Spring 2012

Longitudinal invariance of the Scale of Ethnocultural Empathy

Jerusha Joy Gerstner
James Madison University

Longitudinal Invariance of the Scale of Ethnocultural Empathy

Jerusha J. Gerstner

A thesis submitted to the Graduate Faculty of

JAMES MADISON UNIVERSITY

In

Partial Fulfillment of the Requirements

for the degree of

Master of Arts

Department of Graduate Psychology

May 2012

Dedication

My thesis is dedicated to my parents, Jonathan and Kathy Gerstner, who have always challenged and encouraged me to perform my best in my academic career. None of this work would have been possible without their never-ending love and support.

Acknowledgments

I would like to begin by acknowledging my advisor, Dr. Dena Pastor, for her assistance in the entire process of completing this thesis. Words cannot fully express my gratitude for Dena's guidance, encouragement, and knowledge of methodology, which provided the support needed to complete this manuscript. Dena was always available to answer any questions I had and helped me think through some of the more challenging concepts along the way. I would not have been able to complete this work without her help as my advisor.

I would also like to acknowledge the members of my thesis committee, Dr. Sara Finney and Dr. Matthew Lee. The time and effort both of them spent reading my thesis and providing invaluable comments has helped to strengthen this document. Sara was a great support in ensuring the methodology utilized was appropriate and helping me to think more deeply about these methodological concepts. Matt provided assistance in understanding and reading appropriate literature on cultural matters; his expertise in this area strengthened the more theoretical sections of my thesis.

I would like to acknowledge all of my family for their support over the years. The support and encouragement received from each and every one of them helped me to complete this thesis. Finally, I must acknowledge my friends both old and (relatively) new for their support in this process. Thanks to all my friends from my undergraduate studies for supporting me in this endeavor, even when you did not completely understand the focus of my studies. I would also like to acknowledge my fellow graduate students for their constant support and help with taking much needed breaks. Being surrounded by such dedicated students (and friends) has pushed me to be a better student myself.

Table of Contents

List of Tables

List of Figures

Abstract

The Scale of Ethnocultural Empathy (SEE; Wang et al., 2003) was developed to measure ethnocultural empathy and is a promising tool for assessing change in the construct over time. However, no prior study examines the functioning of the SEE over time. The purpose of the study was to examine the invariance of the SEE utilizing a modified factor model that included a negative wording effect. The SEE was found to exhibit longitudinal measurement invariance over a two-year time span in a sample of undergraduate students. Change over time was examined at the error-free, latent level and only one subscale (EPT) showed a significant increase over time. Test-retest coefficients showed most factors were relatively stable over time (i.e., individuals were changing in the same direction). The results point to the need for further examination of the SEE and of the negative wording factor on the scale, specifically.

I. Introduction

The world is becoming more diverse every day. The US is not a country filled with homogeneous groups of people. On the contrary, it is truly a melting pot of cultures. Because of the increased mixing of cultures, exhibiting a level of empathy toward those from other cultures is imperative for functioning. Wang et al. (2003) reported that minorities are anticipated to comprise one-third of the population of the United States by 2015. Based on U.S. census data, in 2010, racial minorities already comprised 26.67% of the population (U.S. Census Bureau, 2010). Given the increase in diversity anticipated in our country, it is of ever increasing importance that the population is able to exhibit levels of empathy and interpersonal skills toward those of cultures other than their own.

Empathy

Empathy, which is defined as knowing and feeling as another person does, has been a construct studied vastly over the years (Davis, 1983). Research has addressed its role in educational (e.g., E. L. Brown, 2004), medical (e.g., Jolliffe & Farrington, 2006), and counseling (e.g., Chung & Bemak, 2002) settings. Furthermore, numerous scales have been developed to measure general empathy (e.g., Davis, 1983; Mehrabian & Epstein, 1972). Although having the ability to exemplify a form of general empathy is important, of arguably more relevance given the increase in cultural diversity in the country and world is the distinct but related construct of ethnocultural empathy. Ethnocultural empathy is directed toward those from a different ethnic group than one's own group (Wang et al., 2003).

Whereas the general concept of empathy has been studied and related to numerous other variables (e.g., social dominance orientation, altruism), ethnocultural empathy represents a newer area of research. Wang et al. (2003) developed the Scale of Ethnocultural

Empathy (SEE) due to a lack of measures of this construct. It is necessary to have a measure of ethnocultural empathy so its relationship with other variables can be assessed.

The Scale of Ethnocultural Empathy

Wang et al. (2003) developed the SEE in the counseling realm based on a theoretical conceptualization of cultural empathy (Ridley & Lingle, 1996) with the intention to generalize to areas outside of counseling. Cultural empathy is the ability to exemplify empathy toward those from different cultures (e.g., gender, socioeconomic status; Ridley & Lingle, 1996), whereas Wang et al. believed ethnocultural empathy was a more specific form of cultural empathy specific to empathy shown towards those from different ethnocultural backgrounds. Items were written to map onto three domains consistent with Ridley and Lingle's theory: (a) the intellectual domain, (b) the empathic emotions domain, and (c) the communicative domain (Wang et al., 2003). The intellectual domain pertains to the ability to take on the point of view of another. The empathic emotions domain includes feeling emotion as a person from a different cultural background does and the outpouring of empathic emotions towards a person of a different background. The communicative domain includes both probing for insight, which consists of asking necessary questions to enhance understanding of another person's perspective and conveying accurate understanding, which consists of effectively expressing understanding of others. Although the SEE was created to capture these three domains, when the scale was examined by its creators using Principal Components Analysis (PCA), a four-component solution was retained, and four subscales were subsequently developed for the SEE (Wang et al., 2003). The extent to which these subscales relate back to the theoretical conceptualization is unclear, as the names and apparent content covered did not parallel the three originally hypothesized domains. Wang

et al. also conducted more stringent tests of the fit of the models using confirmatory factor analyses with item parcels.

Structural Validity Evidence. Despite having promising qualities, there has been a lack of more stringent research conducted thus far on the factor structure of the SEE. The primary structural validity evidence for the scale was conducted by the scale developers. In order to ensure the validity of the inferences drawn corresponding to ethnocultural empathy, further inspection of the scale's factor structure is warranted. Wang et al. (2003) conducted both exploratory (e.g., PCA) and confirmatory (e.g., CFA) analyses of the SEE to provide some evidence of structural validity. However, they unfortunately used inappropriate methodology to obtain this evidence. PCA is not the appropriate technique to use when attempting to examine a latent construct driving responses to items, as is the case in this scenario. PCA partitions the total variance rather than examining the common variance, which is appropriate when believing a latent construct drives responses to items. Furthermore, the use of item parcels in the CFA prohibits inspection of the functioning of individual items by bundling the items together (Bandalos & Finney, 2001).

Change Over Time. There is a need for the measurement of change in ethnoculutural empathy over time in a variety of substantive domains. For instance, it would be valuable to assess any change in ethnocultural empathy over the course of a physician training program (e.g., Bellini & Shea, 2005), an educational program (e.g., Long, Angera, Carter, Nakamoto, & Kalso, 1999), and over the course of various life developmental stages (e.g., Davis & Franzoi, 1991). Although change in general empathy has been studied in these various contexts, ethnocultural empathy would have likely been more relevant given the added complexity inherent in cross-cultural interactions. In order to utilize the SEE to measure change over time, it is first essential to determine whether the measure is invariant

across time. That is, it is necessary to establish that the measurement properties of the SEE are stable across time by establishing that neither beta change (i.e., the items relate to the factors in the same manner across time) nor gamma change (i.e., students conceptualized the construct in the same manner across time) are occurring. If the measure is noninvariant, it indicates measured change will not be indicative of true change in the construct.

As a result of attending university, it would be anticipated that students would increase in their levels of ethnocultural empathy if exposed to people from other ethnic groups and/or if they learned about people from other cultures in their classes, residence halls, and work settings. That is, many institutions have in place targeted courses or experiences to increase awareness of those of different cultures. However, as of yet, there has been no research examining change in college students' levels of ethnocultural empathy during their first two years of college. Ethnocultural empathy can be assessed more accurately in college students and other populations by first establishing measurement invariance for the SEE and then assessing change over time at the error-free latent level.

Purpose of the Current Study

Given the lack of research thus far on the SEE, it is imperative that more studies are conducted to provide more evidence for the structure of the scale and its stability. As most of the research thus far has been conducted by the developers of the scale, it is important to have an objective analysis conducted by those independent from the development of the scale. Furthermore, it is imperative that appropriate methodology be used when evaluating the factor structure of the SEE. Finally, in order to examine change over time in the construct, measurement invariance should be established for the scale. The following research questions will be addressed in this study.

Research Question #1. Is the SEE an invariant measure of ethnocultural empathy across time? That is, does the SEE measure ethnocultural empathy in the same way across time? The same factor structure was tested at two time points (i.e., configural invariance was tested). If configural invariance is established, then the unstandardized pattern coefficients can be constrained to be equal across time (i.e., metric invariance was tested to assess whether the items related to the factors in the same manner across time). Finally, the intercepts can also be constrained to be equal across time (i.e., scalar invariance was tested to assess whether students with the same level of ethnocultural empathy at both time points would score the same across time on the measure). If measurement invariance of the SEE is established, secondary research questions can be explored.

Research Question #2. Do college students change in ethnocultural empathy over the course of the first two years of their undergraduate career? To examine change over time in ethnocultural empathy, latent mean differences on each SEE subscale and latent Glass' Δ effect size were obtained, in addition to test-retest coefficients. Given the programming in place at universities targeted at increasing ethnocultural empathy, we anticipated an increase in students' levels of ethnocultural empathy over their first two years of college. We also anticipated a moderate test-retest coefficient, which would indicate that students are changing in the same manner over time. We anticipate students would change in the same manner over time, because we expect students to be impacted similarly by the targeted courses.

II. Review of the Literature

Empathy, or the ability to know and feel as another person does, has been a topic discussed and researched for many years (Davis, 1983). It is an area relevant to many different fields. For instance, employees in many careers (e.g., counseling, nursing, education) need to be empathetic towards others for their job. It is necessary to put yourself in the place of another, to understand another's perspective on that deeper, empathetic level to be effective in such careers. In counseling, for example, a counselor will not be able to effectively meet the needs of a client if they cannot understand their client's perspective (Ivey, Ivey, & Simek-Morgan, 1987; Rogers, 1975). Furthermore, in order to interact effectively in society, it is essential to be empathetic with those encountered in everyday life, as well as the workplace.

