called *person ability*. These two latent variables are the key estimates parameters used in the Rasch analysis. The two parameters allow test developers to investigate the performance of each person and item more soundly. Engelhard (2013) highlighted several statistical advantages for using the Rasch model's logistic transformation in measurement development. First, the logistic distribution provides a reliable approximation like a normal distribution. Second, the logistic transformation yields an exponential distribution with

statistically desirable properties (Barndorff-Nielson, 1978; Engelhard, 2013). Lastly, the logit scale, as an interval scale, can be useful for developing a linear scale.

Table 1 shows how the logit scores change according to the proportion of correct responses for items and persons. Specifically, a higher logit score indicates the item is harder and the person is more able. For example, the logit score of 0.00 for an item means that the difficulty level of the item is 0.50, that is, half of test takers would answer correctly to the item. Likewise, the person ability logit of 0.00 indicates that the number of correct answers divided by total number of items for a person is 0.50.

Using two parameters such as person's ability and item's difficulty, the Rasch model calculates the probability of a person (n) with a certain ability to correctly answer an item (i) with a certain difficulty level from the empirical response pattern data. The equation of the probability can be expressed as

$$\mathrm{P}\,(\mathrm{X} = 1|B_n,\ D_i) = \frac{e^{(B_n - D_i)}}{1 + e^{(B_n - D_i)}}$$

where B_n means an individual's ability level, while D_i indicates the difficulty level of an item. Both B_n and D_i are latent variables that are estimated from empirical observations. The Rasch model assumes that both latent variables all have their own S-shaped curve, known as the item characteristic curve, in which the likelihood of answering an item correctly increases monotonically as the test taker's ability increases. On the other hand, the likelihood of a person with the same ability answering an item correctly decreases as the item difficulty increases.

Table 1

Change in Logits according to Person Ability and Item Difficulty

Logit	Item Difficulty	Person Ability
	Hard item	High ability
5.00	0.01	0.99
4.50	0.01	0.99
4.00	0.02	0.98
3.50	0.03	0.97
3.00	0.05	0.95
2.50	0.08	0.92
2.00	0.12	0.88
1.50	0.18	0.82
1.00	0.27	0.73
0.50	0.38	0.62
0.00	0.50	0.50
-0.50	0.62	0.38
-1.00	0.73	0.27
-1.50	0.82	0.18
-2.00	0.88	0.12
-2.50	0.92	0.08
-3.00	0.95	0.05
-3.50	0.97	0.03
-4.00	0.98	0.02
-4.50	0.99	0.01
-5.00	0.99	0.01
	Easy item	Low ability

Since the odds for an event is the ratio of the likelihood of happening over the likelihood of not happening, the odds for the respondent to answer a particular item correctly can be presented as

$$\frac{P}{1-P} = \frac{e^{(B_n - D_i)}}{1 + e^{(B_n - D_i)}} \Big/ \left(1 - \frac{e^{(B_n - D_i)}}{1 + e^{(B_n - D_i)}}\right) = e^{(B_n - D_i)}$$

If we calculate the natural logarithm of the odds obtained above, the equation is as follows

$$\mathrm{L} = \ln\left(\frac{P}{1-P}\right) = B_n - D_i$$

The natural logarithm of the odds, that is, log-odds, is called logits in the Rasch model. The logits are simply defined as a difference between an individual's ability and item difficulty. For instance, if a person has the same ability logit as the item difficulty, the logit score of the person on the item is zero, which means the person has the probability of 50% to answer the item correctly.

Fit statistics

The Rasch analysis, with two parameters (i.e., person ability and item difficulty), estimates the response of a person on an item, and then calculate fit statistics on items and persons by comparing the expected score and the observed response. Fit indices in the Rasch analysis use chi-square fit statistics in analyzing the residual between expected scores and observed scores. The Rasch analysis provides two types of the fit indices for both items and persons. One is called *infit statistics*, and another one is *outfit statistics*. Specifically, the four fit indices that the Rasch provides are infit mean square, outfit mean square, two standardized values for each infit and outfit mean square. The fit statistics indicate the magnitude of the randomness within person responses (Linacre, 2002b). The outfit statistic is unweighted fit value and is more sensitive to unexpected responses, that is, outliers, whereas the infit statistic is information-weighted fit value and is sensitive to response patterns. According to Linacre (2002b), outfit mean squares are influenced by outliers; thus, it is comparatively easy to fix the problem and less threat to measurement; however, infit mean squares' problems related to the response pattern are hard to diagnose and could be greater threat to measurement.

Table 2

Recommended Range of Fit Statistics

Type of Test	Range
Multiple-choice test (High-stakes level)	0.8 – 1.2
Multiple-choice test	0.7 – 1.3
Rating scale	0.6 – 1.4
Clinical observation	0.5 – 1.7
Judged	0.4 – 1.2

If the item perfectly fit the Rasch model, its value of mean square fit index is expected to be 1.0 and that of standardized fit to be 0.0. Bond and Fox (2001) indicated that the item or person with mean squares less than 1.0 and standardized mean squares less than -2.0 is considered to be "overfit"; the item or person with more than 1.3 mean squares fit and 2.0 standardized fit is viewed "underfit". Bond and Fox (2001) also suggested appropriate fit statistic ranges according to the different type of test (Table 2). Specifically, high-stakes level tests like SAT require the strictest range from 0.8 to 1.2; most psychological instruments or surveys using Likert scale require the lenient range from 0.6 to 1.4. Since the developing instrument in this study uses a rating scale, the range from 0.6 to 1.4 is employed in this study as the acceptable criteria for evaluating the fit statistic of each item and person.

Summary

As reviewed above in this chapter, counselor competence is a substantial issue continuously discussed in counseling. The construct is important given the definition of counselor competence guides how to construct the counselor education program and provides a standard for the program evaluation. Thus, there have been many efforts to define counselor competence and identify its components. We reviewed core competencies consisting of professional knowledge, counseling skills, propositions, multicultural counseling competence, and ethical/legal competence.

In addition, we reviewed that counselor competence can be measured in a comprehensive way, by including the perspectives of diverse evaluators around counselor ability and assessing different areas on the construct of interest. However, there is lack of client perception in measuring counselor competence; thus, more work on the inclusion of clients' voice is needs to assess the counselor competence precisely and comprehensively.

Lastly, the characteristics, advantages, and analysis of the Rasch model were explained. The literature showed that applying the Rasch model in developing a new instrument helps test developers to compromise the limitations that the CTT has, in order to create a more valid and reliable measurement.

CHAPTER THREE:
METHODOLOGY

Research methodology provides a road map of specific research steps that the researcher conducted. Specifically, this methodology section presents research design, definition of population and sampling methods, instrumentation, data collecting procedures, and data analysis.

Research Design

The research design in the current study employed the Rasch model measurement theory (Rasch, 1961) for developing the Clients' Rating of Counselor Competence (CRCC) and examine the psychometric properties of the developed scale with a sample of clients in counseling. The Rasch model provides a guideline for developing a linear measure and testing its quality. Specifically, the procedure of developing the CRCC followed the guideline for the scale development using Rasch model, proposed by several researchers (e.g., Engelhard, 2013; Liu, 2010; Wolfe & Smith, 2007).

Population and Sample

The population in the present study is adult clients who receive individual counseling services from counseling trainees among the CACREP-accredited programs in the U.S. The researcher employed a convenience sampling for developing the CRCC. Thus, the researcher recruited participants through a CACREP-accredited program in a large university in the South East of the United States.

In terms of a desired sample size for the Rasch analysis, the study followed the suggestion by Linacre (1994). With the theoretical basis that the modelled standard error

determines the stability of an item calibration, Linacre (1994) calculated the minimum sample size to obtain useful, stable estimates of item calibrations. Bigger sample is needed for polytomies than for dichotomies. For example, dichotomy data requires minimum cases of 30, while polytomous scored data needs the minimum sample of 50 (Linacre, 1994). Thus, the minimum sample size for this study is 50 subjects, which is required to obtain true item difficulty within 1.0 logit with two-tailed 99% confidence. The desired sample size of 150 subjects could be recommended for more precise estimates within 0.5 logits with 99% confidence (e.g., Lamb, Vallett, & Annetta, 2014; Lamb, Annetta, Meldrum, & Vallett, 2012). Therefore, above 150 adult clients were ideally recruited to participate in this study; however, 84 adult clients finally participated in the field test of the CRCC in this study. This sample size was met for the requirement of minimum number of 50 subjects (Linacre, 1994).

Demographic Information of Participants

With the IRB approval from the author's university (Number: SBE-15-11770), 84 participants were recruited through a community counseling center in a large university in the South East of the United States, consisting of 48 females (57.1%), 32 males (38.1%), and one transgender (1.2%); three participants (3.2%) did not provide the demographic information. The sample's mean age was 30.84 (SD = 10.69), ranging from 18 to 62. The majority of participants was 20 to 29-year old adult clients, which was about half of the sample. In terms of the race of the participants, Caucasian (n = 35, 41.7%), Hispanic (n = 21, 25.0%), Black/African (n = 15, 19.0%), Asian (n = 4, 4.8%), and other (n = 5, 6.0%) clients completed the CRCC after session (See Table 3 for more information).

Table 3

Demographic Data of the Study Sample

Category	*N* (84)	Percentage (%)
Gender		
Female	48	57.1 %
Male	32	38.1 %
Transgender	1	1.2 %
Missing	3	3.6 %
Age		
18-19	5	6.0 %
20-29	42	50.6 %
30-39	20	24.1 %
40-49	6	7.2 %
50-59	7	8.4 %
60-62	1	1.2 %
Missing	3	3.6 %
Ethnicity		
Asian	4	4.8 %
African American	16	19.0 %
Caucasian	35	41.7 %
Hispanic	21	25.0 %
Other	5	6.0 %
Missing	3	3.6 %

Instrumentation

Client Ratings of Counselor Competence (CRCC)

The CRCC was designed to assess clients' perceptions of counselor competence used by counselors in sessions. The initial item pool of CRCC was created based on the theoretical framework in counseling field, and the face validity was pilot tested and confirmed by a team of experts. The final scale of the CRCC used for the field test included 36 items using a 4-point

Likert-type scale from 1 (strongly disagree) to 4 (strongly agree). The specific procedure to develop the CRCC is presented in the following section. The Cronbach's alpha of the 36-item CRCC in this study was .927.

Instrument Development Procedures

The study employed a set of specific steps that previous researchers (e.g., Engelhard, 2013; Liu, 2010; Wolfe & Smith, 2007) proposed in developing a measurement scale. The guideline consists of seven processes: (a) definition of what to measure, (b) developing the test specification, (c) generating an appropriate pool of items, (d) determining the scale-type of instrument, (d) conducting an expert review with the initial item pool, (e) conducting a pilot test, and (f) expert reviewing with the revised CRCC.

Step 1: Definition of What to Measure

The first phase of the scale development is to determine what a scale developer wants to measure (Liu, 2010). For the purpose of constructing the CRCC, the construct of interest in this study is defined as the counselor competence, which means the ability to provide effective, professionally ethical counseling service to clients. The literature views counselor competence as the composite of the knowledge, skills, and characteristics required to provide such professional services (McLeod, 1992; Swank, 2010; Wheeler, 2003). Additionally, the literature on counselor competence noted that the competence includes counseling skills, self-awareness, ethical attitude, self-care, and theoretical knowledge. The CACREP (2016) also provides eight-core areas fundamental to a professional, competent counselor: (a) professional identity, (b) social and cultural diversity, (c) human growth and development, (d) career development, (e)

helping relationship, (f) group work, (g) assessment, and (h) research and program evaluation. In regards to the Rasch model pursuing a unidimensional measure, measuring the comprehensive construct of counselor competence is not appropriate for this study. In addition, most of counselors' aforementioned competencies (e.g., knowledge, self-awareness, self-care, group work, career development) are impossible for individual clients to observe them in sessions. As such, the present study focused more on the ability of counselors that is able to be observed as well as evaluated by clients, in terms of the research purpose of developing a new client-rated measure. Therefore, the counselor competence in this study was determined as counselor trainees' competence measurable by clients; thus the latent variable measured in the CRCC included two constructs – counseling skills and counselor's therapeutic attitude.

Step 2: Developing a Test Specification

This study applied the Rasch measurement theory to develop a new client-rated instrument of counselor competence. The Rasch model assumes that a scale measures a unidimensional construct of interest (Bond & Fox, 2001; Engelhard, 2013). Consistent to this perspective, the Rasch model primarily aims to develop a scale to measure only one attribute, which guides a whole process regarding a scale development including item creation, calibration, analysis, and evaluation (Engelhard, 2013). Thus, the Rasch model provides a theoretical framework to generate a hierarchical item pool consisting of a linear measure. Specifically, this study employed a conceptual variable map using the Rasch model, suggested by Engelhard (2013).

Table 4

Hypothesized Variable Map for the CRCC

What is the latent variable?

The latent variable is the counselor competence that a counselor presents in session.

Logit Scale	Cluster	Observations [items]	
			Hard item
5.00	Giving feedback, information	Relevance of solving problem, not directive	
4.00	Self-disclosure	Disclosure related to client's issue	
3.00	Confrontation	Balance of pushing and supporting	
2.00	Reflection of Meaning	Reflecting core value, viewpoint	
1.00	Reflection of Feeling	Normalizing feeling, Validating feeling	
0.00	Summarizing	Summarizing key contents	
-1.00	Reflection of Contents (Paraphrasing)	Reflecting key contents, Accuracy	
-2.00	Questioning	Open question, Clear question, One at once	
-3.00	Encourager	Verbal prompts, reassurance	
-4.00	Therapeutic Attitude	Empathy, Congruence, Positive regards	
-5.00	Nonverbal skills	Eye contact, Physical distance	Easy item

What is the response format or rating scale used?