Given the relevance and necessity of exemplifying empathy, it is not surprising that there is a wide variety of literature focusing on the conceptualization and measurement of this construct. The specific topic of interest in the current study is a precise type of empathy, ethnocultural empathy. Given the recent development of the construct (only developed in 2003 by Wang et al.), general empathy will be discussed first to provide a foundational concept of the construct followed by the more specific research pertaining to ethnocultural empathy in the more recent years.

Affective Empathy

When empathy was first theorized as a topic of interest to study, it was discussed primarily as a human emotion, or feeling. *Einfühlung* was the first known word coined to describe empathy (Vischer, 1873, as cited by Duan & Hill, 1996). By Vischer's (1873) definition, Einfühlung was the subconscious projection of feelings into what is being perceived. Titchener (1924) was first to term the English word, empathy, from Vischer's

theory. Titchener defined empathy as the "process of humanizing objects, of reading or feeling ourselves into them" (Titchener, 1924, p. 417). Although the theoretical conceptualization of empathy has changed over the years, this affective component is still a crucial element of empathy theory.

Cognitive Empathy

The second component of empathy, a cognitive component, was added to the theoretical conceptualization of empathy by Mead (1934). Mead argued that the construct of empathy was incomplete without a cognitive component, which was later defined as "intellectually taking the role or perspective of another person" (Gladstein, 1983, p. 468). In other words, Mead believed it was necessary to know a person (i.e., cognitive component), as well as feel for them (i.e., affective component), to experience empathy. With the addition of a cognitive element of empathy, many theorists adopted a conceptualization that contained both components (e.g., Davis, 1983); whereas others felt empathy was exclusively one or the other components (e.g., Grief & Hogan, 1973).

Measurement of Empathy

The dissent over the components of empathy was evident in the forms of measurement that developed for general empathy. Some researchers who subsequently developed scales created measures that were only reflective of one component of empathy. The Questionnaire Measure of Emotional Empathy (QMEE; Mehrabian & Epstein, 1972), for instance, is a measure strictly of affective empathy. The items on the QMEE reflect the original conceptualization of empathy. That is, they were written to strictly pertain to affective empathy, the act of feeling as another feels, as is consistent with the original conceptualization of empathy by Vischer (1873). On the other hand, some researchers took the opposite perspective, believing empathy was only comprised of an intellectual

component. An example of this is the Hogan Empathy Scale (HES; Hogan, 1969), whose items question the ability to cognitively take on the perspective of another individual.

Finally, there were also those researchers who believed that empathy had both affective and intellectual components. One of the most well-known and widely used scales for measuring empathy is the Interpersonal Reactivity Index (IRI) developed by Davis (1983). The items on the IRI were written to reflect Davis's (1983) belief that empathy was comprised of both an affective and intellectual component. To this point, the IRI has both cognitive (Perspective Taking and Fantasy) and affective (Empathic Concern and Personal Distress) subscales. Over time, consensus arose that empathy was comprised of both cognitive and affective components, and more recent scales developed to measure empathy also reflect this shift in belief (e.g., Hojat, Mangione, Kane, & Gonnella, 2005; Jolliffe & Farrington, 2006).

Relationship to Other Variables

Given the vast number of published measures of general empathy and the many years of research invested into the topic of empathy, there has been quite an extensive body of literature examining the relationship of general empathy to other variables. Empathy has been found to be positively correlated with both prosocial and socially competent behavior in a meta-analysis (Eisenberg & Miller, 1987). Furthermore, Batson (1990) argues theoretically that altruism is only possible through exhibiting empathy for others. Regarding more negative outcomes, low levels of empathy have been shown to be related to higher levels of criminal offending (Jolliffe & Farrington, 2004), relational aggression (Loudin, Loukas, & Robinson, 2003), social dominance orientation (Pratto, Sidanius, Stallworth, & Malle, 1994), and intergroup aggression (Struch & Schwartz, 1989).

Cultural Empathy

Although there is a need for continued research of general empathy, a type of empathy arguably more relevant in our diverse society today is that of cultural empathy, or empathy directed toward those of a different culture. General empathy encompasses empathy towards any individual other than oneself (e.g., toward a parent); whereas cultural empathy is a more specific case of general empathy relative to that shown towards an individual of a different culture than one's own (e.g., toward a person of a different socioeconomic status). Many conflicts and tense situations could be avoided when interacting with people from different cultures if a certain level of cultural empathy is able to be exhibited. Cultural empathy is also relevant in educational, medical, and counseling settings, each of which will be explained in turn.

Educational settings. There are many instances in which the need for cultural empathy, and the measurement of, is especially prevalent. Firstly, in a multicultural awareness course, it is expected that students will develop in levels of cultural empathy (E. L. Brown, 2004). The true purpose of the course is targeted at not only general empathy, but a form of empathy specific to those from different cultures. The course is intending to increase levels of cultural empathy in the students that participate; however, in order to determine if this course is meeting its goals, it is essential to have a firm conceptualization and solid measurement of the construct.

A second area in which cultural empathy is an essential component is study abroad programs. It is important that those students who decide to study abroad are able to empathize with those from a variety of cultural backgrounds (Kitsantas, 2004). Again, this necessitates the development of a construct geared toward not just exhibiting empathy in general, but a specific form of empathy, empathy toward those from different cultures. In

the diverse society we live in, it is expected that everyone will encounter those from different cultures more frequently (Nguyen, 2003). However, when studying abroad, it necessarily dictates by the nature of the program that those from different cultures will be encountered. It may be of use to ensure prior to acceptance to a study abroad program that students have an acceptable level of cultural empathy. If someone has below an acceptable level, there could be training programs designed for students entering into a study abroad environment. It certainly would be of interest to measure growth in cultural empathy as a result of studying abroad. This makes research and measurement of cultural empathy particularly important and highly relevant to study abroad programs (Kitsantas, 2004). Moreover, teachers and researchers in K-12 settings have increasingly been faced with dilemmas concerning the most appropriate way to instruct in increasingly ethnically diverse settings, which could be eased with increased cultural empathy (Cockrell, Placier, Cockrell, & Middleton, 1999; Gay & Kirkland, 2003).

Medical settings. An area where empathy is of grave importance is in physician/patient interactions (Jolliffe & Farrington, 2006). If physicians cannot understand and feel how a patient is feeling, or vice versa, their interactions will suffer. This misunderstanding could lead to a breakdown in communication preventing proper diagnosis or assistance. It is crucial that there is a level of empathy between the physician and patient. Hojat et al. (2005) developed the Jefferson Scale of Physician Empathy (JSPE), after determining that empathy was such an integral component in this environment that a measure specific to physician empathy should be developed. However, an issue mentioned although not explicitly stated by Hojat et al. that is of increasing importance is that of cultural empathy in those environments. Culture adds an additional facet to empathy that creates something of potentially even greater importance to physicians.

Counseling settings. Finally, a significant area where cultural empathy is of the upmost importance is in counseling settings. Chung and Bemak (2002) discussed in detail how the environment in counseling settings is a sensitive one. It is vital that the patient is able to feel understood by their counselor. This allows there to be an open, inviting environment, which is essential for counseling. Again, an additional barrier is raised when empathy needs to be experienced between people from varying cultures. That is, in addition to all the complications that can make experiencing empathy a challenge, there is a challenge of coming from different cultural backgrounds to take into account as well. To this point, Scott and Borodovsky (1990) go as far as to argue that standard counseling settings are not appropriate when counseling is between two people from differing cultures. Despite the challenges of exhibiting cultural empathy, it could help to alleviate stigmas associated with counseling (Sue, Fujino, Hu, Takeuchi, & Zane, 1991; Sue & Zane, 1987). Pamela Hays (1960) also proposed a model of multicultural counseling competency, which she termed "ADRESSING," to highlight the various cultural backgrounds necessary to consider in multicultural counseling (e.g., age, disability, religion). Various techniques are laid out by Parson (1993) for what he terms Ethnotherapeutic Empathy (EthE). These techniques have been designed specifically to aid counselors needing to exhibit cultural empathy towards their patients. The numerous techniques used in counseling demonstrate how important cultural empathy is in counseling sessions. In order to evaluate the effectiveness of techniques, such as the EthE techniques developed by Parson, it is necessary to measure not only empathy, but empathy specific to those from other cultures.

It is evident there are many careers and occupations in which cultural empathy is a significant component. As cultures are merging together in our society in increasing proportions, it is even more essential for the construct of cultural empathy to be developed.

Ridley and Lingle (1996) noticed a lack in research on cultural empathy. In the belief that this was a construct essential to understanding human behavior, Ridley and Lingle developed a thorough conceptualization of the construct of cultural empathy, which closely mirrors that of general empathy, but more narrowly focuses only on empathy exhibited toward those of different cultures (e.g., race, socioeconomic status). The major difference between the conceptualization of cultural empathy and general empathy is the inclusion of a communicative domain in the construct (Ridley & Lingle, 1996). The communicative domain is focused on the expression of empathetic thoughts and feelings. Researchers may argue that the communicative domain is also integral to general empathy; however, it currently is only included in the theoretical conceptualization of *cultural* empathy.

Although many scales have been created for general empathy (e.g., Davis, 1983; Hojat et al., 2001) and the need has been discussed for an examination of empathy through a cultural lens (e.g., Chung & Bemak, 2002; Parson, 1993), the only published scale at this time measuring any form of cultural empathy is the Scale of Ethnocultural Empathy (SEE; Wang et al., 2003). Cultural empathy applies to many types of cultures (e.g., disability status, socioeconomic status), but a more specific type of cultural empathy, ethnocultural empathy, has been defined as "empathy directed toward people from racial and ethnic cultural groups who are different from one's own ethnocultural group" (Wang et al., 2003, p. 221). There have been scales developed to measure intercultural competency and attitudes that contain a subscale devoted to empathy (e.g., Munroe & Pearson, 2006), indicating the belief that empathy is a key component to having the ability to interact competently with those from other cultures. However, despite the need for measurement of this construct, the only known, published scale devoted exclusively to ethnocultural empathy at this time is the SEE.

Scale of Ethnocultural Empathy

Theoretical conceptualization. Wang and her colleagues (2003) developed the SEE due to the lack of measures assessing either cultural or ethnocultural empathy. It was developed by counseling psychologists, but intended to be generalized to areas outside the counseling realm. In developing a theoretical conceptualization for ethnocultural empathy, Wang et al.'s (2003) theory was modeled on the theory developed by Ridley and Lingle (1996) for cultural empathy. As mentioned, Ridley and Lingle's conceptualization also aligns closely with the commonly accepted theoretical conceptualization of general empathy, comprised of both a cognitive and affective domain. Specifically, the creators of the scale developed items to match three theoretical domains of ethnocultural empathy, which parallel Ridley and Lingle's (1996) conceptualization. As was discussed, Ridley and Lingle examined cultural empathy within the context of therapeutic counseling and speculated that there were three domains to this construct: (a) cognitive, (b) affective, and (c) communicative which were termed intellectual, empathic emotions, and communicative, respectively, by Wang et al. (2003). Both sets of authors argue that although these three domains overlap considerably, each represents a distinct aspect of empathy. Each of these domains is described in detail below.

Intellectual. Ridley and Lingle (1996) believed there were separate aspects that comprised each of the domains of cultural empathy. They theorized that the cognitive domain included both the aspects of perspective taking, or the ability to cognitively take on the point of view of another, and cultural self-other differentiation, or the ability to differentiate your own cultural background from that of another. These two aspects combined comprise that of the cognitive domain, or the intellectual component, as Wang et al. (2003) name it, which involves acknowledging and understanding a person from a

different cultural background. This relates highly with the cognitive components added to empathy theory by Mead (1934).

Empathic emotions. Likewise, Ridley and Lingle (1996) theorize the affective domain, which parallels the empathic concern domain of the SEE (Wang et al., 2003). This domain includes both the aspects of vicarious affect, or feeling emotion as the person from a different cultural background does, and expressive concern, or the outpouring of empathic emotions towards the person of a different background. This domain parallels well with the original concept outlined by Vischer (1873) and Titchener (1924) for general empathy.

Communicative. Finally, the communicative domain (so called by both Ridley and Lingle, 1996 and Wang et al., 2003) includes both probing for insight, which consists of the aspects of asking necessary questions to enhance understanding of another person's perspective, and conveying accurate understanding, which consists of effectively expressing your understanding of others.

Scale development. To create the SEE, Wang et al. (2003) recruited a team of six doctoral counseling students to create 71 items, which were mapped to each of the three domains. Backwards translation was conducted on these 71 items to affirm the match between each item and the domain for which it was written. Specifically, three people from varying areas (one journalism master's student, one doctoral student in counseling psychology, and one PhD in counseling psychology) were asked to match all items back to the constructs, rate the match on a scale of one to six, and provide comments. An item was deemed appropriate if the majority of raters (two of three) appropriately matched the item with the construct it was intended to measure and rated the appropriateness of the correct match with an average of at least four out of six on the scale. Based on the results, nine items were removed and six items were revised. The six revised items then proceeded

through the same backwards translation process and were approved for inclusion, leaving a 62-item scale.

A principal components analysis (PCA) with an oblique rotation was then conducted on the scale using a sample of 323 undergraduates from a Midwestern university. Items with skewness or kurtosis greater than an absolute value of 2.0 were removed prior to the PCA. The number of components retained was based on a scree plot, which indicated that one to four components be retained, as well as interpretability, which led to the retention of a four-component model. Items with low factor loadings[1] (less than an absolute value of .40) or items that loaded greater than an absolute value of .30 on more than one factor were removed, leaving a 31-item scale at the conclusion.

It appears that a different four-component PCA solution was obtained using the reduced 31-item scale. Wang et al. (2003) then proceeded to name the four components based on inspecting the contents of the items that loaded on each component. Table 1 presents the definitions of the four components, which were labeled the following (with the acronym and number of items on the component in parentheses): acceptance of cultural differences (ACD; 5), empathic perspective taking (EPT; 7), empathic awareness (EA; 4), and empathic feeling and expression (EFE; 15). Wang et al. only reported the highest "loading" for each item. Therefore, we do not know the extent to which items cross-loaded on components, although it is suspected that the cross-loadings are low given the way in which items were selected for the 31-item scale.

Wang et al. (2003) also report coefficient alphas with values between .71 and .91 for the four subscale scores, which were computed from summing the items that "load" on each component. They also report a full scale alpha of .91 given the inappropriately championed

[1] It is unclear whether Wang et al. are referring to pattern coefficients or structure coefficients by using the term "factor loading".

higher-order structure for the SEE. The correlations among the four components were all moderate in size (ranging from .29 to .59).

Table 1

SEE Subscales and Definitions

Subscale Name	Definition
ACD (5)	"items that center on the understanding, acceptance, and valuing of cultural traditions and customs of individuals from differing racial and ethnic groups"
EPT (7)	"items that indicate an effort to understand the experiences and emotions of people from different racial and ethnic backgrounds by trying to take their perspective in viewing the world"
EA (4)	"items that appear to focus on the awareness or knowledge that one has about the experiences of people from racial or ethnic groups different from one's own"
EFE (15)	"items that pertain to concern about communication of discriminatory or prejudiced attitudes or beliefs as well as items that focus on emotional or affective responses to the emotions and/or experiences of people from racial or ethnic groups different from one's own"

Note. Subscale names and definitions are utilized from Wang et al. (2003), p. 224. ACD = Acceptance of Cultural Differences, EPT = Empathic Perspective Taking, EA = Empathic Awareness, EFE = Empathic Feeling and Expression.

Relation of subscales to theory. One of the primary reasons for using PCA or more appropriately, factor analytic techniques, at this stage in instrument development is to provide evidence that the way the items relate to one another corresponds with theory. However, the extent to which the PCA results align with the three theoretical domains for which the items were written is unclear. According to the authors, the four components closely replicate the three original domains. Unfortunately, the reader lacks the necessary information to assess the level of alignment between the theory and the defined components. For example, it would be helpful to know which of the three domains the items were originally written for and matched to by backwards translation.

Nonetheless, the alignment is of concern because the number of components does not equal the number of theoretical domains and the description of each component (i.e., the component names) and the theoretical domains do not correspond with each other. Wang et al. (2003) briefly suggested that the intellectual aspect seems to include both perspective taking and awareness, but little detail is provided for this link and the other two defined subscales (i.e., ACD and EFE) are not explained in light of the theory by which the items were developed. It is of interest to know how the solution aligns and how it departs from the theoretical conceptualization. The authors suggest one plausible explanation for why the PCA results do not align with theory, which is that ethnocultural empathy is more complex than the original three-domain conceptualization considered. Another explanation is that the original conceptualization was correct and the PCA indicated that some items or components are not aligned with theory. Although there were two conclusions to consider in light of the PCA results, Wang et al. favored only one conclusion: the theoretical conceptualization of ethnocultural empathy consists of four domains, not three. Their bias in favoring this conclusion is evident in the confirmatory factor analyses they executed, which tested the fit of models utilizing the four-domain conceptualization of the construct that emerged in the PCA results. The two models they considered in their CFAs included a four-factor model and a higher-order factor model, where a single higher-order factor was specified to explain the relationships among the four lower-order factors. Absent from their CFAs was a three-factor model, based on the original conceptualization of the construct.

Despite their omission of the three-factor model, the results of their CFAs are useful in that they allowed for a more appropriate and rigorous test of the instrument's structure than the PCAs. Wang et al. (2003) used a sample of 340 undergraduates from two Midwestern institutions in their analyses and found acceptable fit for both the four-factor

and higher-order models. The use of information criteria[2] to compare the fit of the two models led Wang et al. to champion the higher-order model. Although their results do provide some evidence that a higher-order model is appropriate for the data, the value of their results is called into question because of their use of item parceling. Rather than factor analyzing the individual item responses, Wang et al. analyzed aggregated responses based on subsets of items, which are known as item parcels. Wang et al. provide two (unjustified) reasons for their use of item parcels: (a) to simplify the model and (b) to avoid the results being overly influenced by the unique features of the items. Unfortunately, there are several problems associated with item parceling (Bandalos & Finney, 2001). One problem in using item parcels particularly relevant to the SEE, or any scale in the early stages of development, is the inability to assess the functioning of individual items.

Validity evidence. The measurement of ethnocultural empathy is essential in areas such as counseling (Chung & Bemak, 2002), study abroad programs (Kitsantas, 2004), and educational programs (E. L. Brown, 2004). A quality measure of ethnocultural empathy is needed in order to ensure programming aimed at ethnocultural empathy is effective and to assess whether those needing to exhibit ethnocultural empathy in their career (e.g., counseling) have appropriate levels of the construct. However, few studies have examined any form of validity of the SEE. Benson (1998) outlines three stages of construct validation: (a) substantive, (b) structural, and (c) external. The substantive stage of validity was explored when outlining the construct based on Ridley and Lingle's (1996) theory.

Structural. Benson's (1998) second stage of construct validation is the structural stage. This stage examines whether the structure of the scale holds under different circumstances

[2] Wang et al. erroneously concluded that the two models were not nested and therefore did not conduct a chi-square difference test to assess if the more complex model (the four-factor model) fit significantly better than the higher-order model.

(e.g., over time, across samples), often through the use of factor analytic techniques. Some evidence has been found supporting the 4-factor structure determined by Wang et al. (2003) for the SEE. Chan, Blalock, Cardoso, Steven, and Eun-Jeong (2007) conducted a replication of the CFAs employed by Wang et al. (2003). They found the higher-order model with four lower-order factors yielded acceptable fit with scores from a group of counseling students, which provides some structural validity evidence for the SEE; however, as it was a replication, item parceling was used, which could be masking problems with individual items. Rasoal (2009) also conducted some research on a Swedish-translated version of the SEE. Unfortunately, their analysis was, similarly to Wang et al.'s, a principal components analysis. Unlike Wang et al.'s study, they used orthogonal rotation (varimax, specifically), which forces the factors to remain uncorrelated. Given the theoretical conceptualization dictates the components of empathy to be correlated, the use of this rotation calls into question their results. Their results indicated 25 of the 31 items "loaded" on 4 different components, in a manner similar to Wang et al. (Rasoal, 2009).

All published research exploring the SEE's factor structure to this point used a PCA, not an exploratory factor analysis (EFA; Rasoal, 2009; Wang et al., 2003). Although results are similar for PCAs and EFAs under certain conditions, an EFA is the preferred, appropriate technique when investigating the factor structure of a set of items (Fabrigar, Wegener, MacCallum, & Strahan, 1999). Furthermore, although CFAs conducted provide a more rigorous test of the structure of the SEE, inappropriate methodology, namely the use of item parcels, was used in all CFA analyses up to this point (Chan et al., 2007; Wang et al., 2003), prohibiting an inspection of how individual items are functioning (Bandalos & Finney, 2001).