Likert scale are used ranging from 1 (Strongly disagree) to 4 (Strongly agree)

By using the variable map, Table 4 illustrates a blueprint, that is test specification, to construct the latent variable of interest in this study. As defined in previous step, the latent variable that this study intends to measure was the counselor competence that a counselor presents in session. Based on the literature (e.g., Ivey et al., 2013; Young, 2013), the underlying

hierarchy of counselor competence was constructed to consist of 11 different clusters: nonverbal skills, therapeutic attitude, encourager, questioning, reflection of contents, summarizing, reflection of feeling, reflection of meaning, confrontation, self-disclosure, and giving feedback or information.

Step 3: Generating an Appropriate Pool of Items

The next step is to generate an item pool that relates to the measurement of counselor competence. For the item development of the CRCC, the literature (e.g., Egan, 2013; Ivey et al., 2013; Young, 2013) and other instruments (e.g., Barrett-Lennard, 1986; Hill & Kellems, 2002; Lent, Hill, & Hoffman, 2003; Swank, Lambie, & Witta, 2012) related to counseling skills and counselor attitude were reviewed. Based on the test specification, totally 85 items were created in the initial item pool of the CRCC.

Step 4: Determining the Scale Format of Measurement

The fourth step of a scale development involves choosing how to scale the measurement, that is, the type of measuring scale. DeVellis (2012) suggested that Likert scale is widely used in social science, especially in measurements of opinions, beliefs, and attitudes. Thus, the Likert scale ranging from strongly disagree to strongly agree was selected for the CRCC. However, several studies (e.g., Zaporozhets, Fox, Beltyukova, Laux, Piazza, & Salyers, 2015) using the Rasch model indicated a neutral middle point (e.g., uncertain, or not agree/disagree) does not function appropriately in the Rasch scale analysis; therefore, the study finally employed four-point Likert-scale (strongly disagree, disagree, agree, and strongly disagree) without a neutral middle point.

Step 5: Reviewing the Initial Item Pool with Experts

After developing the initial item pool with four-point Likert-scale, a group of experts were asked to review the whole item pool to check the content validity of the scale. This expert reviewing process involved six professionals who are knowledgeable in the counselor competence literature. All experts are holding Ph.D. degree and are full-time faculty in CACREP-accredited counselor education programs. Six experts recruited evaluated each item over three dimensions- (a) importance, (b) relevance, and (c) clarity, with five-point scale. Higher score means better items. They were also asked to provide their opinions over any important factors that the researcher may have failed to include. If the mean score of experts on any criterion for each item were less than 4.0, the items were considered to have some problems, and they were revised or removed. Based on the evaluation by experts, a revision of the item pool was conducted based and 66 items were left for the pilot test.

Step 6: Conducting a Pilot Test

The next step is to conduct a pilot test to detect unexpected errors within instruments and possible problems during administration process. For this goal, a pilot test was conducted through the UCF Community Counseling & Research Center during the spring semester of 2016. In total, 42 adult clients (male = 12, female = 29, other = 1) voluntarily participated in the pilot test for the CRCC, with their mean age of 33.07 (SD = 10.62). The result of factor analysis with the pilot data indicated that the 66-item CRCC was not unidimensional. The researcher reduced the number of sub-clusters into five basic listening skill clusters: reflection of feeling, reflection of contents, questioning, therapeutic attitude, and nonverbal skills, based on the rationale that

such advanced counseling skills as confrontation, advising, and self-disclosure are not frequently used by counselors in every session. Table 5 shows the revised blue print for the CRCC. And then, the items with higher factor loading for each cluster were selected and revised if needed; thus, five to nine items were included for each cluster. Finally, through this item revision, the CRCC had 36 items ready for the field test.

Table 5

Revised Variable Map for the 36-item CRCC

What is the latent variable?

The latent variable is the counselor competence that a counselor presents in session.

Logit Scale	Cluster	Observations [items]
		Hard item
5.00	Reflection of Feelings	Normalizing feeling
		Validating feeling
4.00		Using diverse feeling expression
3.00	Reflection of Contents	Accuracy of reflection
		Reflecting key contents
2.00		Good summarizing
1.00	Questioning	Appropriate exploration question
		Clear question
0.00		One question at once
-1.00	Therapeutic Attitude	Empathy
		Congruence (=Genuineness)
-2.00		Positive regards
-3.00	Nonverbal skills	Eye contact
		Physical distance
-4.00		Head nodding
-5.00		Easy item

What is the response format or rating scale used?

Likert scale are used ranging from 1 (Strongly disagree) to 4 (Strongly agree)

Step 7: Expert Review with the Revised CRCC

Another expert review was conducted with this 36-item CRCC. Four experts among the previous expert group participated in this review. This review also examined each item in terms of three dimensions- (a) importance, (b) relevance, and (c) clarity. As a result of this expert review, there was no item that was rated with less than 4.0 for three criteria.

Data Collection

The data in the study was collected during two different times- spring semester and summer semester of 2016. Before conducting any data collection, the researcher received the approval for collecting human-related data from the Institutional Review Board (IRB) of University of Central Florida (UCF). The following section presents specific steps of the data collection process including IRB approval, recruitment, and incentives.

Institutional Review Board Approval

After conducting the expert-review with the initial item pool, the researcher had submitted the current research protocols including informed consent and actual scale items to the Institutional Review Board (IRB) of University of Central Florida in November, 2015 and the IRB approval (IRB Number: SBE-15-11770). Approval for the data collection for this investigation was obtained on December 3rd, 2015. After receiving the IRB approval from the UCF IRB office, the first data collection for the pilot test was conducted during spring semester of 2016. In addition, the items were reduced and revised through this pilot study and it was also approved by the UCF IRB that the new-version instrument with revised items was used for the

further data collection during summer semester of 2016.

Pilot Data Collection

Participants for the pilot test were recruited from the Community Counseling Research and Center of University Central Florida through spring semester of 2016. The participants were adult clients who met student counselors in practicum training. The clients participated voluntarily in the pilot test study and had no incentives regarding the participation of this study. Administration of the survey took place after finishing their counseling session. Each participant received a packet of paper-written survey consisting of informed consent, the initial form of the CRCC with 66 items, and a demographic questionnaire, and completed the forms. It took about 20 minutes for completing all surveys. The survey packet was given to participants by their counselors and the counselors left the counseling room during the administration. And then, the completed instruments were collected via a locked research box in the clinic center, so that the counselors could not see the answers of their clients. In total, 42 adult clients participated in the pilot study; their average age was 33.07 (SD = 10.62); and gender was female (n = 29), male (n = 12), and transgender (n = 1). The collected data was analyzed and used for revising the items.

Main Data Collection

After revising items based on the pilot test results, the second, main data collection was conducted using the revised version of the 36-item CRCC through summer semester of 2016. The goal of the second data collection was aimed at conducting the Rasch analysis as well as evaluating the psychometric properties of the CRCC from Rasch model context. For this purpose, 83 participants were recruited from the Community Counseling Research and Center of

University Central Florida. Similar to the pilot test, the packet of instruments involved an informed consent, an instrument form of the revised 36-item CRCC, and a demographic questionnaire (i.e., age, gender, race). The participation of the study was voluntary and anonymous. There was no incentive for participating in this study.

The researcher visited all seven practicum sections and educated master's student counselors in each practicum about conducting the survey. The administration of the CRCC instrument to participants was conducted after finishing third counseling session with the clients. If there would be some difficulty for conducting the survey, the data collection after 4^{th} or 5^{th} session was allowed. Like the pilot data collection, the packet of instruments was given to clients by their counselors and the counselor left the room after a brief explanation of this survey. Each participant was asked to complete a set of survey forms, which took less than 15 minutes. The completed forms were collected via the same locked research box used for the pilot study, so that the counselors could not see the ratings of their clients. The test results were not shared with the participants and their counselors. All data collected was entered into the IBM SPSS 21.0 and the WINSTEPS (Linacre, 2016) for the data analysis.

Data Analysis

IBM SPSS 21.0 was used to conduct the descriptive analysis of the sample and the 36 items of CRCC. Some items (i.e., item 2, 4, 13, 15, 22) were reverse coded before performing the analysis. Additionally, this study used the WINSTEPS 3.92 version (Linacre, 2016) to evaluate the underlying psychometric properties of the 36 items of the CRCC in Rasch model: its content evidence, substantive evidence, structural evidence, generalizability evidence, and

interpretability evidence (Wolfe & Smith, 2007). The rating scale model (RSM) originally developed by Andrich (1978) was employed to analyze the data in this study. The RSM was developed to analyze the rating scale data with more than two response categories (Engelhard, 2013). The RSM is appropriate when all items use an equal rating scale and the distance between the response categories is intended to be same for all items (Kim & Hong, 2004; Ludlow et al., 2014). Since the CRCC is using a 4-point rating responses (i.e., strongly disagree, disagree, agree, and strongly agree) and all items are using the same rating scale, the Rasch analysis using the RSM was appropriate for this study. Specifically, the CRCC was analyzed for the properties of the item-total correlation and fit statistics of each item, rating scale functioning, person fit indices, appropriateness of item difficulty hierarchy, unidimensionality, reliability, differential item functioning across subgroups, and interpreting via person-item map.

Content Evidence

The content validity means the extent to which items of an instrument can represent the content of what it is intended to measure (Wolfe & Smith, 2007). Expert review and test specification are commonly provided as the content evidence. In terms of content validity, the Rasch model can evaluate the technical quality of each item with item-measure correlation and fit statistics (Wolfe & Smith, 2007).

Item-total Correlation

The item-total correlation refers to Pearson's correlation coefficient between the score on a single item and the total score of the remaining items. This value shows how an individual item is consistent with other items in the instrument. In order to demonstrate that each item

measures the same construct, the item-measure correlation should be positive and greater than .40 for polytomously scored items (Wolfe & Smith, 2007). In this study, the cutoff value of .40 was used to evaluate the item-measure correlation of each item in the CRCC.

Item Fit Statistics

As mentioned in the literature review section, item fit statistics indicate the degree to which items fit the model. That is, item fit indices detect abnormal patterns for each item by comparing expected response and observed response. The Rasch analysis provides two types of item fit indices: infit and outfit mean square statistics. In Rasch, infit index are weighted and sensitive to abnormal responses made by persons on the items that match their ability, while outfit mean square statistics are unweighted and sensitive to outliers. If an item has the fit mean square value of 1.0, it indicates that the item fit the model perfectly. When evaluating item fit indices, a range from 0.6 to 1.4 is considered to be acceptable for the rating scale model (Linacre, 2005). Thus, the current study used this range between 0.6 and 1.4 to detect misfitting items. For instance, if an item's fit mean square statistics fell out of the range, the item was considered misfitting.

Substantive Evidence

The substantive validity appraises how well theoretical frameworks underlying an instrument can work for respondents as the test developer intended (Wolfe & Smith, 2007). In the context of the Rasch, that evidence can take the investigation of rating scale functioning, examination of person fit statistics, and the degree to which the observed item difficulty hierarchy fits to the anticipated hierarchy.

Rating Scale Analysis.

The Rasch analysis provides a tool of examining rating category function. There are four requirements for appropriate rating scale (Linacre, 2005). Specifically, category frequencies, fit statistics for each category, observed average measure, and the thresholds of each category can be investigated. First, each rating category should contain a minimum of 10 observations to ensure the precision of the relevant indices. Second, the fit statistics for each category should be less than 2.0 (Linacre, 2004; Wolfe & Smith, 2007). Third, the average measures of each category should increase monotonically as the response categories move up. The average measure for each category is the empirical mean of the ability of the people who respond in that category (Linacre, 2005; Liu & Lee, 2015). Lastly, the thresholds, that is the intersection between adjacent categories, should increase monotonically and the measure difference between thresholds should be within 1.4 to 5.0. According to these criteria, how the 4-point rating scale of the CRCC functions to the sample was evaluated.

Person Fit Statistics

The Rasch model assumes that an examinee's guessing, carelessness, and misunderstanding could cause the person misfit. Thus, higher proportion of misfitting persons indicates that there is more noise for respondents to appropriately respond to items. This study determined the individuals with more than 2.0 infit or outfit statistics as misfitting persons (Linacre, 2016; Zaporozhets et al., 2015). The proportion of misfitting persons was reported in this study.

Item Difficulty Hierarchy

The Rasch analysis estimates the difficulty level for each item based on the observed responses of examinees. In this study, item difficulty indicates the extent to which clients agree with counselors' behaviors or attributes described in each item. As such, more difficult items describe the less agreeable behaviors of counselors regarding counselor competence. In other words, from client's perspectives, there are few counselors presenting good competence described in the difficult level item, which means that more difficulty items are related to more advanced characteristics of counselor competence compared to easier items. Therefore, with Rasch analysis's item difficulty concept, it is possible to investigate whether the observed item difficulty is consistent with that predicted hierarchy from the underlying theory upon which the instrument was developed. This study examined how well the empirical item difficulty hierarchy agreed with the variable map that had been created during the test specification phase.

Structural Evidence

The structural aspect of validity shows how the variable measured in an instrument is internally constructed. As the classical test theory provides the structural evidence for any instrument with the result of factor analysis, the Rasch model also offers that evidence to test developers with dimensionality analysis, evaluating whether the developing instrument is unidimensional or multidimenstional.

Dimensionality Analysis

The Rasch model originally views that a measurement should measure only one variable; thus, the intended structure within any measurement is unidimensional. Concerning this

investigation, WINSTEPS software provides the principle component analysis based on standardized residuals. The residual-based principal components analysis was conducted to examine the dimensionality of the CRCC. According to Linacre (2016), a measure should explain a minimum of 50.9% of variance to ensure the unidimensionality of the measure. If any contrast in unexpected variance has more than 3.0 eigenvalue, it indicates that the presence of the additional dimension needs to be investigated. This study employed these cut-off scores to examine the CRCC's unidimensionality.

Generalizability Evidence

Generalizability addresses how well tests maintain their function across diverse measurement contexts (e.g., subgroup's characteristics, administration environment). The Rasch model offers two types of evidence related to the generalizability aspect of validity: reliability for both item and person, and differential item functioning across subgroups.