Given the lack of research conducted on the SEE and because the analyses conducted thus far were based on questionable methodology, Gerstner and Pastor (2011a) conducted an EFA on the SEE. Their results revealed a similar factor structure to Wang et al.'s subscales, but with some distinct differences (Gerstner & Pastor, 2011a). Most notably, one item (item 29) had a pattern coefficient greater than .30 (i.e., loaded) on a different subscale than in Wang et al.'s original scale, and one item (item 2) did not load on any factor. There were also a number of items on the Empathic Feeling and Expression subscale with low pattern coefficients (less than an absolute value of .40). Gerstner and Pastor (2011b) replicated these results on an independent sample, which again revealed the same problematic items in an exploratory framework.

In order to conduct a more rigorous test of the factor structure using appropriate methodology, Gerstner (2011) examined the SEE using confirmatory factor analysis *without* item parcels. This represented the first study that allowed for the examination of individual item functioning in a confirmatory framework. All specified a priori models, including those based on Wang et al.'s (2003) original model and the adjustments suggested by Gerstner and Pastor (2011a, b), showed vast local misfit. Therefore, post hoc modifications were made by examining correlation residuals and item content with consideration for the theoretical conceptualization under which the SEE was developed. This led to the championing of a revised 19-item scale fit to a bifactor model (with a negative wording factor). It is necessary to replicate this study using an independent sample prior to drawing any strong conclusions regarding the structure of the scale; however, these results do indicate the revised 19-item SEE is a promising measure of ethnocultural empathy.

External. The third stage of Benson's (1998) model, namely the external stage, involves testing relations of the scale with other constructs, or examining known group

differences. That is, this stage examines whether the scale relates to and/or measures known differences in predictable ways. Studies examining external validity of a scale may be premature if the factor structure is not supported. However, there have been a number of studies supporting, with some departures, Wang et al.'s (2003) structure for the SEE (e.g., Gerstner & Pastor, 2011; Rasoal, 2009), warranting the examination of external validity. The only published external validity evidence for the SEE has been found by the scale developers (Wang et al., 2003). Specifically, Wang et al. found that the SEE related in predictable ways to both the IRI (Davis, 1983) and the Miville-Guzman Universality-Diversity Scale (M-GUDS; Miville et al., 1999). Furthermore, the Balanced Inventory of Desirable Responding (BIDR; Paulhus, 1984) was used to indicate whether students' responses were influenced by social desirability bias. There was no relationship between scores on the BIDR and the SEE, suggesting students did not respond out of a need to have socially desirable responses. That is, this provides further support that scores on the SEE are indicative of levels of ethnocultural empathy.

Change Over Time

There has been some validity evidence found using Benson's (1998) three stages of construct validation, which supports the functioning of the SEE at a single time point. However, it is also necessary to examine the instrument's functioning across multiple time points, given interest in how empathy changes over time. At the early stages of general empathy research, there was a debate among researchers as to whether empathy was a trait or a state. Those arguing empathy was a trait argued individuals had innate levels of empathy that was not influenced by the environment (Hogan, 1969; Mead, 1934). On the other hand, other theorists argued that empathy was a state, influenced by the situations that individuals were placed in (e.g., Rogers, 1975).

Traditionally, psychoanalytic theorists, psychotherapy researchers, and social and developmental psychologists believed that empathy was a trait, a general ability that is not altered by situations (Duan & Hill, 1996). Holding this belief led to research focused on how innate levels of empathy relate to other characteristics like altruism (e.g., Batson, 1990). On the other hand, those arguing empathy was an alterable state researched changes in empathy levels over time (e.g., before and after a therapy session). This view also lends itself to the necessity of developing programs targeted at impacting empathy positively.

Over time, consensus arose that empathy is an alterable state (Duan & Hill, 1996). As the theoretical conceptualization of cultural empathy is based heavily in empathy theory, it is of no surprise that these theorists as well also noted the belief that cultural empathy is an alterable state, influenced by situations and environments encountered by individuals (Rasoal, Eklund, & Hansen, 2011; Ridley & Lingle, 1996; Wang et al., 2003). Given cultural empathy is believed to change over time, this increases the need and desire for measuring ethnocultural empathy at various time points.

As ethnocultural empathy is a newer area of research, there have been *no* studies examining the way it changes over time or across various situations. Authors discussing ethnocultural empathy, including the developers of the theoretical construct (Wang et al., 2003) are in consensus that it is a dynamic state, molded by the situations individuals are placed in (Rasoal et al., 2011). However, despite this conclusion, there have been no studies examining how individuals do change over time in ethnocultural empathy. Nonetheless, much of the theoretical conceptualization of ethnocultural empathy mirrors that of general empathy, and changes in empathy have been researched over the years. Specifically, changes in empathy over time have been a topic researched in medical, educational, and developmental research.

Medical. Bellini and Shea (2002) were interested in examining changes in empathy over the course of physician's training. Empathy is a critical component of a physician's career, as their job provides constant human interaction. There has been concern raised that medical students' levels of empathy will decrease over the course of their physician training (Bellini & Shea, 2002; Hojat et al., 2009). This could lead to desensitization when interacting with patients. Although earlier research found there was no decline in empathy across the time of medical training (Markham, 1979; Zeldow & Daugherty, 1987, as cited in Hojat et al., 2004), a host of more recent research has found that empathy does decrease over the period of time in which future physicians are in medical school (Bellini & Shea, 2005; Chen, Lew, Hershman, & Orlander, 2007; Hojat et al., 2004; Hojat et al., 2009; Newton, Barber, Clardy, Cleveland, & O'Sullivan, 2008). Although disconcerting, this provides compelling evidence that empathy is a state, altered by situational factors.

Educational. In a more positive light, changes in empathy have also been observed as a result of empathy training programs. Long, Angera, Carter, Nakamoto, and Kalso (1999) conducted a longitudinal study analyzing the impacts of a 10-hour training program for couples designed to train them in feeling empathy for their partner. As expected, they found that empathy did increase over a six-month period (Long et al., 1999). This indicates targeted programming can cause a change in empathy, providing further support that empathy is an alterable state.

Developmental. Furthermore, research conducted on adolescent populations using the IRI (Davis, 1983) indicated that they increase in some aspects of empathy, namely Perspective Taking, a cognitive subscale, and Empathic Concern, an affective subscale (Davis & Franzoi, 1991). Davis and Franzoi (1991) also report that adolescents decreased in

predictable ways on an affective subscale, namely Personal Distress. These results demonstrate how natural development can result in changes to levels of empathy.

All the current examples regarding change used general empathy measures; however, these studies may have been even more suited to use ethnocultural empathy. In many cases, measuring ethnocultural empathy could address the topic of interest more directly than the studies that were conducted on general empathy. Although research has examined changes over time in general empathy in the areas mentioned, many areas greatly need research on changes in ethnocultural empathy over time.

For instance, no research is available directly assessing changes in ethnocultural empathy due to counseling, but the need is especially great in this area. Chung and Bemak (2002) argue that cultural empathy is an essential component to multicultural counseling. Even though they do not go as far as to say changes must be measured based on intentional programming, they call the reader's attention to the necessity for an increase in cultural empathy in these situations.

Forster (2006) describes in detail training programs instituted for cultural empathy implemented for those living outside of their home country. In order to assess the effectiveness of such programs, it is essential to have an instrument that can reliably and validly measure change in ethnocultural empathy over time. Similarly, Kitsantas (2004) outlines a case for assessing the development of cross-cultural skills (of which ethnocultural empathy would surely fall under) as a function of study abroad programs. In another educational setting, E. L. Brown (2004) has researched what factors lead to change in awareness of other cultures as a result of a multicultural course. It is clear the desire is great for a measure of ethnocultural empathy.

Measuring change. It is evident that there is a need to measure change in ethnocultural empathy over time. However, prior to utilizing the SEE to measure change over time, it is necessary to first establish that the measurement properties of the SEE remain invariant across measurement occasions. It should be ensured that the SEE is assessing the same construct on the same metric over time. In other words, it is important to establish that the SEE has measurement invariance. Determining the SEE is invariant across measurement occasions can strengthen the claim that changes in scores are attributable to changes in respondents' levels of ethnocultural empathy. Without ensuring the instrument maintains relatively stable in its measurement properties, it is unclear whether any observed changes in ethnocultural empathy are due to true change in the construct or to changes in the functioning of the measure or the conceptualization of the construct. The literature on measurement invariance has defined three types of change that can occur: (a) alpha, (b) beta, and (c) gamma (Riordan, Richardson, Schaffer, & Vandenberg, 2001; Vandenberg & Lance, 2000). These three types of change will be discussed in turn.

Alpha change. Alpha change represents a true change in the population, or in other words, an actual change in the levels of the construct over time (Golembiewski et al., 1976; Riordan et al., 2001). Although some specific research studies may be interested in examining other forms of change, this is the form of change most often sought by those in research (Riordan et al., 2001). That is, most researchers want to examine whether a population is exhibiting change in a construct. For example, if a program is developed to change students' empathy, researchers would be interested in measuring an actual change in the construct of empathy. However, there is not a direct way to test the presence of alpha change. Instead, it is necessary to rule out the other two feasible types of change – beta and gamma (T. A. Brown, 2006). If testing indicates that beta or gamma change has not

occurred (i.e., the measure exhibits longitudinal measurement invariance), then it supports the conclusion that changes in the scores reflect true changes in the construct (i.e., alpha change).

Beta change. The second type of change is beta change. Beta change occurs when participants respond differently on the scale at the time points being examined despite having identical levels of the construct. Riordan et al. (2001) defined beta change as occurring when "respondents recalibrate the intervals anchoring the measurement continuum" (p. 53). In other words, although at both time points a respondent has an identical amount of the construct (i.e., empathy), the respondent records different responses across the time points. Beta change can manifest itself in two ways: (a) different pattern coefficients on the factor, or (b) different intercepts. If the pattern coefficients associated with an item are different across time points, it indicates this item was more or less salient (i.e., the item has a higher or lower relationship with the factor, respectively) across the two time points. This difference in pattern coefficients signifies noninvariance in the measure, preventing the measure from accurately measuring change over time. For instance, if an item is more salient at one time point than the other, any observed changes in the average item response may not be reflective of true change in the construct, but merely a change in the saliency of the indicator. If the intercept of an item is different across time, it indicates respondents were utilizing the range of the scale differently at the time points (i.e., responding higher at one time point than another), although their true level of the construct did not change.

Gamma change. The final type of change, gamma change, refers to a change in the conceptualization of the construct itself (T. A. Brown, 2006; Riordan et al., 2001). It could be, for example, that as participants mature over time, they view a construct as being

multidimensional versus unidimensional. If one has little knowledge or exposure to a construct, it could be the conceptualization of the construct broadens over time, exposing new facets/dimensions previously unknown or unacknowledged.

Without assessing longitudinal measurement invariance of a scale, any conclusions made about change over time could be attempting to compare "apples to spark plugs" (Vandenberg & Lance, 2000, p. 9). Therefore, prior to measuring changes over time in ethnocultural empathy using the SEE, the longitudinal measurement invariance of the scale will be examined in a structural equation modeling (SEM) framework, as modeled by Vandenberg and Lance (2000). Specifically, testing measurement invariance utilizing a SEM approach involves testing a series of progressively more stringent models to determine if measurement invariance holds across time points by ruling out the occurrence of beta or gamma change through tests of configural, metric, and scalar invariance.

To test for gamma change, *configural invariance* is examined, which means evaluating fit of the same factor structure at all time points (Vandenberg & Lance, 2000). If the same factor structure fits at all time points, then it can be concluded that students are conceptualizing the construct in the same way across time (i.e., gamma change does not appear to be occurring) and beta change can be examined. Beta change is examined in a SEM framework in two tests: metric and scalar invariance. *Metric invariance* is tested by constraining the unstandardized factor pattern coefficients to be the same at all time points (Vandenberg & Lance, 2000). If metric invariance is violated, this suggests that beta change is occurring, because the saliency of items does not remain the same across the time points. However, if metric invariance is indeed upheld, a stricter test of beta change, scalar invariance, can be tested. *Scalar invariance* is tested by constraining the intercepts to be equivalent across time (Vandenberg & Lance, 2000). If this is upheld, it is concluded

respondents are not utilizing the range of the scale differently across time. That is, observed changes are determined to be indicative of changes in the levels of the construct, and longitudinal measurement invariance is established for the scale.

Other theorists have termed three types of invariance as weak, strong, and strict (Vandenberg & Lance, 2000). Prior to establishing any of these three forms of invariance, configural invariance must be established. *Weak invariance* then refers to metric invariance, ensuring the items have the same saliency to the factor across time. The second type, *strong invariance*, corresponds with scalar invariance, ensuring respondents utilize the scale in the same manner over time. Finally, *strict invariance* goes beyond the previous three types of invariance discussed to test whether the residual variances of individual items (i.e., item variance unexplained by the factor) are equivalent across time. The level of invariance required for any given study is dependent on the types of questions being investigated (Steenkamp & Baumgartner, 1998). However, most researchers agree that change in the construct over time can be evaluated if *strong* invariance is established. Thus, in order to establish longitudinal measurement invariance of a scale, the equality of residual item variances is not necessary.

If longitudinal measurement invariance holds across time points, then changes in ethnocultural empathy can be examined. If measurement invariance is not satisfied, it will indicate that the SEE cannot be used to accurately measure change in the level of the construct over time. That is, the measure functions differently across time, so any observed changes are not reflective of true, alpha change in the construct. To date, there have been no studies examining the longitudinal invariance of the SEE. Although previous studies provide some supporting validity evidence for the SEE scores at a single time point, it is

important to assess the extent to which the psychometric properties of the instrument hold over time to draw accurate conclusions about changes over time.

Recommended Future Research of Ethnocultural Empathy

As ethnocultural empathy is still a relatively new and developing construct, it is of the upmost importance that further validity evidence is provided for it. The construct is crucial to many domains (e.g., counseling, education), so future research on the SEE, the only published measure of ethnocultural empathy, will greatly aid research on these programs. Given the focus of many organizations hoping to measure change in ethnocultural empathy over time, it is especially imperative to examine the longitudinal invariance of the SEE. That is, it is crucial to assess whether the factor structure holds at each time point and if so, whether the parameters of the factor model remain invariant across time. Longitudinal invariance studies are crucial prior to using the scale to assess change over time.

III. Method

A series of models were fit to data collected at two time points in order to assess the longitudinal measurement invariance of the Scale of Ethnocultural Empathy (SEE). The bifactor model[3] (see Figure 1) proposed by Gerstner (2011) was assessed at both time points using a sample of over 500 college students from which data were collected in August 2008 and again in February 2010. In the sections that follow, the sample, data screening techniques, and measurement invariance evaluation procedures will be explained. Although the primary purpose of this study is to assess the longitudinal measurement invariance of the SEE, if measurement invariance is established, changes over time can be examined at the latent level.

Participants

The sample was comprised of a matched sample of 557 students from a midsized, mid-Atlantic university. Participants with at least one item response to the scale at both time points were retained because maximum likelihood allows for the inclusion of missing data. The sample was 67.5% female and racially, it was 80.2% White, 5.5% Asian American, 3.0% Black or African-American, 3.4% Hispanic, 0.2% Pacific Islander, 0.4% American Indian, and 7.3% unspecified. The students in this sample were required to complete a battery of instruments, one of which included the SEE, as part of a university-wide assessment day in August 2008 and again in February 2010. The students completed the SEE in August 2008 the Friday prior to the start of classes their freshman year. The February 2010 assessment occurred after these students had completed between 45 and 70 credit hours. Based on the last two digits of their student identification number students were assigned to testing rooms, which had different batteries of instruments. Because students were randomly

[3] This factor structure modified Wang et al.'s (2003) original model.

selected to tests the demographic make-up of this sample is reflective of the demographics of all incoming students at the university. Assessment day is used to collect data for program assessment purposes at the university. All students are required to attend; if they do not, a hold is placed on their record. However, for the majority of assessments, including the SEE, they do not receive scores or feedback on their performance. The SEE was administered for the purpose of collecting data for the Sociocultural Domain of the General Education program. The SEE is relevant to this domain, because one of the goals of this area of General Education is that "students gain an understanding of the relationship between the individual and a diverse community" (http://www.jmu.edu/gened/cluster5.shtml).

Data Screening

Data were screened independently at time point 1 (August 2008) and again for time point 2 (February 2010) for all following procedures. Changes suggested by the data at one time period were also implemented at the other time period (e.g., removal of items or cases). Maximum likelihood was chosen as the estimation procedure for the longitudinal measurement invariance models because it provides the most unbiased, efficient, and consistent estimates of our parameters (Olsson, Foss, Troye, & Howell, 2000). Because maximum likelihood estimation assumes multivariate normality of the item responses, data were screened for normality prior to fitting the models to the data. Skewness and kurtosis for the item responses were evaluated to examine univariate normality with values less than an absolute value of 2 and 7, respectively, indicating that the responses were approximately univariate normal. Multivariate normality was examined utilizing Mardia's normalized kurtosis. An established cut-off has not been determined, but values greater than 3 may indicate nonnormality in the data (Bentler & Wu, 2003). If nonnormality was present in our

data, the Satorra-Bentler scaled chi-square and robust standard errors would be used to adjust for kurtosis.

To screen for multivariate outliers, estimates of Mahalanobis distances were calculated with DeCarlo's (1997) macro. Those observations having larger Mahalanobis distance values relative to the other cases were considered for removal from the data set after examining their response patterns. Recall, changes suggested at one time point were carried out on the other time points, as well. Therefore, if a case was a multivariate outlier at one time point, they were removed from both data sets entirely. Because excessive relationships among items can be problematic in factor analysis, multicollinearity was also examined using Pearson bivariate correlation coefficients, with values greater than an absolute value of .85 indicating problematic multicollinearity. If correlations exceeded an absolute value of .85, one of the two items was removed after inspection of item content. Multicollinearity was also assessed using tolerance, which represents the proportion of variance in the item not shared with the other variables. Items with tolerance values less than .10 were considered to overlap considerably with the remainder of the items on the scale and were removed prior to analysis.

Procedure

A series of competing models (described in detail below) was tested in order to examine the longitudinal measurement invariance of the SEE. For each of these models, the same factor structure was used, which follows a revised 19-item bifactor model championed by Gerstner (2011) for the SEE (see Figure 1). The four factors are as follows (with the number of items on the factor in parentheses): empathic feeling and expression (7), empathic perspective taking (4), acceptance of cultural differences (4), and empathic awareness (4). There were also 8 items that cross-loaded onto a negative wording factor in order to model

the variability the negatively-worded items shared above and beyond their relationships to the substantive factor. The measurement invariance models were tested using this general factor structure. All substantive factors were allowed to covary because theory dictated the factors would be intercorrelated. However, the negative wording factor was only permitted to correlate with itself across time (i.e., the correlations between the negative wording factor and all substantive factors were constrained to be zero, because theory dictates the method factor represents systematic variability unrelated to the substantive factors). In addition, the errors of the same items across time points were allowed to covary, because the items are presumed to share unique variance that is temporally stable (T. A. Brown, 2006). For each model described below, a variety of indices were consulted to assess the fit of the model to the data.

Fit indices. The robust root mean square error of approximation (RMSEA) and the standardized root mean square residual (SRMR) were used as measures of absolute model-data fit. The robust RMSEA evaluates the extent to which the proposed model deviates from the observed data per degree of freedom with values less than or equal to .05 indicating good model fit, values between .05 and .08 indicating acceptable fit, and values between than .08 and .10 indicating poor fit (Little et al., 2007). The SRMR represents on a standardized correlation metric the square root of the average squared residuals between the model-implied and reproduced covariances, with values equivalent to or lower than .08 indicating good model fit (Hu & Bentler, 1999). The robust CFI, or comparative fit index, was also obtained. The CFI provides a measure of the incremental fit of our model to the data (i.e., the comparative fit of the proposed model to a null model; Yu & Muthén, 2002). CFI values greater than .90 indicate good model fit (Little et al., 2007). Furthermore, standardized covariance residuals (i.e., standardized discrepancy between original and model-implied

covariance matrix), correlation residuals (i.e., difference between original and model-implied correlation matrix), and modification indices were examined to diagnose any misfit in the model. Because standardized covariance residuals can be sensitive to sample size, both the standardized covariance residuals and correlation residuals, which represent an effect size of the difference, were examined to compare fit. Fit indices were also used to compare fit across the models. To test whether a parsimonious model fit significantly worse than a more complex model, $\Delta\chi^2$ and ΔCFI were examined with a nonsignificant $\Delta\chi^2$ and a ΔCFI less than .01 (Little et al., 2007) indicating the more parsimonious model did not fit significantly or practically significantly worse than the more complex model.