Reliability

Investigation of reliability is the most commonly used method to ensure the generalizability of any measure (Wolfe & Smith, 2007). The internal consistency reliability of the CRCC was examined by two Rasch statistics of separation index and separation reliability for both person and item. Firstly, separation index shows the degree to which the measure is able to differentiate persons or items on the measured variables. The 2.0 or greater separation index is considered to be acceptable (Linacre, 2016). In addition, separation reliability, equivalent to Cronbach's alpha addresses the internal consistency reliability. The range of separation reliability is between 0.0 and 1.0. Separation index and separation reliability can be transformed

from each other. The separation index of 2.0 is equal to the separation reliability of 0.8. Thus, the separation reliability over 0.8 in this study was acceptable to ensure the consistency of a measure.

Differential Item Functioning (DIF)

The DIF statistics addresses whether the function of items is able to be maintained across subgroups of respondents and across time. The Rasch analysis examines the DIF index to determine whether or not individual items of the instrument work differently across the different groups or contexts. This study investigated the DIF statistics for each item to evaluate if there is any item functioning differently across gender of respondents. Specifically, the size of the difference in average measure between male and female group and its significance were assessed with the DIF statistics.

Interpretability Evidence

The interpretability validity addresses the degree to which the meaning of measures is clearly communicated to those who want to interpret the measures (Wolfe & Smith, 2007). The Rasch model offers various figures (e.g., person-item map, kid map) that provide a great deal of information concerning the interpretation of the observed data. Among them, this study included the interpretability of the CRCC using the person-item map.

Person-item Map

The result of Rasch analysis using the WINSTEPS produces person-item map, which graphically illustrate how person ability distribution overlaps with item difficulty distribution (Liu & Lee, 2015). If there is a sufficient overlap between person ability and item difficulty

distribution, it would demonstrate that the item difficulty level is appropriate to measure the ability of persons in the sample. Additionally, considering a person's ability level and item difficulty level together helps predict how the person would answer each item. This study examined the distribution of counselor competence perceived by clients and item difficulty on the person-item map. The interpretability of a counselor's competence measure assessed by clients with the CRCC was discussed with the person-item map.

Summary

In summary, the data analysis in the current study was performed to point out diverse aspects of validity of the CRCC through the Rasch analysis. Specifically, this data analysis involved the content, substantive, structural, generalizability, and interpretability evidences. Table 6 shows the types of validity evidences analyzed for the CRCC in this study (Wolfe & Smith, 2007).

As mentioned earlier, the Rasch analysis is able to provide various statistics to examine diverse aspects of validity in the developing measure; the classical test theory cannot analyze most of those aspects. This study used the latest version of WINSTEPS software (Linacre, 2016) to investigate those validity evidences in the CRCC. Given the main goal of this study in using the Rasch model for developing a valid instrument, the subsequent chapters will focus on the discussion and present ways of examining and interpreting the results provided by the Rasch analysis.

Table 6

Rasch Analysis Evidence Relevant to Validity Aspects

Validity Aspect	Rasch Analysis Evidence
Content	• Item Technical Quality (Item-total correlation, Item fit statistics)
Substantive	• Rating Scale Function Analysis • Person Fit Statistics • Item Difficulty Hierarchy
Structural	• Dimensionality Analysis
Generalizability	• Reliability (Person/ Item separation index) • Differential Item Functioning
Interpretability	• Person-item Map

CHAPTER FOUR:
FINDINGS

The purpose of this study was to develop a valid and reliable instrument to measure clients' perception of counselor competencies within a therapeutic environment. A second part of the study was to examine the psychometric properties of the developed instrument, using the Rasch analysis. This chapter presents the results of the data analyses regarding development and validation of the newly developed measure, the Client Ratings of Counselor Competence (CRCC). This chapter includes item response frequency as well as diverse validity evidences of the CRCC that the Rasch analysis provided. According to the classification of evidence suggested by Wolfe and Smith (2007), the CRCC's content evidence, substantive evidence, structural evidence, generalizability evidence, and interpretability evidence were presented using various statistics and figures of the Rasch analysis.

Descriptive Statistics

Table 7 illustrates the frequency of responses on each items in the CRCC. This table shows that most participants positively responded to all the items, answering strongly agree (n = 2015, 66.7%), agree (n = 892, 29.5%), disagree (n = 64, 2.1%), strongly agree (n = 26, 0.9%), no response (25, 0.8%). This result means that participants in this study gave to their counselors good scores on most items in the CRCC.

Table 7

Frequency of Responses to the CRCC

Item	No Response	Strongly Disagree	Disagree	Agree	Strongly Agree
1	0	0	1	13	70
2 R*	2	2	2	36	42
3	0	0	1	11	72
4 R	0	0	2	11	71
5	1	0	0	13	70
6	0	1	0	35	48
7	0	1	0	19	65
8	0	1	2	26	55
9	0	0	1	26	57
10	1	0	3	38	42
11	0	0	1	26	57
12	1	0	1	24	58
13 R	1	0	3	25	55
14	0	0	2	24	58
15 R	0	4	0	15	65
16	1	3	0	40	40
17	1	0	0	22	61
18	0	1	1	10	72
19	1	0	2	21	60
20	0	0	1	12	71
21	0	1	1	21	61
22 R	1	4	6	12	61
23	0	0	2	23	59
24	3	1	3	23	54
25	1	0	3	32	48
26	0	0	3	44	37
27	0	0	3	24	57
28	0	1	3	41	39
29	1	0	2	18	63
30	0	0	2	34	48
31	0	0	1	25	58
32	0	2	2	29	51
33	2	0	4	27	50
34	2	0	2	35	44
35	2	1	2	27	51
36	4	3	2	30	45
Total	25 (.8%)	26 (.9%)	64 (2.21%)	892 (29.5%)	2015 (66.7%)

* Reverse items are listed with an R.

There were 5% or less missing responses across the CRCC items, which did not influence the analysis results based on the < 5% rule of thumb in statistical analysis (Tabachnick & Fidell, 2007). Specifically, the largest number of missing data was only three responses (3.8%) for both item 24 and item 36. Furthermore, the Rasch analysis is robust for missing data because it uses empirical response patterns rather than raw test scores; thus, the missing data was not an issue in this study.

Content Evidence

Technical Quality of Items

The technical quality of the items in CRCC was evaluated based on the field test responses. When evaluating the quality of any item, the Rasch analysis reveals two types of indices: the item-total correlation and the fit mean square statistics (Wolfe & Smith, 2007). Table 8 shows the item-measure correlation and item fit statistics of each item. The item-total correlation, also called the point-biserial correlation, means the Pearson correlation coefficient between the item and the total raw score. According to the recommendation of Wolfe and Smith (2007), the item-total correlations should be positive and more than .40 for polytomously scored items. In terms of this cutoff value, although the item-measure correlation of all items is positive, item 1, 4, 15, 16, and 22 had .40 or less value of their item-measure correlation, indicating that these five items should be removed or revised from the scale.

Table 8

Item Misfit Statistics and Item-Total Correlation

No	Item	Infit	Outfit	ITC
1	maintained good eye contact with me.	1.07	1.82	0.37
2	did *not* detect my deeper feelings.	1.30	1.88	0.42
3	was open to talking about any feeling that I expressed.	1.02	0.86	0.41
4	made me feel interrogated by his/her questions.	1.31	1.14	0.36
5	maintained a gentle tone of voice.	0.82	0.60	0.45
6	was honest and frank.	0.70	0.64	0.64
7	responded to me warmly.	0.72	0.51	0.56
8	organized my thinking about what happened in the session.	0.95	0.81	0.57
9	asked me to give more details about my topic.	0.82	0.64	0.58
10	normalized the feelings I was having.	0.85	1.21	0.56
11	was curious about hearing my story.	0.71	0.60	0.61
12	gave me enough time to think after questioning.	0.67	0.59	0.62
13	asked too many questions at the same time.	1.22	1.95	0.42
14	summarized what I said so that I could understand my situation more clearly.	0.69	0.55	0.64
15	used *inappropriate* head nodding.	2.70	3.49	0.23
16	mirrored the key content of what I said.	1.63	1.65	0.38
17	showed open and welcoming gestures.	0.70	0.53	0.59
18	actively listened to what I said.	1.23	0.66	0.49
19	asked me questions in a clear way.	0.92	1.00	0.53
20	seemed to be genuine with me.	0.85	0.56	0.50
21	explored important issues with me.	1.00	1.02	0.54
22	imposed his/her values on me.	2.80	3.09	0.30
23	summarized the main points of what we discussed.	0.81	0.79	0.58
24	used more open questions than “yes or no” questions.	1.30	1.82	0.48
25	helped me identify my underlying feelings.	0.62	0.58	0.69
26	used a variety of feeling words to describe my emotions.	0.84	1.00	0.57
27	asked questions that helped me explore what I was thinking or feeling.	0.66	0.52	0.67
28	helped me label my feelings.	0.86	1.11	0.58
29	provided a comfortable physical distance between us.	0.95	0.59	0.56
30	understood exactly what I meant.	0.54	0.53	0.71
31	seemed to think what I said was important.	0.63	0.51	0.65
32	validated my feelings.	1.25	1.80	0.52
33	fully understood my unique situation and values.	0.82	0.83	0.63
34	precisely identified my feelings.	0.46	0.53	0.74
35	accurately rephrased what I said in his/her own words.	0.97	0.95	0.60
36	repeated back a concise version of what I said.	1.34	1.33	0.55

Note

Next statistics investigated were the item fit statistics. By applying the acceptable range of 0.6 to 1.4 (Bond & Fox, 2001), the infit mean squares were examined to identify items that were misfitted to the Rasch model. That is, the items with less than 0.6 or more than 1.4 infit mean squares were considered to be misfit items. Applying this criterion, items 15, 16, 22, 30, and 34 were identified as misfitted items, indicating that total five items should be removed or revised from the scale. Given both results of the item-measure correlations and item fit indices, totally seven items were problematic: item 1, 4, 15, 16, 22, 30, and 34.

Substantive Evidence

Rating Scale Analysis

The current CRCC instrument used a 4-point rating scale. To determine whether the categories (i.e., strongly disagree, disagree, agree, strongly agree) functioned as intended, the functioning of the CRCC's four response categories were diagnosed.

Table 9

Summary of the Rating Scale Category Structure for the Original 4-Point Rating Scale

Category Label	Frequency Count	Observed Average	Infit MNSQ	Outfit MNSQ	Andrich Threshold
1. Strongly disagree	25	1.81	1.99	4.55	None
2. Disagree	64	.76	1.01	1.38	-.60
3. Agree	892	1.72	.86	.76	-1.35
4. Strongly Agree	2015	3.65	.91	.93	1.95

The properties of the response categories of the CRCC are presented in Table 9. The frequencies in all category responses exceeded the recommended minimum number of 10 (Linacre, 2002a). However, although both infit and outfit mean squares for other three categories (i.e., disagree, agree, strongly disagree) were less than the cutoff value of 2.0 (Linacre, 2002a), the infit and outfit mean-squares for the "strongly disagree" category were 1.99 and 4.55, suggesting that the category included noise that would have brought misinterpretation (Linacre, 2002a). In addition, the average measure of the "strongly disagree" was higher than that of "disagree", even "agree", which indicates that they were not functioning properly. Under the Rasch model, the threshold estimates should increase theoretically as the category order and the difference between two thresholds should be more than 1.4 logits. The threshold is the intersection where two adjacent category probability curves meet. For example, the first threshold is the point where Category 1 (strongly disagree) and Category 2 (disagree) meet. Likewise, the second threshold is the intersected point between Category 2 and 3, the third threshold between Category 3 and 4. The result shows that the first threshold (-.60) was higher than the second threshold (-1.35), as well as the gap (.75 logits) between both thresholds was less than 1.4 logits, which indicates that the 4-point scale category used in the CRCC did not function well.

Figure 1 shows the category probability curve for each category in the CRCC, illustrating the probability of responding to a specific category given the differences in estimates between person trait scores and item difficulties. Figure 1 demonstrates that Category 2 (disagree) had a

low probability to be endorsed at any given point of the measure, indicating that the Category 2 did not function as a distinct rating scale structure.