Longitudinal measurement invariance models. A series of steps were evaluated to assess whether the measurement properties of the SEE were invariant across time. The process involved fitting progressively more stringent models to the data.

Configural invariance. The first step in testing the invariance of the SEE is to attempt to fit the same factor structure at both time points simultaneously. Latent standardization was utilized to set a metric for our latent variable (factor means were set to zero and factor variances were set to one; Little et al., 2007). If an acceptable level of fit was found, it was determined that students were conceptualizing the construct of ethnocultural empathy in the same manner over time, and metric invariance was examined.

Metric invariance. This step in invariance testing involves constraining the unstandardized pattern coefficients to be equal across time. That is, metric invariance measures whether the same relationship exists between each individual item and the factor at both time points. If so, it indicates the item has the same level of saliency to the factor at both time points, and the measure exhibits metric invariance. All factor means were set to zero, and the variances of the factors at the first time point were set to one to standardize

the latent variable, whereas the variance of factors at the second time point were freely estimated (Little et al., 2007). The fit of this model was compared to the configural invariance model. If this model fit significantly worse than the configural model and did not fit well in an absolute sense, then the methods outlined by Rensvold and Cheung (2001) would be utilized to pinpoint which items have invariant and variant loadings across measurement occasions. Metric invariance was examined in two stages given the presence of a negative wording factor. Metric Model A consisted of constraining the unstandardized pattern coefficients for only the substantive factors (the unstandardized pattern coefficients for the method factor were free). If this metric invariance model fit, Metric Model B was examined, which kept the unstandardized substantive pattern coefficients constrained, but also constrained the unstandardized pattern coefficients of the negative wording factor. This two-step approach tested if the loadings for the negative wording factor could be invariant across time. If full metric invariance was upheld, scalar invariance could be examined. Partial scalar invariance could be examined if partial metric invariance was upheld, which means that if only a few items were not metrically invariant across time, the rest of the items could still be examined for scalar invariance.

Scalar invariance. The final step in determining longitudinal measurement invariance of the SEE is scalar invariance. This model tests that the intercepts are equivalent across time. If this holds, it indicates that those respondents with a same level of the factor at both time points are providing the same score on the response scale at both time points. That is, they are utilizing the various response options similarly when no true change in the construct occurs. The intercepts for the items were constrained to be equivalent across time for all items that were determined to be metrically invariant. The latent variables were standardized by setting the mean of the factors at the first time point to zero and the

variance of the factors to one; factor means and variances were freely estimated at the second time point in order to scale the factors (Little et al., 2007). If this model did not fit significantly worse than the metric model, the SEE was determined to exhibit measurement invariance across time, and change in time was assessed.

Latent mean and rank-order differences. If the SEE demonstrated partial scalar invariance, latent mean differences could be examined. Specifically, a latent *t*-test, latent Glass' Δ effect size[4] (i.e., standardized mean difference across time), and a latent test-retest coefficient were obtained. The latent test-retest coefficient provides an estimate of the stability of students' rank-order across time. That is, it provides a measure of whether students change in a consistent manner over time. Although these tests could be obtained at an observed level, evaluating these differences in a structural equation modeling (SEM) framework has the advantage of obtaining error-free estimates. Therefore, SEM allows a more accurate estimate of the true difference across time.

Software. The data screening analyses were performed in SPSS v.19, and the measurement invariance models were fit to the data using Mplus v. 6.1.

[4] $Glass'\Delta = \frac{\mu_{T_1} - \mu_{T_2}}{\sigma_{T_1}}$

IV. Results

Data Screening

Data were screened independently for time 1 and time 2. Therefore, results will be presented separately for each time point.

Normality. Data were screened for normality. Although skewness and kurtosis values were within the desired cutoffs for both times (see Table 2), Mardia's multivariate kurtosis was 38.64 and 56.82 at time 1 and 2, respectively. This indicated the data were multivariately nonnormal, and the Satorra-Bentler adjustment was utilized to correct the χ^2, standard errors, CFI, and RMSEA for nonnormality (Satorra & Bentler, 1994, 2001).

Outliers. No univariate outliers were detected. However, several cases at both time points had high Mahalanobis distance values. After examining the response patterns, one case from time 1 and three cases from time 2 were removed for having nonsensical response patterns (e.g., responding the same to a reverse scored item and non-reverse scored item). The final, effective sample size was 553 after removing these outliers. Recall, these 553 cases contained missing data, because full information maximum likelihood estimation can utilize cases with missing data (Schafer & Graham, 2002).

Multicollinearity. An examination of Pearson product-moment correlation coefficients did not reveal any items that related too highly with one another. Furthermore, all tolerance values were greater than .10, indicating each item had a significant amount of variance to contribute that was not shared by other items.

Descriptive Statistics

Descriptive statistics for all items are presented in Table 2 below. Most responses are near the midpoint of the 6-point response scale. Some increases can be seen in the raw scores from time 1 to time 2 (e.g., Item 1). On average, 9 of the 19 items showed slight

increases in the observed scores. The remaining 10 items showed slight decreases across time, on average; however, no large differences were observed in the scores across time points. All standard deviations are around 1.0 indicating that item scores vary from the mean by about 1 point.

Table 2
Descriptive Statistics for SEE Items

Item[a]	Subscale[b]	Time 1				Time 2			
		M	*SD*	Skew	Kurtosis	*M*	*SD*	Skew	Kurtosis
1	ACD	3.76	1.41	-0.09	-0.86	3.79	1.36	-0.07	-0.85
5	ACD	4.99	1.06	-1.02	0.50	4.97	0.97	-1.10	1.65
10	ACD	4.43	1.35	-0.60	-0.47	4.39	1.27	-0.51	-0.56
29	ACD	3.96	1.31	-0.17	-0.75	4.10	1.25	-0.22	-0.65
4	EPT	2.90	1.81	0.52	-1.19	3.18	1.75	0.26	-1.35
19	EPT	3.22	1.54	0.27	-0.95	3.29	1.41	0.25	-0.75
28	EPT	3.60	1.33	0.06	-0.63	3.75	1.26	0.01	-0.76
31	EPT	3.27	1.27	0.29	-0.51	3.56	1.21	0.19	-0.55
7	EA	4.14	1.36	-0.70	-0.19	4.15	1.27	-0.81	0.09
20	EA	4.10	1.20	-0.57	0.02	4.09	1.17	-0.61	0.25
24	EA	4.87	1.00	-0.88	0.84	4.80	0.94	-0.84	1.07
25	EA	4.72	0.96	-0.71	0.79	4.57	0.92	-0.89	1.34
11	EFE	4.67	1.02	-0.65	0.51	4.56	1.00	-0.71	0.78
13	EFE	4.55	1.02	-0.52	0.09	4.58	0.85	-0.39	0.61
14	EFE	4.45	1.09	-0.88	1.38	4.49	0.98	-0.78	1.31
15	EFE	4.90	1.03	-1.05	1.40	4.85	0.90	-0.89	1.66
17	EFE	4.47	1.27	-0.60	-0.37	4.39	1.19	-0.50	-0.32
19	EPT	3.22	1.54	0.27	-0.95	3.29	1.41	0.25	-0.75
21	EFE	4.75	1.15	-0.87	0.31	4.65	1.15	-0.90	0.43

Note. [a]Item numbers correspond to Wang et al.'s (2003) original scale. [b]Subscale abbreviations correspond to Wang et al.'s. Values calculated on $N = 520$ utilizing listwise deletion. Responses with missing data were utilized for the SEM models, because full information maximum likelihood can utilize cases with missing data. However, skewness and kurtosis could only be obtained using listwise or pairwise deletion. Therefore, listwise deletion was used to obtain descriptive statistics for simplicity.

Longitudinal Measurement Invariance Models

A series of structural equation models (SEM) were estimated for the data. A variety of fit indices were obtained to compare the global and relative fit of each model. Results for these analyses are presented in Table 3.

Table 3
Fit of the Tested Longitudinal Measurement Invariance Models, N = 553

Model	M-L χ^2	S-B χ^2	*df*	$\Delta\chi^2$	Δdf	*p*	SRMR	CFI	ΔCFI	RMSEA
Configural	979.24	849.30	601	--	--	--	.043	.959	--	.027
Metric[a]										
Model A	995.38	861.05	616	11.76	15	.697	.046	.959	.000	.027
Model B	1007.88	871.87	623	22.57	22	.426	.046	.959	.000	.027
Scalar	1057.28	919.37	637	47.50	14	.000	.047	.953	.006	.028

Note. [a]Metric model A constrained only the loadings of the four substantive factors to be equivalent across time; whereas metric model B also constrained the loadings for the negative wording factor.

Step 1: Configural Invariance. The configural model allowed the factor pattern coefficients and intercepts to be freely estimated across time, but constrained the same bifactor model to the data at each time point (see Figure 1). Both global and incremental fit indices revealed acceptable fit. Furthermore, an examination of standardized covariance residuals revealed no localized areas of unacceptable misfit. Although there were correlation residuals greater than a commonly accepted absolute value of .10, they represented less than three percent of the correlation residuals and none of them exceeded .16, indicating there was not substantial localized misfit and configural invariance holds. Moreover, none of the areas of slight, local misfit replicated from previous examinations of the scale.

Step 2: Metric Invariance. Because the model includes a method factor, this stage of metric invariance will be conducted using a two-phase approach consistent with the literature (A. R. Brown & Finney, 2011). First, the unstandardized pattern coefficients of the four substantive factors were constrained to be equal across time (Metric Model A). Then, the unstandardized pattern coefficients of the negative wording factor were constrained (if

Model A demonstrated acceptable fit) to determine if its unstandardized pattern coefficients were also invariant across time (Metric Model B).

Metric Model A. After constraining the unstandardized pattern coefficients of the substantive factors to be equivalent across time, there was not a significant reduction in fit. The lack of reduction of fit was exemplified through the nonsignificant $\Delta\chi^2$, as well as no change in CFI values between the two models. Moreover, correlation residuals greater than .10 were only found in less than four percent of the correlation residuals. All measures of fit indicated that there was metric invariance for all four substantive factors.

Metric Model B. With all unstandardized pattern coefficients constrained for the method factor in addition to the substantive factors, all fit indices were favorable, indicating adequate model-data fit. Furthermore, there was a nonsignificant $\Delta\chi^2$ and no change in CFI values between the fully constrained and partially constrained metric model. An examination of correlation residuals did not indicate gross areas of localized misfit (less than 4% of correlation residuals were greater than .10 and only one exceeded .16). These results indicate that the unstandardized pattern coefficients of the negative wording factor were also invariant across time.