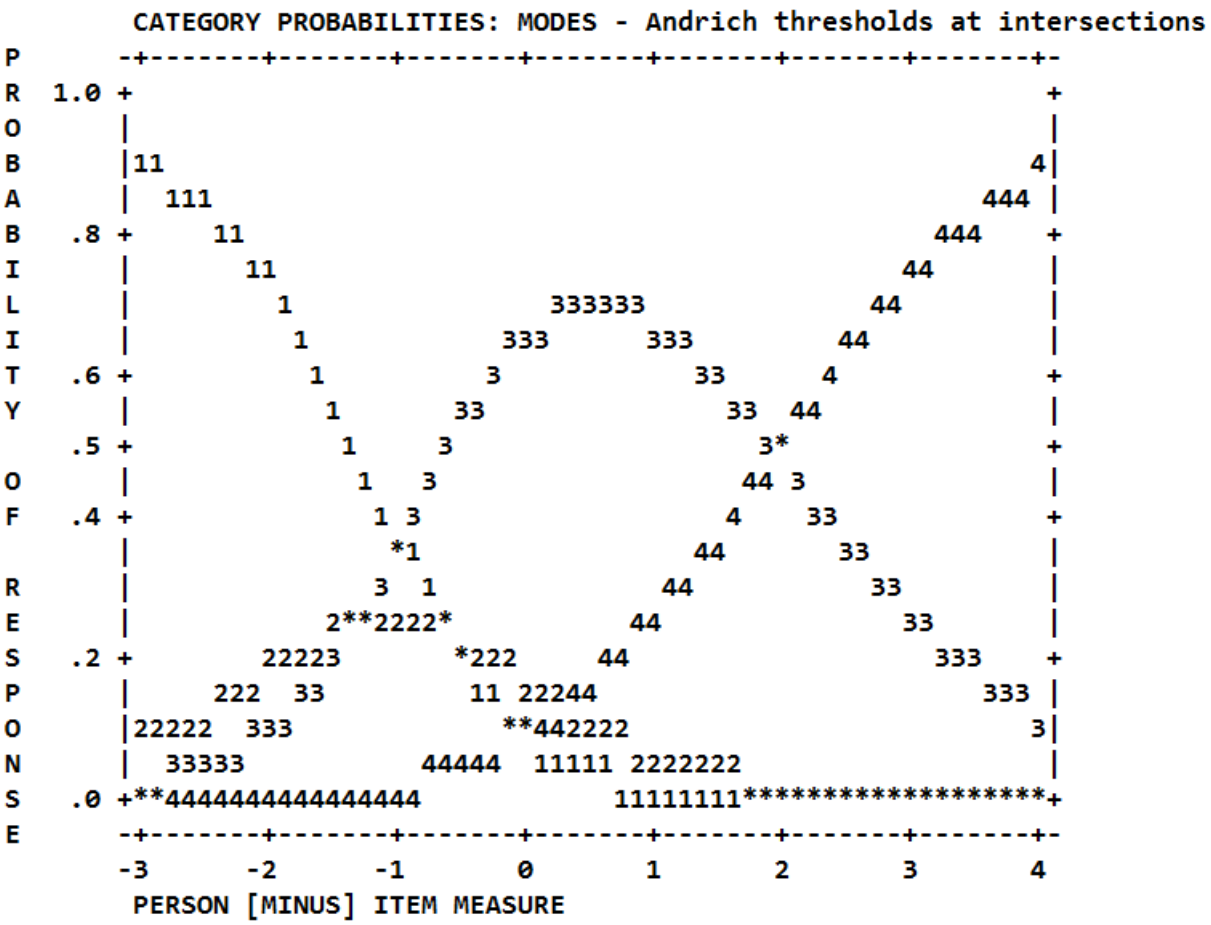


Figure 1

Rating Scale Probability Curves for the Original 4-Point Rating Scale

After trying several options to collapse the categories, the researcher selected combining

the original Category 1 (strongly disagree) with Category 2 (disagree) to optimize the rating scale functioning. After recoding the data, the rating scale functioning analysis for the revised 3-point scale format indicated better values for most statistics. In specific, the frequencies for each category became relatively uniform, as well as was met for the requirement of the minimum of 10 observations for each category (see Table 10). The observed average of categories also increased monotonically. For the threshold, the distance between two thresholds was within the required range from 1.4 to 5.0.

Table 10

Summary of the Rating Scale Category Structure for the Revised 3-Point Rating Scale

Category Label	Frequency Count	Observed Average	Infit MNSQ	Outfit MNSQ	Andrich Threshold
1. Strongly disagree & Disagree	89	.42	1.38	2.93	None
3. Agree	892	1.15	.88	.74	-1.31
4. Strongly Agree	2015	2.92	.91	.95	1.31

As shown in Figure 2, each response category had the highest probability at some points. Moreover, the fit statistics for the original Category 3 (agree) and 4 (strongly agree) were within expectation, whereas that of the newly collapsed rating scale category still showed the misfit result, with the outfit mean square of 2.93. This result suggests that the way of collapsing the original rating scale Category 1 and 2 could be an alternative to improve the rating scale of CRCC.

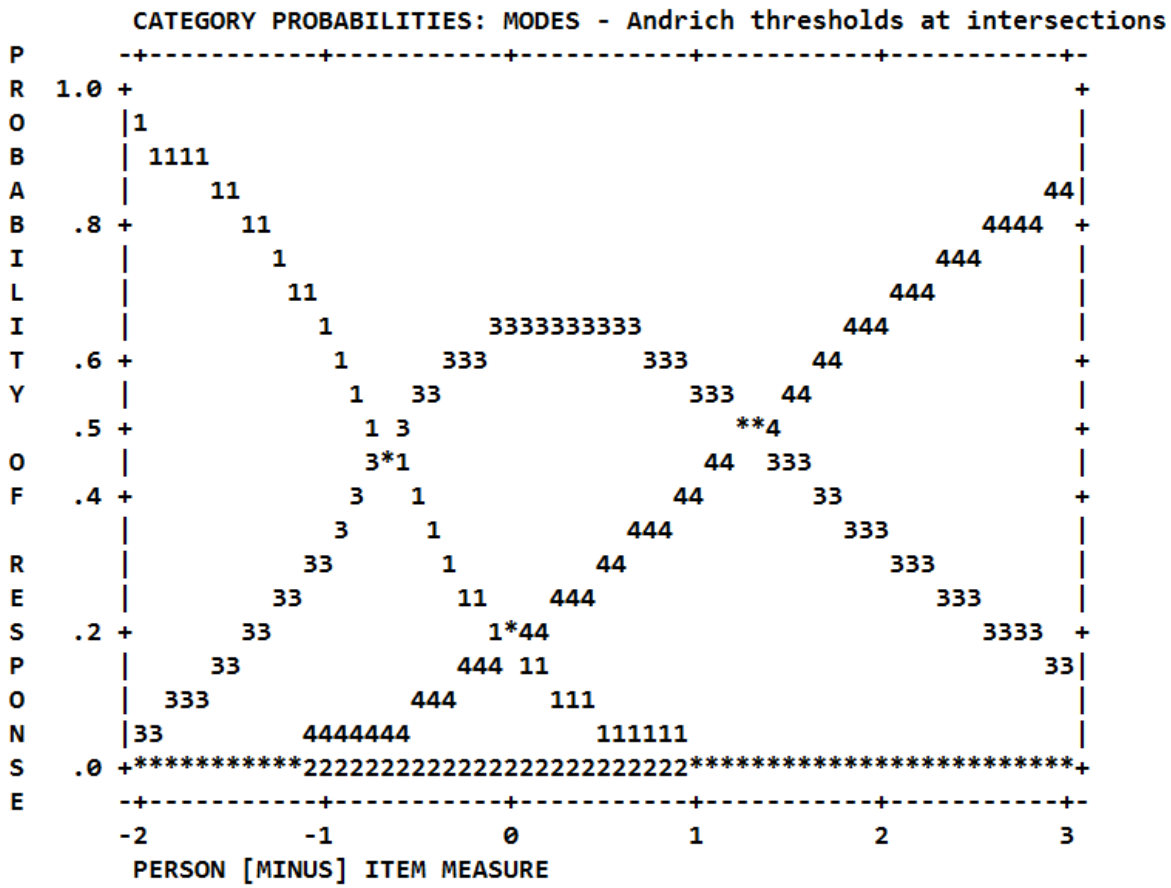


Figure 2

Rating Scale Probability Curves for the Revised 3-Point Rating Scale

Person Fit

Person fit statistics, the agreement between the expected responses and the observed responses of a respondent can provide additional evidence to support the substantive validity of a measurement in the Rasch model (Wolfe & Smith, 2007). By investigating the person fit indices, the test developers can confirm that examinees' response process is consistent with the

framework constructed by the test developers (Wolfe & Smith, 2007). Under the Rasch context, person misfit could happen due to unexpected responses of examinees such as guessing, carelessness, item bias, and specialized knowledge. In this study, people with the infit and outfit mean squares over 2.00 (Linacre, 2012b) were identified to misfit persons.

The result showed that 12% of the respondents (n = 84) showed misfit, with the mean square infit and outfit statistics above 2.0. This result indicates that 10 clients may have difficulties to clearly understand the CRCC items as intended by this scale developer, suggesting that more investigation on the items with the person's misfit response is needed.

Item Difficulty Hierarchy

According to Wolfe and Smith (2007), another way to provide substantive evidence using Rasch model is to examine the hierarchical structure of item difficulty parameters. The underlying hierarchy framework identified in the test specification may specify that items should be calibrated in a particular hierarchy of difficulty based on an expected linear sequence (Wolfe & Smith, 2007). That is, the adequacy of the item calibration can be inspected with each item's difficulty level measures. In this study, item difficulty parameter refers to willingness of the participant to agree with the statements in the instrument (Liu, 2010), indicating items with lower difficulty level are more likely to be agreed positively by participants. Given the CRCC intends to measure counselor competence, easy items describe the behaviors or attributes of counselor competence that most counselors show in counseling sessions. Aversely, difficult items in the CRCC states the competence attributes that a few counselors demonstrate in the sessions, meaning that more difficulty items are related to more advanced counselor competence.

Table 11 shows the ordering of item difficulties for the 36 items in the CRCC. Two items to measure the reflection of feeling- Item 26 "My counselor used a variety of feeling words to describe my emotions" and Item 28 "My counselor helped me label my feelings", are the highest (the least frequent behavior) logits (1.02 logits) among the 36 items, whereas the item 3, "My counselor was open to talking about any feeling that I expressed" (-1.29 logits) is the lowest (the most frequent behavior). Overall, the nonverbal counseling skills (e.g., eye contact, physical distance, tone of voice) were the easiest cluster among the five types of counseling skills, while the reflection of feeling were the most difficult cluster. This indicates that there were many counselors showing good nonverbal skills, while few counselors showed good reflection of feeling skills. This result was consistent with the test specification structure in this study, viewing that reflection counseling skills may be located higher than questioning and nonverbal skills. However, for the therapeutic attitude cluster, the overall item difficulty levels for items in this cluster were not consistent with the assumed hierarchical level. In particular, Item 6 (was honest and frank), Item 30 (understood exactly what I meant), Item 22 (imposed his/her values on me), and Item 33 (fully understood my unique situation and values) showed much higher item difficulty level than expected, which was similar to the level of reflection of contents, even the level of reflection of feeling. This wrong item calibration suggests that these items could be excluded. Likewise, Item 3 (D = -1.38) with a wrong item calibration was out of the range value because it was supposed to be located in the highest level of "Reflection of Feeling". This finding supports that the item 3 may have some problems and could be excluded, too. In addition, Item 15 had too higher difficulty level, compared to those of other items in the same

cluster. When the item difficulty level for all items in the CRCC in the same way, this result suggested totally eight items including item 3, 4, 6, 15, 22, 30, and 33 should be excluded or revised in the next revision of the CRCC scale. Although Items 4, 15, 22, and 30 were already detected from item technical quality analysis, Item 3, 6, and 33 were newly detected via this assessment of item difficulty hierarchy.

The item-person map in Figure 3 graphically illustrates the relative level of person ability and item difficulty parameters on the CRCC linear scale using logits. Specifically, persons with higher level of counseling abilities are located in higher place in the map, meaning higher logit scores. Similarly, items with more difficult level are calibrated at higher locations. For instance, items 28, 26, and 16 are the three most difficulty items, while items 3, 5, and 20 are the three least difficult items in the CRCC.

Table 11

Item Difficulty Hierarchy of CRCC: Measure Order

Item	Level	Item	Measure (Logits)	SE
26	5	used a variety of feeling words to describe my emotions.	1.02	0.20
28	5	helped me label my feelings.	1.02	0.20
16	4	mirrored the key content of what I said.	1.00	0.20
2*	5	did *not* detect my deeper feelings.	0.90	0.21
36	4	repeated back a concise version of what I said.	0.79	0.22
10	5	normalized the feelings I was having.	0.77	0.21
34	5	precisely identified my feelings.	0.59	0.22
25	5	helped me identify my underlying feelings.	0.53	0.22
32	5	validated my feelings.	0.52	0.22
6	2	was honest and frank.	0.47	0.22
30	2	understood exactly what I meant.	0.47	0.22
22*	2	imposed his/her values on me.	0.42	0.23
33	2	fully understood my unique situation and values.	0.38	0.23
35	4	accurately rephrased what I said in his/her own words.	0.32	0.24
24	3	used more open questions than “yes or no” questions.	0.25	0.24
8	4	organized my thinking about what happened in the session.	0.21	0.24
13*	3	asked too many questions at the same time.	0.15	0.24
27	3	asked questions that helped me explore what I was thinking or feeling.	0.04	0.24
9	3	asked me to give more details about my topic.	-0.09	0.25
11	2	was curious about hearing my story.	-0.09	0.25
14	4	summarized what I said so that I could understand my situation more clearly.	-0.09	0.25
15*	1	used *inappropriate* head nodding.	-0.15	0.25
23	4	summarized the main points of what we discussed.	-0.15	0.25
31	2	seemed to think what I said was important.	-0.15	0.25
12	3	gave me enough time to think after questioning.	-0.16	0.26
21	3	explored important issues with me.	-0.21	0.26
19	3	asked me questions in a clear way.	-0.23	0.26
17	1	showed open and welcoming gestures.	-0.44	0.27
29	1	provided a comfortable physical distance between us.	-0.49	0.27
7	2	responded to me warmly.	-0.73	0.29
1	1	maintained good eye contact with me.	-1.08	0.31
4*	3	made me feel interrogated by his/her questions.	-1.08	0.31
18	2	actively listened to what I said.	-1.08	0.31
20	2	seemed to be genuine with me.	-1.18	0.32
5	1	maintained a gentle tone of voice.	-1.19	0.33
3	5	was open to talking about any feeling that I expressed.	-1.29	0.33