Step 3: Scalar Invariance. The $\Delta\chi^2$ between the full metric invariance model and scalar invariance model was significant, indicating scalar invariance was violated. However, the change in CFI was less than the guideline of .01, indicating that scalar invariance was upheld. Moreover, this model exemplified good fit from a noncomparative standpoint (e.g., all fit indices were adequate given suggested cut-offs). Additionally, there were no problematic mean residuals. Mean residuals represent discrepancies between the observed and model-implied means and indicate the magnitude of misfit in the mean structure of the model. The largest mean residual was deemed nonproblematic (a .08 difference on a 6-point

scale). These results indicated that scalar invariance was upheld, and latent mean differences and test-retest coefficients were examined.

Examining the Final Scalar Model

As the scalar model was determined to yield the most parsimonious fit to our data, various pieces of information were examined for this model. Intercorrelations among the factors, test-retest coefficients, reliability, and variance extracted (presented in Table 4) will be discussed followed by a discussion of latent mean differences and the estimated parameters of the model.

Intercorrelations among factors. The intercorrelations among the substantive factors ranged from .34 to .65 at time 1 and .42 to .68 at time 2 (see Table 4). Within each time point, the ACD and EFE factors were highly correlated. The lowest correlation at both time points was the correlation between the EPT and EFE factors. There were also moderate correlations among the different factors across time (e.g., correlation between ACD_1 and EPT_2), as would be expected given the hypothesized correlations among factors within a single time point based on theory.

Test-retest coefficients. The test-retest correlations, which indicate the consistency in students' rank-order across time, ranged from .68 to .86 for the substantive factors (see Table 4). The lowest correlation across time was present for the EFE factor, indicating there were more differences in rank-order on this factor than the others. However, this is still a moderately high correlation, so the majority of students are maintaining the same rank-order across time. For the EPT factor on the other hand, scores were fairly stable ($r = .86$) across time.

Table 4
Intercorelations among the Factors and Reliability Estimates, N = 553.

Factor	1	2	3	4	5	6	7	8	9	10
1. ACD_1	—									
2. EPT_1	.35	—								
3. EA_1	.41	.43	—							
4. EFE_1	.67	.34	.65	—						
5. $Negative_1$	—[a]	—[a]	—[a]	—[a]	—					
6. ACD_2	**.81**	.45	.38	.56	—[a]	—				
7. EPT_2	.36	**.86**	.35	.31	—[a]	.45	—			
8. EA_2	.33	.38	**.69**	.52	—[a]	.51	.46	—		
9. EFE_2	.45	.37	.47	**.68**	—[a]	.64	.42	.68	—	
10. $Negative_2$	—[a]	—[a]	—[a]	—[a]	**.56**	—[a]	—[a]	—[a]	—[a]	—
ω_1[b]	.52	.72	.71	.78		.55	.75	.76	.78	
ω_2[c]	.65	.74	.71	.79		.68	.78	.76	.80	
Variance Extracted	.32	.43	.40	.31		.33	.46	.44	.35	

Note. Correlations are disattenuated for measurement error because they were calculated in a structural equation modeling framework. All correlations are significant at the $p < .01$ level. Subscripts indicate the time point. Correlations in boldface represent test-retest coefficients. [a]Correlation constrained to be zero. [b]Estimates of omega were based on inclusion of the variance explained by the negative wording factor in the total variance (Green & Yang, 2009; Johnston & Finney, 2010).[5] [c]Estimates of omega were also calculated without the negative wording factor included in the total variance.[6] Given the lack of consensus regarding the calculations of variance extracted with multidimensional items, the values presented were calculated using only the substantive factor variance and the error variance (not including the method factor variance) in the denominator.

Given that some of the correlations (e.g., EA, EFE) were not particularly high, average observed subscale scores from a random sample of 25 students were plotted to examine how individuals were changing over time (see Figure 2). It is evident from the graphs that there is the most change in rank-order in the EA subscale, which corresponds to the second lowest correlation ($r = .69$). The lowest test-retest coefficient was shown for the EFE subscale ($r = .68$). The plots of 25 random cases for the EA and EFE subscales show

[5] The formula utilized to calculate this version of ω was as follows (Green & Yang, 2009; Johnston & Finney, 2010): $\frac{(\sum b_i)^2}{(\sum b_i)^2+\sum e_i+(\sum b_j)^2}$
b_i = the unstandardized pattern coefficients for the substantive factor, b_j = the unstandardized pattern coefficients for the method factor, and e_i = the error terms from the model.

[6]The formula utilized to calculate this version of ω was as follows: $\frac{(\sum b_i)^2}{(\sum b_i)^2+\sum e_i}$
b_i = the unstandardized pattern coefficients for the substantive factor and e_i = the error terms from the model.

some students decreasing substantially whereas others increase substantially, which results in a low test-retest coefficient because students change in rank-order across time. In contrast, the change in students' scores over time tends to be more similar for the ACD (r = .81) and EPT subscales (r = .86).

Reliability. Reliability was adequate for the factors, ranging from .52 to .78 at time 1 and .55 to .78 at time 2 (see Table 4). The lowest reliability was present for the ACD factor, which is logical given all the items on this factor share variance with the substantive ACD factor and the negative wording factor. Therefore, there was less variability available to be partitioned to the true score variance in the ACD factor (i.e., all items have the variance partitioned three ways: common variance, negative method variance, and error variance). Because there was a large difference between the two calculations of ω (other than for the EA factor where none of the items load on the negative wording factor), it was evident that the negative wording factor has strong relationships to the items. The other three substantive factors show more adequate reliability.

Variance Extracted. The variance extracted for each of the substantive factors was less than desirable. A common guideline is that the variance extracted is estimated to be .50 or higher for the factor; this would indicate that a higher amount of variance in the items was attributable to the factor rather than to measurement error (Bandalos & Finney, 2010). Although none of the factors exceed this, the EPT and EA factors at both time points are nearing this value. Moreover, there was still a substantial amount of variance explained by the factors, ranging from .31 to .46.

Latent Mean Differences. Latent mean differences were examined and are presented in Table 5 below. Observed score means are also presented for comparison. An examination of the observed means reveals an increase in ACD and EPT factors across time

and slight decreases in the EA and EFE factors. This pattern holds at the error-free, latent level depicted by Glass' Δ (a latent effect size). A latent *t*-test was also conducted and the standard error and *p*-values are reported for these *t*-tests. The only statistically significant difference across time was in the EPT factor; this difference was also practically significant. Glass' Δ indicates there was on average approximately a .2 *SD* change on the EPT subscale between the two time points (a small effect size). The EPT factor also reported a high test-retest coefficient ($r = .86$). Therefore, students are increasing across time, and this indicates they are increasing in a consistent pattern (i.e., maintaining their rank-order across time). A smaller effect size and nonsignificant *decrease* was observed in the latent difference for the EA factor across time. This decrease was contrary to prediction, as it was anticipated that students would increase in ethnocultural empathy over time. However, this difference was *nonsignificant*, and the EA factor had only a moderate test-retest coefficient ($r = .69$), indicating many students were changing in rank-order across time. Both ACD and EFE showed minuscule differences in average latent scores across time. Neither of these differences were statistically significant and their effect sizes were quite small.

Table 5
Observed and Latent Mean Differences

	Observed Score Means		Latent Means		
	Time 1	Time 2	Glass' Δ[a]	*SE*	*p*
ACD	4.285	4.312	0.019	0.087	.825
EPT	3.432	3.647	0.183	0.045	< .001
EA	4.456	4.404	-0.082	0.044	.062
EFE	4.623	4.581	-0.038	0.045	.392
Negative	—[b]	—[b]	-0.006	0.128	.960

Note. The range of scores goes from 1 to 6. Observed score means were calculated using listwise deletion on an effective $N = 520$. [a]This effect size represents the standardized difference in subscale means across time and is equivalent to the difference in factor means across time, because the means of all factors at Time 1 were constrained to be 0. [b]Means for the negative wording factor can only be calculated in a structural equation modeling framework at the latent level.

The observed (not latent) changes over time were also examined in proportions (see Table 6). That is, the frequency and percentage of students who increased, decreased, and stayed constant on average for each subscale was examined to determine the proportion of students showing these trends. The average latent level of change in the factor can mask how individual students are changing. For example, if some students decreased a substantial amount (e.g., four points), but the majority of students increased slightly (e.g., less than one point on average), the average scores for each subscale across all participants could be negative despite more students showing a slight increase rather than a decrease.

Table 6

Frequencies and Percentages of Observed Change over Time for Entire Sample

	ACD			EPT			EA			EFE		
	n	%	Δ[a]	*n*	%	Δ[a]	*n*	%	Δ[a]	*n*	%	Δ[a]
Positive	246	44.5	.68	294	53.2	.79	209	37.8	.68	215	38.9	.52
Negative	206	37.3	.73	170	30.7	.71	236	42.7	.68	262	47.4	.50
None	94	17.0	.00	65	11.8	.00	99	17.9	.00	64	11.6	.00
Total	546	98.7	.58	529	95.7	.67	544	98.4	.56	541	97.8	.45
Missing	7	1.3		24	4.3		9	1.6		12	2.2	

Note. Because ML estimation utilizes cases with missing values, pairwise deletion was utilized to obtain the most information possible for all cases. [a]This represents the average absolute value change over time for the individual groups.

Examining the proportions makes it clear that although there is a slight decrease on average for the EA and EFE subscales, there is a fairly even split between those increasing and those decreasing, as well as a sizeable percentage showing no change over time. On the other hand, the EPT subscale proportions show that 23% more students are increasing than decreasing. This percentage difference explains the statistically significant increase on average. Finally, the ACD subscale (which showed a slight increase over time) shows that a higher percentage (roughly 7% more) of students showed positive rather than negative change over time. However, 17% of students on this subscale (the second largest percentage of the subscales, following EA) showed no change over time. The absolute value of the

average changes for each of the groups show that on average each group (both showing positive and negative change) shows small amounts of change (less than 1 point on a 6-point scale).

Parameter estimates. The estimates for the various parameters in the final scalar model are presented in Table 7 below.