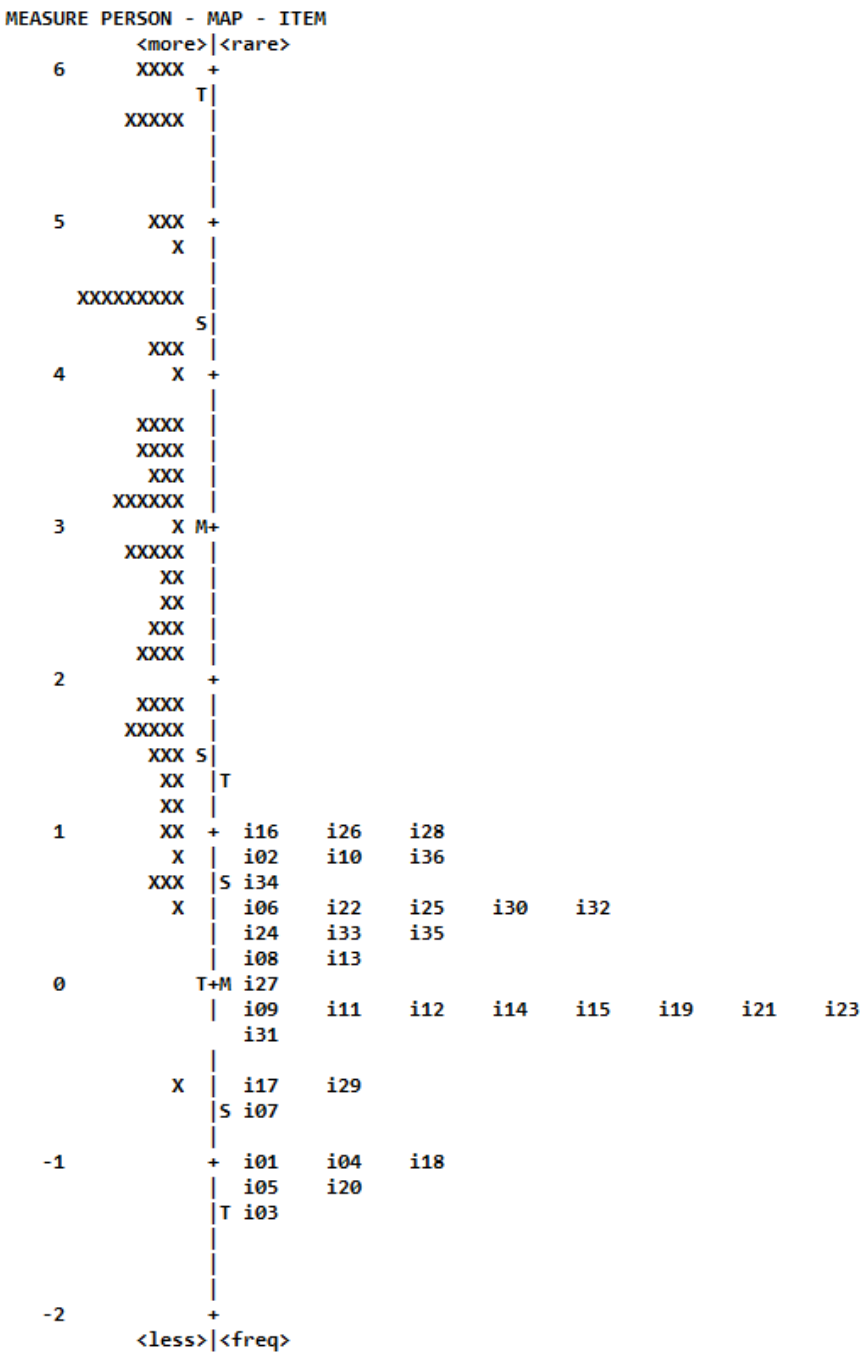


Figure 3

Item-Person Map in the CRCC

With the map, we are easy to identify the cluster of items having similar difficulty level. Based on Figure 1, items with the same location mean that their difficulty levels are very close to each other, even equal. For instance, the item-person map shows that Items 6, 22, 25, 30, and 32 had a common item difficulty level. Item 9, 11, 12, 14, 15, 19, 21, and 23 were also considered to be another item group with a similar difficulty level. In the Rasch, including many items with a similar item difficulty level causes item redundancy in a measure. Thus, it is recommended in the Rasch model that items showing similar item difficulty level be reduced to lessen the item redundancy. Therefore, by considering both the item difficulty measure and its attribute cluster, reducing the number of items in the CRCC is needed. The specific procedure for this is presented in the discussion.

Structural Evidence

Dimensionality Analysis

The WINSTEPS software provides the automated process using principal component analyses to determine whether the measure is unidimensional. After extracting the explained variance of the primary component from data, the WINSTEP performs additional principal component analysis with the standardized residuals to investigate the possibility of additional dimensions in the data (Linacre, 2016). According the guideline of Linacre (2016), contrasts with 3.0 or higher eigenvalues implied the possible presence of additional dimension that can explain substantial variance in the data. In addition, the minimum criterion of 50.9% (Linacre, 2016) was used to determine the unidimensionality of the CRCC.

Table 12

Summary of Dimensionality Analysis

	Eigenvalue Units	%	Variance Unexplained (%)
Total raw variance in observations	55.38	100.0	
Raw variance explained by measures	19.38	35.0	
Raw unexplained variance (total)	36.00	65.0	100.0
Unexplained variance in 1st contrast	2.96	5.4	8.2
Unexplained variance in 2nd contrast	2.50	4.5	7.0
Unexplained variance in 3rd contrast	2.33	4.2	6.5
Unexplained variance in 4th contrast	2.02	3.7	5.6

Table 12 shows the result of the principal component analysis of the residuals in the CRCC. There were no contrasts with 3.0 or higher eigenvalues; however, the proportion (35.0%) of the variance explained by measures was less than the minimum value of 50.9%, not supporting unidimensionality of the CRCC. This fact implies that that the CRCC may not be unidimensional and there may be another dimension in the current CRCC. Therefore, more investigation is needed after removing misfit items or persons, or revising items with some issues.

Generalizability Evidence

The Rasch analysis provides two evidence for the generalizability validity. One is

reliability index and another one is the differential item functioning (DIF) index. For reliability, the WINSTEPS using Rasch analysis provides separation index and separation reliability index for both persons and items, which shows how well and consistently a measure can discriminate persons and items. The DIF was investigated for gender groups.

Reliability

Reliability analysis measures the consistency of instrument across scoring designs, similar to internal consistency reliability by Cronbach's alpha. The current CRCC's internal consistency reliability was assessed with *separation index* and *separation reliability*. The summary of person and item reliability estimates is shown in Table 13.

Table 13

Person and Item Reliability Summary Statistics

Parameter	Average Measure	True SD	RMSE	Separation	Reliability
Person	3.15	1.50	.67	2.25	.84
Item	0.00	.61	.27	2.24	.83

In specific, person separation reliability estimate for the CRCC was .84, indicating that the CRCC can adequately differentiate individuals. The value of item separation reliability was observed to be .83, suggesting that the separation of item difficulty is reliable. Additionally, the separation indices for both persons and items were greater than 2.00;

specifically, person separation index was 2.25 and item separation index was 2.24. These values suggest adequate separation between persons as well as between items on the CRCC.

Differential Item Functioning

The functioning of items should be maintained regardless of sub-groups of respondents or measurement time, in order to appropriately generalize the results of the measure. The differential item functioning (DIF) in the Rasch analysis determines whether individual items of instrument work differently across the different groups or contexts. That is, the DIF index examines whether each item functions in the same way across sub-groups or contexts in the sample.

This study examined the DIF of each item across gender, and the result indicates that the DIF for most items were not significant, except for Items 6, 15, and 19, 22, and 33 (see Table 14). The logit gender difference for these five items was ranged from 1.04 to 1.56, which was statistically significant ($p < .05$). This result indicates that most items in CRCC functioned in the same manners, regardless of the gender of a respondent, while only five items (i.e., items 6, 15, 19, 22, 33) worked differently according to the respondent's gender. In specific, for Items 6, 19, and 33, female group's average measure was significantly higher that male group's. For Items 15 and 33 items, the mean of male clients was significantly higher than that of female clients. Thus, this analysis suggests that Items 6, 15, 19, 22, and 33 should be removed or revised from the scale. While other items showed some problems in previous evaluations, the item 19, "My counselor asked me questions in a clear way", was newly detected from this DIF investigation.

Table 14

Differential Item Functioning (DIF) Size

Item No.	Item	Male	Female	DIF Size	P
	(During this session, my counselor…)				
6	was honest and frank.	-.25	.95	-1.20	< .05
15	used inappropriate head nodding.	.44	-1.22	1.56	< .01
19	asked me questions in a clear way.	-1.18	.34	-1.52	< .05
22	imposed his/her values on me.	.99	-.25	1.23	< .05
33	fully understood my unique situation and values	-.29	.79	-1.04	< .05

Interpretability Evidence

Person-item map

The person-item map conveyed a great deal of information concerning the appropriateness of the items for the target population (see Figure 4). Overall, in the CRCC, the distribution of the item difficulty measures is lower than counselors' latent trait, indicating that most practicum counselor students used good counseling skills in sessions.

If using the person-item map, any person can interpret the result of the CRCC easily. For example, the ability level rated by the person 44, who had the lowest CRCC measure, was located similar to the item difficulty in questioning skill level. This result means that the use of nonverbal skills was appropriate, while the use of other higher skills like reflection skills was

poor in session. The developmental level of counseling skills for the counselor rated by Person 44 could be around the stage to need questioning skills. Thus, more training on questioning skills might is needed for the counselor.

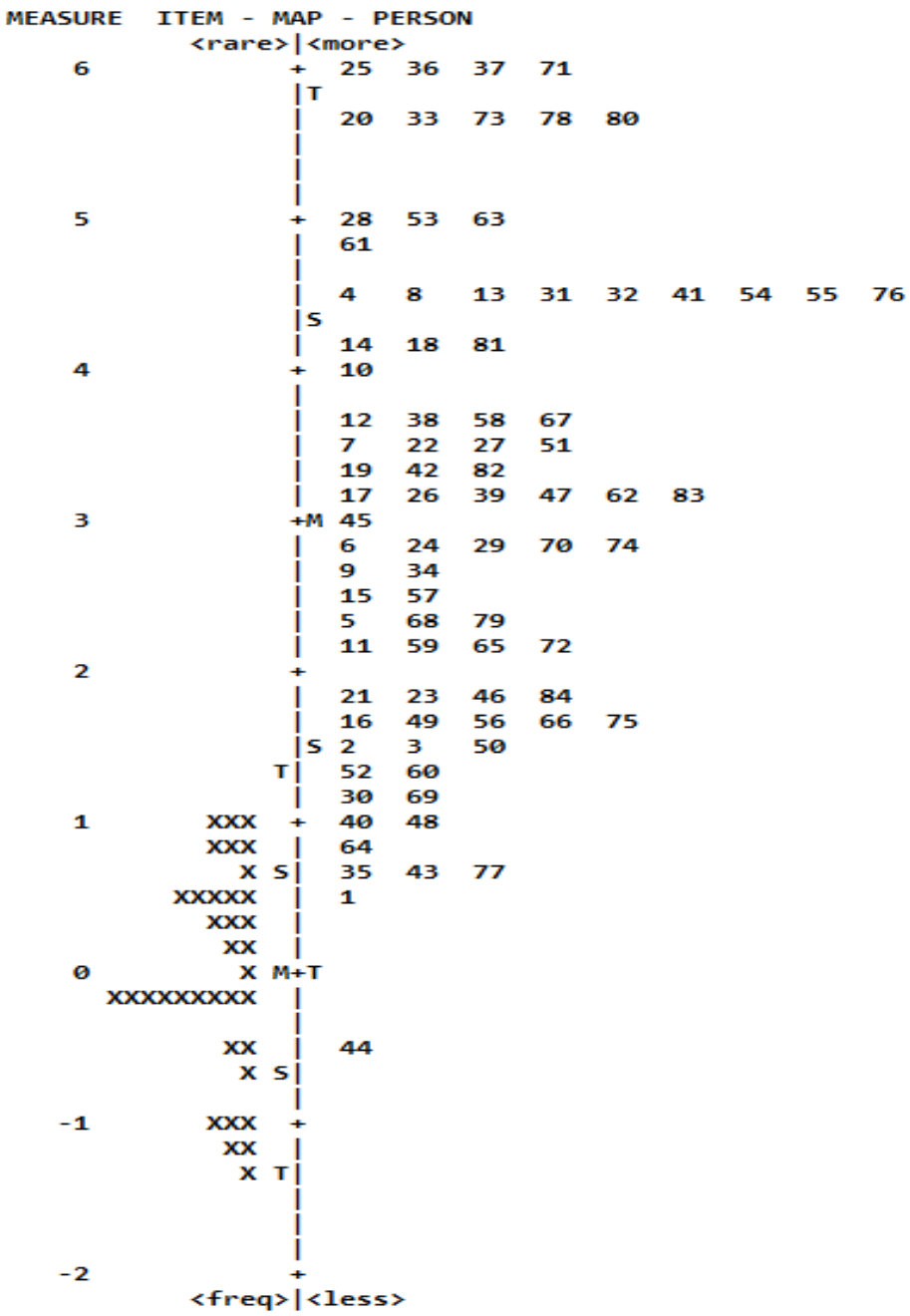


Figure 4

Person-Item Map in the CRCC

CHAPTER FIVE: DISCUSSION AND CONCLUSION

The main purpose of this study was to examine the psychometric properties of the Client Ratings of Counselor Competence (CRCC) via the Rasch measurement model. This chapter discusses the validity evidences of the CRCC that we found through the Rasch analysis. In addition, the practical implications of this research, its limitations, and the recommendations for future research are discussed.

Discussion of Results

Content Evidence

n terms of the content aspect of validity in the CRCC, two item statistics−the item-total correlations and the fit statistics for each item−were evaluated to determine how well individual items represent the variable to be measured and fit to the model.

Firstly, the investigation regarding the item-total correlation indicates that Items 1, 4, 15, 16, and 22 have the problematic item-total correlation lower than the acceptable value of .40. Given three items (i.e., Items 4, 15, 22) among those five items are negatively worded, it is possible that negative wording could produce some noise in delivering the meaning as intended.

Secondly, the result of item fit statistics revealed several misfitting items that were unable to provide meaningful information for the model. Specifically, this result indicates that Items 15, 16, 22, 30, and 34 need to be excluded or revised from the scale. When taking a closer look at the descriptions of the misfitting items, it is observed that two highest misfitting items− Items 15

and 22−are reverse items, suggesting that negatively phrased statements do not express the intention of the descriptors well. This result is consistent with that of Liu and Lee (2015). Although those reverse items were included for obtaining more genuine response of respondents, the negative wording items seem to rather arise elicit wrong or misfit responses from respondents or even interrupt respondents' clear understanding of items. This result supports the recommendation that social science questionnaires should avoid the use of negative wording because respondents tend to disagree with items that describe negative behaviors to be in accordance with social desire or preference (Liu & Lee, 2015). However, it is uncertain whether item wording or other construct-irrelevant contents cause the misfitting items. To elaborate this, Liu and Lee (2015) suggested further study to investigate how the revised items functions after rewording all negative wordings into positive wordings.

Both investigations (i.e., item-measure correlation, fit statistics) of item quality were able to detect that Items 15, 16, and 22 have some problems. However, Items 1 and 4 were detected only from evaluating item-total correlations, while Items 30 and 34 were identified from investigating each item's fit indices, indicating that each statistic (i.e., item-total correlation, item fit statistics) can examine different aspects of item quality. Item-total correlation widely used for evaluating tests comes originally from the CTT's theoretical model, whereas the item fit statistics is the Rasch model's unique concept. Using the item fit statistics of the Rasch is more beneficial when items' functions are not enough empirically validated (Christensen, Engelhard, & Salzberger, 2012), especially in developing a new measure or instrument. Therefore, this result suggests that when evaluating technical quality of items in a scale, both item-measure correlation

and item fit statistics could complement each other, so to evaluate the item quality more precisely and more comprehensively.

Substantive Evidence

Rating scale function, person fit, and item difficulty hierarchy were examined to investigate the substantive validity of the CRCC. The rating scale analysis in the result indicates that the current four-point rating response category in the CRCC did not function well. The percentage of category 3 (agree) and category 4 (strongly agree) was 29.77% and 67.26%; both were almost 97% of all responses, indicating that most respondents gave good scores on every item in the CRCC. This result might be related to the setting of the data collection because all clients participating in this study received free counseling services from counselor trainees, and this fact could put some pressure on the participant giving good score to their counselors-in-training. More investigations with diverse samples in different settings will be needed for verifying this hypothesis. Moreover, the average measure and the threshold for each category did not increase monotonically, suggesting that the category structure functioned inappropriately to the respondents. According to the recommendation by Linacre (2016), the researcher tried to collapse the category 1 (strongly disagree) and category 1 (disagree) that did not work well in the original scale analysis. The revised 3-point scale by collapsing two categories showed better rating scale function, although there was a minor problem in the fit index for the first scale category. This fact suggests that revising the scale format (i.e., wording, number of point) in the CRCC need to be considered. For instance, the 3-point scale format rating with poor, moderate, and good could be a possible alternative format.

In terms of evaluation of person fit, the result shows that 12% respondents answered unexpectedly to items, which indicates that more than 10% of participants were unable to clearly understand the meaning items as intended by the scale developers. When more closely investigating the response of those misfitting persons, their most misfitting responses were related to Items 15, 16, 22, and 36, which are the same items detected in item-fit statistics except for Item 36. This fact indicates that such items as Items 15, 16, 22, and 36 in the CRCC were not clearly understood by clients in this study. Specifically, negative wordings like "inappropriate", "impose" as well as unclear wordings like "mirrored", "concise" might be some issues for the clarity to the respondents.

Additionally, the result of item difficulty hierarchy in the CRCC shows that the items were hierarchically located by reflection skills, questioning, and nonverbal skills, which is consistent with Ivey et al. (2013)'s theoretical framework. This empirical evidence could support the conceptual framework that micro counseling skills are hierarchically located (Ivey et al., 2013). However, such an interpretation is posited cautiously, since item difficulty hierarchy just means that a counselor's behaviors related to reflection skills are more difficult for clients to observe in session, compared to the use of questioning or nonverbal skills. Less observations of reflection skills could be associated with other aspects not related to counselor competence (e.g., counseling theory orientation, difficulty to perceive). In other words, we cannot say that less observable behaviors are more advanced or developed ones with only this result. Thus, in order to confirm the fact that reflection skills are a higher attribute on a linear scale of counselor competence, further investigations are needed to examine the relationship between the CRCC

and other instruments to measure the same construct. This type of validity evidences is classified as external validity by Wolfe and Smith (2007). If there exists a strong, positive correlation between them, it could support the adequateness of the hierarchical attributes found in this study. Nevertheless, it is still meaningful that the Rasch model can be a useful tool to illustrate the underlying conceptual structure in any latent variables like counseling skills and attitude in this investigation.

Considering the item difficulty parameters, several items (e.g., Items 9, 11, 12, 14, 15, 19, 21, 23) seem to be functioning in a similar manner, thus serving as repetitive measures. For example, Items 12, 19, and 21 all measure the same person trait level, and among them, these items are that describe questioning skills. In other words, items are likely to function so similarly that little new information could be extracted from individual items. Therefore, this result suggests that such items be reduced in order to cut the item redundancy, a major limitation of the classical test theory. This method of utilizing the item difficulty level could help reduce the number of items in CRCC, with sacrificing less explained variance.

Structural Evidence

The result of dimensionality analysis revealed that the variance explained by the CRCC was 35%, which did not reach the suggested 50.9% (Linacre, 2016) for the measure to be accepted unidimensional. Furthermore, the unexplained variance of the first contrast accounted for slightly over 5% with an eigenvalue of 2.9, indicating that there might be a possible presence of second dimension in the model. The four items showing considerably high correlations with the potential second dimension were items 36, 16 (positively correlated), 11, and 18 (negatively

correlated).

Item 36: My counselor repeated back a concise version of what I said.

Item 16: My counselor mirrored the key content of what I said.

Item 11: My counselor was curious about hearing my story.

Item 18: My counselor actively listened to what I said.

It is possible that there might be additional dimension to differentiate paraphrasing skills and active listening attitude in the CRCC. Even though this explanation could be reasonable, further investigation is still needed for this.

Generalizability

In spite of a small sample size in this study, the 36-item CRCC showed adequate separation index between persons (2.25) and between items (2.24), which were greater than the cutoff value of 2.0 (Linacre, 2016). In addition, person and item separation reliability estimate for the CRCC were .84 and .83, indicating that the CRCC with 36-items was able to appropriately differentiate the persons as well as items on the counselor competence being measured. Unlike the internal consistency reliability in the CTT, the Rasch model's reliability estimate is on ratio scale and provides the measurement error for each separation estimate of person and item (Schumacker & Smith, 2007). This difference enables researchers to more precisely compare the reliability values from the same data, or even different samples. Reliability is a major consideration in psychometrics (Schumacker & Smith, 2007); further

analysis is needed for examining whether or not the adequate quality of reliability in the CRCC can be maintained after removing several bad items identified via the previous analysis.

This study examined the differential item functioning (DIF) across gender for 36 items in the CRCC, in order to detect whether there was any item that functioned differently according gender. The result showed that Items 6, 15, 19, 22, and 33 in the CRCC had statistically significant difference in average measure between male and female. In particular, Item 19 was newly detected as an additional wrong item via DIF investigation, while other items were already mentioned as the items needed to remove or revise. This fact indicates that DIF can evaluate a distinct aspect of psychometric properties in the CRCC. When taking a close look at the result, for Items 6, 19, and 33, the average measure of male was significantly lower than that of female, whereas for Items 15 and 22, male participants' mean was significantly higher than female's. For instance, the item 19, "My counselor asked questions in a clear way" was about the degree to how clearly counselors asked questions in a session. The result indicates that male clients were more reluctant to give good scores on this item 19 than female ones. In other words, male clients were more likely to perceive their counselor's questions as unclear than female clients did. It is possible that this difference might be due to gender difference in brain (Kimura, 1992; Ingalhalikar et al.,2014). Even though this issue about sex difference of brain functioning is still controversial, a counselor's questions without specific facts could be perceived as value questions by male clients more than by female clients. Because of the small sample size, whether this tendency will maintain in bigger sample needs to more investigations.

Interpretability

The person-item map graphically illustrates the relative level of affirmation for items and persons on the Rasch calibrated scale in logits (Linacre, 2016). The person-item map (see Figure 4) in the CRCC shows that the ability levels of most counselors, perceived by their clients were higher than the difficulty levels of the CRCC items, indicating that the current items in the CRCC was unable to appropriately measure the counselor competence of practicum-level counselors-in-training. In other words, it can be interpreted that the range of competence level for most practicum counselors were above the adequate level that the current items in CRCC could measure. The current version of the CRCC focused on measuring beginner-level counselor competence such as basic counseling skills and therapeutic attitude, not including advanced counseling skills (e.g., confrontation, meaning) and other sub-competencies (e.g., multicultural competence, assessment, research, case management).

For lower level counseling skills, Ivey et al. (2013) defined *basic listening sequence* with five basic counseling skills, including attending skills, observation skills, questioning, and reflection skills. Young (2013) also put questioning, clarifying response, paraphrasing, reflecting, and summarizing together under *nonjudgmental listening cycle*. Based on the result that counselor competence level perceived by clients in this study were above the range of the CRCC item difficulty level, most practicum-level counseling students seem to possess the basic competence related to "basic listening sequence" addressed by Ivey et al. (2013) or "nonjudgement listening cycle" by Young (2013) through the training that students received in their counseling program for past 1.5 to 2 years.

Practical Implications

The development of the CRCC and the investigation of psychometric properties in CRCC from the Rasch measurement model suggests several practical implications in measurement and assessment in counselor education. First, the CRCC could be used to assess clients' perspectives of counselor competence. As mentioned in the introduction part, there has been lack of clients' voice in assessing counselor competence (Tate et al., 2014). Adding the perception of clients will result in more comprehensive assessment of counselor competence.

Second, the CRCC suggested some possibility that it can be used as a screening tool for counselor trainees, after more revisions of the current version CRCC and further validation studies. Most counselor training program expects their counselor trainees to build up basic counselor competence enough to perform as a professional in practice during their counseling program. Most practicum-level counselor trainees rated in this study showed the higher competence level than the ability measured by the CRCC; this means that the CRCC could be utilized when counselor educators want to assess whether or not the trainees develop the competence above the expected level, especially from clients' perspective.

Third, this study could encourage the application of the Rasch model to develop more valid or more reliable instruments in counseling field. Although the Rasch model is widely used as an alternative model in other field, there has been a very few research (e.g., Cooke et al., 2015; Kim & Hong, 2004; Ludlow, 2014; Seol, 2007; Zaporozhets et al., 2015) in counseling. This research presented specific procedures concerning how the Rasch model was applied to develop the CRCC and to investigate diverse validity aspects of the developing instrument. This

presentation can support researchers to develop a new linear scale with better validity and reliability, above the CTT's theoretical limitations. In addition, the Rasch model could be widely used to re-evaluate the items in the original instruments developed based on the CTT model, as well as to develop useful, but valid short-form instruments with revising and reducing items in the original instruments widely used in counseling field.

Lastly, this study suggests that negative wording be not used in social science instruments, consistent with previous research (e.g., Liu & Lee, 2015). The result in this study also showed that some negatively described items in the CRCC had problems with several validity investigations. Although the reversely coded items are still widely used in counseling-related instruments, the use of negatively worded items might have some risks not only to be understood differently by respondents, but also to include the variance not related to the variable to be measured.

Limitations

There were several limitations in the present study. As seen from the results, the current items in the CRCC did not sufficiently measure the wide range of ability of counselor competence. This result might be because the current CRCC was developed as a unidimensional scale measuring counselor competence's attributes observable by clients during every session. For this reason, the current CRCC items include only counseling skills and therapeutic attitude. Adding new items to measure the wider level of counselor competence is needed for the CRCC instruments to precisely assess the counselor competence from excellent to poor level. Addition of new items will improve the CRCC with differentiating the competence of counselors.

Moreover, another limitation in this study is related to the current sample. One limitation in the sample was a relatively small sample. A minimum sample size of 150 is commonly recommended for more precise estimations in the Rasch measurement model (Linacre, 1994). However, the participants in this study were 84 adult clients. To obtain more precise estimates, more subjects need to be collected, ideally from more diverse settings. In addition, the subjects measured by the CRCC were rather homogeneous since they all were 2- or 3-year student counselors training in practicum; thus their levels of counselor competence might be within a limited range. Thus, the sample with more diverse range of counselor competence need to be additionally collected so that the CRCC will be able to become more reliable and more valid instrument.

Lastly, this result did not include any external validity evidence for the CRCC. This study addressed only internal evidences of validity within the CRCC, using the Rasch model. External validity of any instrument is considered to be arguably the most important aspect in the traditional measurement (Wolfe & Smith, 2007). For instance, convergent validity and discriminant validity are the commonly used evidence for external validity. In addition, predictive validity to examine the relationship between the test score and the consequence of test score was not investigated in this study. In the context where the CRCC can be used, for example, the relationship between the CRCC and some client outcome tests (e.g., OQ 45.2) needs to be investigated. The correlation value, as a predictive evidence, will demonstrate whether the CRCC can predict the outcome of counseling service by counselors.

Future Research

The present study limitations lead to several anticipated recommendations for future research. First, the limitation regarding the small, homogeneous sample suggests more data collection from counselors with wide range of counselor competence (e.g., 1-year-level, 2-year-level, practicum, internship, interned, licensed) would be needed for a study of this kind. In addition, the sample in this study was recruited from one community counseling center, where students in a CACREP-accredited counseling program received their practicum training. Further research will need to be conducted from different settings, that is, additional validation studies will be needed at other counseling programs in different location, so that the results can be more valid and more generalizable.

Second, another limitation requires future research with the revise items that will re-confirm the results in this study and investigate external validity of the revised CRCC such as convergent validity and predictive validity. It is possible that such a research will administer the revised version of CRCC with reduced items to a new sample and conduct the Rasch analysis with the newly collected data. The research will also examine the correlation between the new CRCC and other instruments related to counselor competence. From the recurring revisions and validations, the CRCC will become a more useful instrument with the valid and reliable psychometric quality enough for research and practice.

Third, there is a need to add items of other advanced counseling skills to the current scale to provide more useful diagnostics for learners with higher levels of counselor competence. As previously addressed, the current items in the CRCC included only lower-level counseling skills

in the hierarchy model of micro-skills proposed by Ivey et al. (2013), and those existing items was not able to appropriately measure the counselor competence that practicum students presented in sessions. When the CRCC is viewed as a ruler, the current ruler of the CRCC has a limited scale, not measuring the higher level of counselor competence. Therefore, further research will involve the generation of new items related to more advanced counseling skills and the Rasch-evaluation on the items.

Lastly, future research is possible to compare different perspectives of counselor competence between supervisors, peer counselors, counselors themselves, and clients. In addition to clients, the CRCC could be rated with same items by diverse raters such as supervisors, peers, and counselors themselves. As such, with the CRCC, diverse perceptions around the same performance of a counselor can be assessed and compared. This comparison research will bring a comprehensive assessment on counselor competence presented by a counselor.

Conclusions

The current research presented how to use the Rasch measurement model for developing the new client-rated measure of counselor competence, the CRCC and examining diverse aspects of psychometric properties of the developed CRCC. The use of Rasch model to assess the psychometric properties of the CRCC scale makes the study results more valid and reliable than using the classical test theory (CTT) because theoretically IRT model overcomes the major weakness of CTT which has circular dependency of item statistics (Fan, 1998). To elaborate, the Rasch analysis provided the validity evidence such as item fit statistics, item difficulty hierarchy,

item-person map, person fit, reliability, and differential item function for the 36-item CRCC; thus, it helped evaluate the developing scale in item level, beyond CTT's group statistics from diverse aspects of validity (Bond & Fox, 2001; Engelhard, 2013; Wolfe & Smith, 2007).

The investigations of the CRCC in this study was able to detect several wrong items: misfitting items to the model, items functioning differently across gender, and items with wrong item calibration. The rating scale function used in the CRCC also was evaluated. Those results suggested how to improve the current items and rating scale functioning in the CRCC, in order to produce a valid, linear measure. In addition, theoretically ordered clusters underlying counseling skills are mostly consistent to the result of the item calibration in this study, except for the "Theoretical Attitude" cluster. This can be the evidence to support that the latent variable "basic listening sequences" consists of 'Reflection of Feeling', 'Reflection of Contents', 'Questioning', and 'Attending Behaviors' in a hierarchical way that Ivey et al. (2013) conceptualized. Like this result, the use of the Rasch analysis can be a useful tool to empirically demonstrate whether a theoretical concept or model, especially with hierarchical or developmental structure exist with real data. Moreover, the study presented that the item-person map in the Rasch model can provide useful information regarding evaluating the instruments and interpreting the test scores.

In summary, this study addressed the use of the Rasch model through developing and validating procedures of the newly developed CRCC measure. The researcher hopes this study could contribute to more application of Rasch model in counseling field, in order to produce more valid, reliable instruments.

APPENDIX A: UNIVERSITY OF CENTRAL FLORIDA INSTITUTIONAL REVIEW FORM

University of Central Florida Institutional Review Board
Office of Research & Commercialization
12201 Research Parkway, Suite 501
Orlando, Florida 32826-3246
Telephone: 407-823-2901 or 407-882-2276
www.research.ucf.edu/compliance/irb.html

Approval of Exempt Human Research

From: **UCF Institutional Review Board #1**
FWA00000351, IRB00001138

To: **Hang Jo**

Date: **December 03, 2015**

Dear Researcher:

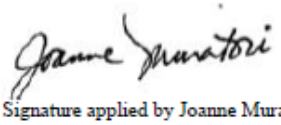
On 12/03/2015, the IRB approved the following activity as human participant research that is exempt from regulation:

Type of Review:	Exempt Determination
Project Title:	A Pilot Study for Developing the Client Ratings of Counselor Competence: A Client-Rated Measure of Counselor Competence
Investigator:	Hang Jo
IRB Number:	SBE-15-11770
Funding Agency:	
Grant Title:	
Research ID:	N/A

This determination applies only to the activities described in the IRB submission and does not apply should any changes be made. If changes are made and there are questions about whether these changes affect the exempt status of the human research, please contact the IRB. When you have completed your research, please submit a Study Closure request in iRIS so that IRB records will be accurate.

In the conduct of this research, you are responsible to follow the requirements of the Investigator Manual.

On behalf of Sophia Dziegielewski, Ph.D., L.C.S.W., UCF IRB Chair, this letter is signed by:



Signature applied by Joanne Muratori on 12/03/2015 09:27:24 AM EST

IRB Manager

APPENDIX B: UNIVERISTY OF CENTRAL FLORIDA INSTITUTIONAL REVIEW FORM ADDENDUM

University of Central Florida Institutional Review Board
Office of Research & Commercialization
12201 Research Parkway, Suite 501
Orlando, Florida 32826-3246
Telephone: 407-823-2901 or 407-882-2276
www.research.ucf.edu/compliance/irb.html

Approval of Exempt Human Research

From: **UCF Institutional Review Board #1**
FWA00000351, IRB00001138

To: **Hang Jo**

Date: **May 20, 2016**

Dear Researcher:

On 05/20/2016, the IRB approved the following minor modifications to human participant research that is exempt from regulation:

Type of Review:	Exempt Determination
Modification Type:	A revised study instrument has been uploaded in iRIS. In addition the total number of study participants is being increased from 60 to 150 and data collection will extend through July 2016. A revised protocol has been uploaded in iRIS and a revised consent document has been approved for use.
Project Title:	A Pilot Study for Developing the Client Ratings of Counselor Competence: A Client-Rated Measure of Counselor Competence
Investigator:	Hang Jo
IRB Number:	SBE-15-11770
Funding Agency:	
Grant Title:	
Research ID:	N/A

This determination applies only to the activities described in the IRB submission and does not apply should any changes be made. If changes are made and there are questions about whether these changes affect the exempt status of the human research, please contact the IRB. When you have completed your research, please submit a Study Closure request in iRIS so that IRB records will be accurate.

In the conduct of this research, you are responsible to follow the requirements of the Investigator Manual.

On behalf of Sophia Dziegielewski, Ph.D., L.C.S.W., UCF IRB Chair, this letter is signed by:

Signature applied by Joanne Muratori on 05/20/2016 04:09:33 PM EDT

IRB Manager

APPENDIX C: EXPLANATION OF RESEARCH

Summary Explanation for Exempt Research

EXPLANATION OF RESEARCH

Title of Project: Development of the Client Ratings of Counselor Competence: A Client-Rated Measure of Counselor Competence

Principal Investigator: Hang Jo, M.A.

Other Investigators: N/A

Faculty Supervisor: K. Dayle Jones, Ph.D.

You are being invited to take part in a research study. Whether you take part is up to you.

- The purpose of this research is to develop a new client-rated scale of counselor competence, named Client Ratings of Counselor Competence.
- You will be asked to fill out a written-paper instrument to rate your counselor's responses in session. The administration of the survey will be conducted only once after a counseling session.
- The survey will take from 10 to 15 minutes to complete. You do not have to answer every question. After completing the survey, you will be asked to turn in the completed form to an assigned place in the center. Your answer sheet will not be viewed by your counselor.

You must be 18 years of age or older to take part in this research study.

Study contact for questions about the study or to report a problem: If you have questions, concerns, or complaints, talk to: Hang Jo, Doctoral Candidate, Counselor Education Program, College of Education and Human Performance, (407) 222-6105 or by email at hang.jo@knights.ucf.edu or Dr. K. Dayle Jones, Faculty Supervisor, College of Education and Human Performance at (407) 823-2401 or by email at daylejones@ucf.edu.

IRB contact about your rights in the study or to report a complaint: Research at the University of Central Florida involving human participants is carried out under the oversight of the Institutional Review Board (UCF IRB). This research has been reviewed and approved by the IRB. For information about the rights of people who take part in research, please contact: Institutional Review Board, University of Central Florida, Office of Research & Commercialization, 12201 Research Parkway, Suite 501, Orlando, FL 32826-3246 or by telephone at (407) 823-2901.

APPENDIX D: CLIENT RATINGS OF COUNSELOR COMPETENCE (CRCC) FINAL FORM

CRCC Scale

Instructions: Please answer each of the following statements based upon your experience in THIS WEEK's counseling session.

During this session, my counselor…	Strongly Disagree	Disagree	Agree	Strongly Agree
1. maintained good eye contact with me.	☐1	☐2	☐3	☐4
2. did *not* detect my deeper feelings.	☐1	☐2	☐3	☐4
3. was open to talking about any feeling that I expressed.	☐1	☐2	☐3	☐4
4. made me feel interrogated by his/her questions.	☐1	☐2	☐3	☐4
5. maintained a gentle tone of voice.	☐1	☐2	☐3	☐4
6. was honest and frank.	☐1	☐2	☐3	☐4
7. responded to me warmly.	☐1	☐2	☐3	☐4
8. organized my thinking about what happened in the session.	☐1	☐2	☐3	☐4
9. asked me to give more details about my topic.	☐1	☐2	☐3	☐4
10. normalized the feelings I was having.	☐1	☐2	☐3	☐4
11. was curious about hearing my story.	☐1	☐2	☐3	☐4
12. gave me enough time to think after questioning.	☐1	☐2	☐3	☐4
13. asked too many questions at the same time.	☐1	☐2	☐3	☐4
14. summarized what I said so that I could understand my situation more clearly.	☐1	☐2	☐3	☐4
15. used *inappropriate* head nodding.	☐1	☐2	☐3	☐4
16. mirrored the key content of what I said.	☐1	☐2	☐3	☐4
17. showed open and welcoming gestures.	☐1	☐2	☐3	☐4
18. actively listened to what I said.	☐1	☐2	☐3	☐4
19. asked me questions in a clear way.	☐1	☐2	☐3	☐4
20. seemed to be genuine with me.	☐1	☐2	☐3	☐4
21. explored important issues with me.	☐1	☐2	☐3	☐4
22. imposed his/her values on me.	☐1	☐2	☐3	☐4
23. summarized the main points of what we discussed.	☐1	☐2	☐3	☐4
24. used more open questions than "yes or no" questions.	☐1	☐2	☐3	☐4
25. helped me identify my underlying feelings.	☐1	☐2	☐3	☐4
26. used a variety of feeling words to describe my emotions.	☐1	☐2	☐3	☐4
27. asked questions that helped me explore what I was thinking or feeling.	☐1	☐2	☐3	☐4
28. helped me label my feelings.	☐1	☐2	☐3	☐4
29. provided a comfortable physical distance between us.	☐1	☐2	☐3	☐4
30. understood exactly what I meant.	☐1	☐2	☐3	☐4
31. seemed to think what I said was important.	☐1	☐2	☐3	☐4
32. validated my feelings.	☐1	☐2	☐3	☐4

Please Continue on the Back

Instructions: Please answer each of the following statements based upon your experience in THIS WEEK's counseling session. **During this session, my counselor…**	Strongly Disagree	Disagree	Agree	Strongly agree
33. fully understood my unique situation and values.	☐1	☐2	☐3	☐4
34. precisely identified my feelings.	☐1	☐2	☐3	☐4
35. accurately rephrased what I said in his/her own words.	☐1	☐2	☐3	☐4
36. repeated back a concise version of what I said.	☐1	☐2	☐3	☐4

37. What is your age?

_____

38. What is your gender?

☐ Male

☐ Female

☐ Transgender

☐ Other: _____

39. Race / ethnicity / cultural background:

☐ Asian ☐ Black/African American ☐ Caucasian ☐ Native American

☐ Hispanic/Latino ☐ Multiracial/multicultural ☐ Other: _____

40. How many sessions have you had with this counselor?

_____ sessions

**** Thank you for your time in completing this questionnaire. ****

APPENDIX E: CRCC EXPERT REVIEW FORM

Expert Review Form of CRCC

Subscale	Items During this session, my counselor…	Clarity (1 to 5)	Importance (1 to 5)	Relevance (1 to 5)
Nonverbal Skills	1. provided a comfortable physical distance between us.			
	2. maintained good eye contact with me.			
	3. showed open and welcoming gestures.			
	4. maintained a gentle tone of voice.			
	5. used *inappropriate* head nodding. (Reverse)			
Therapeutic Attitude	6. actively listened to what I said.			
	7. understood exactly what I meant.			
	8. fully understood my unique situation and values.			
	9. responded to me warmly.			
	10. imposed his/her values on me. (Reverse)			
	11. seemed to think what I said was important.			
	12. was curious about hearing my story.			
	13. was honest and frank.			
	14. seemed to be genuine with me.			
Questioning	15. asked me questions in a clear way.			
	16. gave me enough time to think after questioning.			
	17. asked questions that helped me explore what I was thinking or feeling.			
	18. asked multiple questions at the same time. (Reverse)			
	19. used more open questions than “yes or no” questions.			
	20. made me feel interrogated by his/her questions. (Reverse)			
	21. asked me to give more details about my topic.			
	22. explored important issues with me.			

Subscale	Items During this session, my counselor…	Clarity (1 to 5)	Importance (1 to 5)	Relevance (1 to 5)
Reflection of Contents	23. repeated back a concise version of what I said.			
	24. accurately rephrased what I said in his/her own words.			
	25. mirrored the key content of what I said.			
	26. summarized what I said so that I could understand my situation more clearly.			
	27. summarized the main points of what we discussed.			
	28. organized my thinking about what happened in the session.			
Reflection of Feelings	29. was open to talking about any feeling that I expressed.			
	30. precisely identified my feelings.			
	31. did *not* detect my deeper feelings. (Reverse)			
	32. helped me identify my underlying feelings.			
	33. used a variety of feeling words to describe my emotions.			
	34. helped me label my feelings.			
	35. normalized the feelings I was having.			
	36. validated my feelings.			

LIST OF REFERENCES

Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied psychological measurement, 21*(1), 1-23.

Andrich, D. (1978). Scaling attitude items constructed and scored in the Likert tradition. *Educational and Psychological Measurement, 38*(3), 665-680.

American Counseling Association. (2014). *Code of ethics*. Alexandria, VA: Author.

Barndorff-Nielsen, O. (1978). Hyperbolic distributions and distributions on hyperbolae. *Scandinavian Journal of statistics*, 151-157.

Barrett-Lennard, G. T. (1986). The Relationship Inventory now: Issues and advances in theory, method, and use.

Bernard, J. M., & Goodyear, R. K. (2014). *Fundamentals of clinical supervision*. (5th ed.). Upper Saddle River, NJ: Pearson Education.

Blount, A. J. (2015). *The Helping Professional Wellness Discrepancy Scale (HPWDS): Development and validation*. (Unpublished doctoral dissertation). University of Central Florida, Orlando, FL.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: fundamental measurement in the human sciences*, Mahwah, NJ: Erlbaum.

Cattel, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*, 245-276.

Christensen, K. B., Engelhard, G., & Salzberger Jr, T. (2012). Ask the experts: Rasch vs. factor

analysis. *Rasch Measurement Transactions, 26*(3), 1373-1386.

Cole, E. M., Piercy, F., Wolfe, E. W., & West, J. M. (2014). Development of the Multicultural Therapy Competency Inventory-Client Version. *Contemporary Family Therapy, 36*(4), 462–473. doi:10.1007/s10591-014-9320-8

Cooke, D., Marais, I., Cavanagh, R., Kendall, G., & Priddis, L. (2015). Differences between mothers' and fathers' ratings of family functioning with the family assessment device the validity of combined parent scores. *Measurement and Evaluation in Counseling and Development*, 48(3), 226-237.

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, & Evaluation, 10*(7). Retrieved from http:// pareonline.net/pdf/v10n7.pdf

Council for Accreditation of Counseling and Related Educational Programs. (2016). *2016 standards*. Retrieved from http://www.cacrep.org/2016standards.html

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando, FL: Harcourt Brace Jovanovich, Inc.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334.

DeVellis, R. F. (2012). *Scale development: Theory and applications*. Thousand Oaks, CA: Sage.

Dill-Standiford, T. J., Stiles, W. B., & Rorer, L. G. (1988). Counselor-client agreement on session impact. *Journal of Counseling Psychology, 35*(1), 47.

Egan, G. (2013). *The skilled helper: A problem-management and opportunity-development approach to helping*. Cengage Learning.

Engelhard Jr, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. Routledge.

Eriksen, K., & McAuliffe, G. (2003). A measure of counselor competency. *Counselor Education and Supervision, 43*(December), 120–133. Retrieved from http://ww2.odu.edu/~gmcaulif/documents/Articles/MeasureCNSComp.pdf

Fairburn, C. G., & Cooper, Z. (2011). Therapist competence, therapy quality, and therapist training. *Behaviour research and therapy, 49*(6-7), 373–8. doi:10.1016/j.brat.2011.03.005

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and psychological measurement, 58*(3), 357-381.

Fox, C. M., & Jones, J. A. (1998). Uses of Rasch modeling in counseling psychology research. *Journal of Counseling Psychology, 45*(1), 30.

Gall, M. D., Gall, J. P., & Borg, W. R. (2007). *Educational research: An introduction*. (8th ed.). Boston, MA: Pearson Education, Inc.

Gazda, G. M. (1999). *Human relations development: A manual for educators*. Boston: Allyn & Bacon.

Hatcher, R. L., & Gillaspy, J. A. (2006). Development and validation of a revised short version of the working alliance inventory. *Psychotherapy Research, 16*(1), 12–25. doi:10.1080/10503300500352500

Hill, C. E. & Kellems, I. S. (2002). Development and use of the Helping Skills Measure to assess client perceptions of the effects of training and of helping skills in sessions. *Journal of Counseling Psychology, 49*(2), 264-272.

Hughes, G. (2014). *Competence and self-care in counselling and psychotherapy*. New York, NY: Routledge.

Ingalhalikar, M., Smith, A., Parker, D., Satterthwaite, T. D., Elliott, M. A., Ruparel, K., ... & Verma, R. (2014). Sex differences in the structural connectome of the human brain. *Proceedings of the National Academy of Sciences, 111*(2), 823-828.

Ivey, A. E. (1971). *Microcounseling: Innovations in interviewing training*. Springfield, IL: Thomas.

Ivey, A. E., Ivey, M. B., & Zalaquett, C. (2013). *Intentional interviewing and counseling: Facilitating client development in a multicultural society*. Cengage Learning

Johnson, E., Baker, S. B., Kopala, M., Kiselica, M. S., & Thompson, E. C. (1989). Counseling self-efficacy and counseling competence in prepracticum training. *Counselor Education and Supervision, 28*, 205-218.

Kimura, D. (1992). Sex differences in the brain. *Scientific American, 267*(3), 118-125.

Kagan, N., Krathwohl, D. R., & Miller, R. (1963). Stimulated recall in therapy using video tape: A case study. *Journal of Counseling Psychology, 10*(3), 237.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*, 141-151.

Kaplan, D. M., Tarvydas, V. M., & Gladding, S. T. (2014). 20/20: A Vision for the Future of Counseling: The New Consensus Definition of Counseling. *Journal of Counseling & Development, 92*(3), 366–372.

Kim, B. S. K., & Hong, S. (2004). A psychometric revision of the Asian values scale using the Rasch model. *Measurement and Evaluation in Counseling and Development, 37*(1), 15-37.

Kohrt, B. A, Jordans, M. J. D., Rai, S., Shrestha, P., Luitel, N. P., Ramaiya, M. K., & Patel, V. (2015). Therapist competence in global mental health: Development of the ENhancing Assessment of Common Therapeutic factors (ENACT) rating scale. *Behaviour research and therapy, 69*, 11–21.

LaFromboise, T., Coleman, H., & Hernandez, A. (1991). Development and factor structure of the Cross-Cultural Counseling Inventory–Revised. *Professional Psychology: Research and Practice, 22*, 380–388.

Lamb, R. L., Annetta, L., Meldrum, J., & Vallett, D. (2012). Measuring science interest: Rasch validation of the science interest survey. *International Journal of Science and Mathematics Education, 10*(3), 643-668.

Lamb, R. L., Vallett, D., & Annetta, L. (2014). Development of a short-form measure of science and technology self-efficacy using Rasch analysis. *Journal of Science Education and Technology, 23*(5), 641-657.

Lambert, M. J., Hansen, N. B., Umphress, V. J., Lunnen, K., Okiishi, J., Burlingame, G. M., et al. (1996). *Administration and scoring manual for the OQ 45.2*. Stevenson, MD: American Professional Credentialing Services.

Lambert, M. J., & Shimokawa, K. (2011). Collecting client feedback. *Psychotherapy, 48*(1), 72.

Larson, L. M., Suzuki, L. A., Gillespie, K. N., Potenza, M. T., Bechtel, Bechtel, M. A., & Toulouse, A. L. (1992). Development and validation of the counseling self-estimate inventory. *Journal of Counseling Psychology, 39*(1), 105-120.

Leigh, I. W., Smith, I. L., Bebeau, M. J, Lichtenberg, J. W., Nelson, P. D., Portnoy, S., Rubin, N. J., & Kaslow, N. J. (2007). Competency assessment models. *Professional Psychology: Research and Practice, 38*, 463–473.

Lent, R. W., Hill, C. E., & Hoffman, M. A. (2003). Development and validation of the Counselor Activity Self-Efficacy Scales. *Journal of Counseling Psychology*, *50*(1), 97–108. doi:10.1037/0022-0167.50.1.97

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

Linacre, J. (2002a). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, *3*(1), 85-106.

Linacre, J. (2002b). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*, 878.

Linacre, J. (2016). WINSTEPS (Software and user's guide). Chicago, IL: Winsteps.

Liu, X. (2010). *Using and developing measurement instruments in science education: A Rasch*

modeling approach. Charlotte: Information Age.

Liu, H. H., & Lee, Y. (2015). Measuring self-regulation in second language learning: A Rasch analysis. *Sage Open*, 1-12.

Ludlow, L. H., Matz-Costa, C., Johnson, C., Brown, M., Besen, E., & James, J. B. (2014). Measuring Engagement in Later Life Activities Rasch-Based Scenario Scales for Work, Caregiving, Informal Helping, and Volunteering. *Measurement and Evaluation in Counseling and Development, 47*(2), 127-149.

McAuliffe, G. & Eriksen, K. (Ed.). (2011). *Handbook of counselor preparation: Constructivist, developmental, and experiential approaches*. Thousand Oaks, CA: Sage.

McLeod, J. (1992). What do we know about how best to assess counsellor competence? *Counselling Psychology Quarterly, 5*(4), 359-372.

McLeod, J. (1996). Counsellor competence in Bayne, R., Horton, I. and Bimrose, J. (eds), *New Directions in Counselling*. London: Sage.

Miller, S. L., Duncan, B. L., Sorrell, R., & Brown, G. S. (2005). The partners for change outcome management system. *Journal of Clinical Psychology, 61*, 199–208.

Mullen, P. R., Lambie, G. W., & Conley, a. H. (2014). Development of the Ethical and Legal Issues in Counseling Self-Efficacy Scale. *Measurement and Evaluation in Counseling and Development, 47*(1), 62–78. http://doi.org/10.1177/0748175613513807

Parsons, R. D., & Zhang, N. (2014). *Becoming a skilled counselor*. Thousand Oaks, CA: Sage.

Ponterotto, J. G., Rieger, B. P., Barrett, A., & Sparks, R. (1994). Assessing multicultural competence: A review of instrumentation. *Journal of Counseling & Development, 72*, 316–322.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Paedagogiske Institute.

Reese, R. J., Usher, E. L., Bowman, D. C, Norsworthy, L.A., Halstead, J. L., Rowlands, S. R., & Chisholm, R. R. (2009). Using client feedback in psychotherapy training: An analysis of its influence on supervision and counselor self-efficacy. *Training and Education in Professional Psychology, 3*(3), 157-168.

Rogers, C. R. (1967). *Person to person: The problem of being human*. Moab, UT: Real People Press.

Sammet, K. (2012). *A Rasch Item Response Modeling Approach to Validation: Evidence Based on Test Content and Internal Structure of the Life Effectiveness Questionnaire*. (Doctoral dissertation, Emory University)

Sampson, J. P., & Peterson, G. W. (1996). *Career thoughts inventory*. Psychological Assessment Resources.

Schumacker, R. E., & Smith, E. V. (2007). A Rasch perspective. *Educational and Psychological Measurement, 67*(3), 394-409.

Seol, H. (2007). A Psychometric Investigation of the Marlowe-Crowne Social Desirability Scale using Rasch measurement. *Measurement and Evaluation in Counseling and Development, 40*, 127-149.

Smith, E. V., Conrad, K. M., Chang, K., & Piazza, J. (2002). An introduction to Rasch measurement for scale development and person assessment. *Journal of Nursing Measurement, 10*(3), 189-206.

Soska III, P. J. (2012). *Use of Rasch Rating Scale Modeling to Develop and Validate a Measure of District-Level Characteristics and Practices Identified to Improve Instruction and Increase Student Achievement* (Doctoral dissertation, Bowling Green State University).

Swank, J. M. (2010). *Assessing the psychometric properties of the Counseling Competencies Scales©: A measure of counseling skills, dispositions, and behaviors*. (Unpublished doctoral dissertation). University of Central Florida, Orlando, FL.

Swank, J. M. (2014). Assessing counseling competencies: A comparison of supervisors' ratings and student supervisees' self-ratings. *Counseling Outcome Research and Evaluation, 5*(1), 17–27. doi:10.1177/2150137814529147

Swank, J. M., & Lambie, G. W. (2012). The assessment of CACREP core curricular areas and student learning outcomes using the Counseling Competencies Scale. *Counseling Outcome Research and Evaluation, 3*(2), 116–127. doi:10.1177/2150137812452560

Swank, J. M., Lambie, G. W., & Witta, E. L. (2012). An exploratory investigation of the Counseling Competencies Scale: A measure of counseling skills, dispositions, and behaviors. *Counselor Education and Supervision, 51*(3), 189-206.

Sue, D. W., Arredondo, P., & McDavis, R. J. (1992). Multicultural counseling competencies and standards: A call to the profession. *Journal of Counseling & Development, 70*(4), 477-486.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics, 5th*. Needham Height, MA: Allyn & Bacon.

Tate, K. A., Bloom, M. L., Tassara, M. H., & Caperton, W. (2014). Counselor competence, performance assessment, and program evaluation: Using psychometric instruments. *Measurement and Evaluation in Counseling and Development*, *47*(4), 291–306. doi:10.1177/0748175614538063

Tate, K., Torres Rivera, E., Conwill, W., Miller, M., & Puig, A. (2013). Conceptualizing group dynamics from our clients' perspective: Development of the Conceptualization of Group Dynamics Inventory. *Journal for Specialists in Group Work, 38*, 146–168.

Thompson, B., & Hill, C. (1993). Client Perceptions of Therapist Competence. Psychotherapy *Research*, *3*(2), 124–130. doi:10.1080/10503309312331333729

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

Urofsky, R. I., & Bobby, C. L. (2012). The Evolution of a Student Learning Outcomes Focus in the CACREP Standards in Relation to Accountability in Higher Education. *Counseling Outcome Research and Evaluation, 3*(2), 63–72. doi:10.1177/2150137812452562

Wheeler, S. (2003). *Training counsellors: the assessment of competence*. Thousand Oaks, CA: Sage.

Wolfe, E., & Smith, E. (2007). Instrument development tools and activities for measure validation using Rasch models: Part II – Validation activities. *Journal of Applied Measurement, 8*(2), 204-234.

Wright, B., & Masters, G. (1982). *Rating scale analysis*. Chicago, IL: MESA Press.

Young, M. E. (2013). *Learning the art of helping: Building blocks and techniques*. Upper Saddle River, NJ: Pearson.

Zaporozhets, O., Fox, C. M., Beltyukova, S. A., Laux, J. M., Piazza, N. J., & Salyers, K. (2015). Refining change measure with the Rasch model. *Measurement and Evaluation in Counseling and Development, 48*(1), 59-74.