Table 7
Final Scalar Model Path Estimates

Factor	Item	Intercept	Pattern Coefficients		Error (1-R^2)		Auto-correlation
			Substantive Factors	Negative Wording	Time 1	Time 2	
ACD	1	3.78	0.92 (0.67, 0.63)	0.40 (0.29, 0.31)	.47	.48	.43
	5	4.98	0.47 (0.44, 0.87)	0.33 (0.31, 0.37)	.71	.65	.26
	10	4.40	0.84 (0.61, 0.64)	0.48 (0.35, 0.40)	.51	.43	.33
	29	4.04	0.47 (0.35, 0.51)	0.43 (0.32, 0.37)	.77	.73	.28
EPT	4	2.95	1.10 (0.61, 0.63)		.63	.61	.45
	19	3.15	1.24 (0.80, 0.87)		.36	.24	-.27
	28	3.60	0.80 (0.60, 0.64)	0.53 (0.40, 0.45)	.48	.39	.05
	31	3.36	0.62 (0.49, 0.51)	0.29 (0.23, 0.25)	.71	.68	.14
EA	7	4.19	0.72 (0.53, 0.60)		.72	.64	.18
	20	4.13	0.76 (0.63, 0.70)		.60	.51	.11
	24	4.86	0.60 (0.60, 0.67)		.64	.55	.27
	25	4.67	0.71 (0.74, 0.81)		.45	.34	.12
EFE	3	4.58	0.52 (0.49, 0.53)		.76	.72	.31
	11	4.62	0.54 (0.50, 0.57)		.75	.68	.37
	13	4.61	0.62 (0.63, 0.75)		.62	.44	.22
	14	4.50	0.72 (0.67, 0.75)		.56	.44	.12
	15	4.89	0.67 (0.66, 0.76)		.56	.42	.12
	17	4.44	0.52 (0.41, 0.45)	0.33 (0.26, 0.30)	.77	.71	.30
	21	4.71	0.55 (0.48, 0.49)	0.30 (0.26, 0.28)	.70	.69	.27

Note. Unstandardized coefficients presented followed by standardized coefficients in parentheses (Time 1 followed by Time 2). Factors were scaled by constraining the mean of all factors at time 1 to 0 and the variance to 1.0.

Pattern coefficients. The unstandardized pattern coefficients at Time 1 represent the unit change in the item for every standard deviation change in the factor, whereas the unstandardized pattern coefficients at Time 2 represent the unit change in the item for every one unit change in the factor (although the unstandardized pattern coefficients were

constrained to be equal in the scalar model). The standardized pattern coefficients at both time points represent the standard deviation change in the item for every standard deviation change in the factor. All the pattern coefficients indicate positive relationships to the factor, as anticipated. The standardized pattern coefficients for the negative wording factor are sizeable (ranging from $\lambda = 0.23$ to 0.40). It is important to note, however, that the standardized pattern coefficients for the substantive factors are greater than the negative wording pattern coefficients. Therefore, the negative wording factor is representing a substantial component of variability in the responses to the item, but all items are still more strongly related to the substantive factor, which is desirable.

Error variances. The error variances indicate the amount of unexplained variability in the item and range from .34 to .77. In all cases except for item 1, the error is higher for the item at time 1 than at time 2, which indicates there is more unexplained variability at time 1 than at time 2. Because the SEE exhibited metric invariance, this means that there is more total variability (explained and unexplained) at time 1 than at time 2, because the relationship of the items to the factor is equivalent across time. That is, after accounting for the equivalent relationship between the items and the factor, at time 1 there is more variability left unaccounted for (i.e., error) than at time 2. This different amount of variability at the two time points can be seen by examining the descriptive statistics, as well (Table 2).

Autocorrelations. The autocorrelations (i.e., correlations of the errors across time) varied substantially for items, as well. Item 28 had a minuscule autocorrelation, $r = .05$; however, item 1 had a moderate correlation of its errors across time, $r = .43$. These correlations indicate there is variability in the items not explained by the factors that in some cases have moderate relationships across time. The intercepts for the items (i.e., the value for the item when the level of the factor is 0) ranged from 2.95 to 4.89. This difference in

intercepts indicates that for some items, an individual would score much higher on the SEE with the factor at a level of 0. This difference is important, because it shows that on average students are scoring lower on some subscales (e.g., EPT) than others (e.g., EFE).

V. Discussion

This study sought to contribute to our understanding of the construct of ethnocultural empathy and the Scale of Ethnocultural Empathy (SEE; Wang et al., 2003) more specifically. Longitudinal measurement invariance was examined to determine whether or not the SEE could be used to accurately assess change in ethnocultural empathy over time. Prior to using the SEE to measure change over time, it was important to examine whether the instrument measured ethnocultural empathy in the same manner at two time points. If the measure functions differently at the two time points, differences in observed scores on the scale could reflect differences in measurement rather than differences in ethnocultural empathy, as intended, which would prevent a researcher from drawing accurate conclusions from the scores. Two research questions were posed for this study concerning the SEE and the construct of ethnocultural empathy. These will be examined in turn.

Research Question #1

The first research question concerned whether or not the SEE exhibited longitudinal measurement invariance (i.e., whether the SEE is an invariant measure of ethnocultural empathy over time). Invariance was examined by testing multiple nested models in a SEM framework. The models all exhibited acceptable fit to the data, leading to the support of the scalar model, which is the most parsimonious model, indicating that measurement invariance was indeed upheld. These results provide evidence that those students completing the SEE did not have a different conceptualization of the construct over time (i.e., configural invariance). As well, the findings support the conclusions that the measurement model parameters (e.g., unstandardized pattern coefficients, intercepts) for the SEE were invariant across time. Thus, items are equally salient across time (i.e., metric invariance) and students

with a score of zero on the construct have the same average responses to the items across time (i.e., scalar invariance).

Because configural invariance was upheld, there is evidence that gamma change is not occurring. Given the length of time between testing occasions, it would not have been surprising to find that some college students grew in their understanding of ethnocultural empathy, therefore altering their conceptualization of the construct. Gamma change may have manifested itself by students realizing that there are more facets to the construct than previously believed. In the present study, however, there is evidence that college students do not have a different conceptualization of the construct over time. The lack of gamma change found in the present study should be replicated and examined with other samples.

Because tests for both metric and scalar invariance were upheld, there is evidence that beta change was not occurring. Not only did the factor structure (i.e., conceptualization) remain consistent across time (i.e., no gamma change), but the relationships between the items and the factors did not change over time (i.e., no beta change). If these relationships did change, then these changes could result in observed changes in the scores over time even though the latent level of the construct remained constant. In ruling out the presence of beta change in this study, we have support that alpha change is what is being measured by the change in scores on the SEE. Although the implications of not having beta change in this study, namely that change could be examined, were desirable, there were no strong a priori hypotheses concerning whether or not this would have occurred.

It is also important to note that consistent with previous research, the negative wording factor was also invariant across the time points (Motl & DiStefano, 2002). Because the negative wording factor was invariant, it provides validity evidence for its presence, as it

is systematically appearing in studies over time and relating to the items in the same manner across time. These results are consistent with Motl and DiStefano's (2002) claim that "method effects should be considered of potential substantive importance rather than simply substantively irrelevant noise" (p. 571). Because the SEE exhibits longitudinal measurement invariance, latent means and test-retest coefficients were examined.

Research Question #2

The second research question addressed what the differences were across time at the latent level and was pursued given that the SEE was found to be invariant. We were examining a time period of two years over which most students were required to complete a targeted course addressing empathy towards those of other ethnicities. Given there is some intentional programming in place, we had anticipated an increase over time in ethnocultural empathy. However, this was only found for one factor (EPT). The Empathic Perspective Taking (EPT) factor showed an increase, but not a large one; the effect size for the latent difference was small. This small change is logical as there is not a strong treatment in place, so there is not a compelling reason to anticipate a large growth, although it would not be unwanted. Despite exhibiting an increase in the EPT factor, students' scores were the lowest on this subscale at both time points relative to all other subscales. Moreover, as anticipated, the test-retest coefficient for the EPT factor was quite high ($r = .86$) indicating that most students consistently increased in their ethnocultural empathy (i.e., were maintaining the same rank-order). Researchers in intercultural competency should consider why some substantive factors were more stable than others.

The significant increase and high test-retest coefficient shown in the EPT factor did not hold for the other factors. Consistent with prediction the ACD, EA, and EFE factors demonstrated moderate to high correlations, $r = .81$, .69, and .68, respectively. However,

unlike the EPT factor, no significant difference was demonstrated in their latent means. It is important to note the lack of significant change across time was not due to a ceiling effect (i.e., students were not scoring at the highest point on the scale at time 1). These moderate scores could reflect it is more challenging to have high levels of ethnocultural empathic perspective taking than some of the other facets of ethnocultural empathy (e.g., feeling for others).

The ACD factor had the second highest test-retest correlation (following EPT), indicating the rank-order stayed fairly consistent across time; whereas, EA and EFE showed slightly more change in rank-order of students across time. These lower test-retest coefficients (relative to the other factors) provide evidence that some students are changing differentially on average across time. Bowman (2011) argues this may be occurring due to differential experiences of college students, which may make it difficult to examine change over time in some constructs. ACD showed, on average, a slight increase in scores (not statistically significant), which paired with the high test-retest correlation indicates on average most people stay in the same ordering and increase only slightly. On the other hand, both the EA and EFE factors showed a slight decrease on average over time (albeit nonsignificant) with more variation in individual changes. The moderate test-retest coefficients provide evidence that ethnocultural empathy is indeed an alterable state, as there are individual differences in how students change over time. If the construct of ethnocultural empathy was a trait, we would not expect to see test-retest coefficients as low as they are ($r \approx .70$ to .90).

Limitations

There are some limitations to the results from the study. We selected a two-year time period over which to examine college students' change over time. There were not *large*

differences in college students' levels of ethnocultural empathy in this time period, but some change was exhibited. However, more time might be needed and/or more direct treatment in order to greatly alter students' levels of ethnocultural empathy. Furthermore, we were interested in examining what the changes were over time without strong hypotheses as to whether students should change in two years. That is, given there was no strong treatment in place, we did not have a compelling reason to believe students should increase, although we would not like to see decreases, as empathy is a desirable characteristic. Our results did not show any significant decreases in the subscales, but three of the four subscales did not demonstrate a meaningful average difference in scores over time. Therefore, it would be interesting to consider, more intentionally, a time period or treatment where students would certainly be anticipated to change. It is also important to note that although there was not a significant increase in most subscale scores, students scored at or above the midpoint on the scale (indicating a positive level of ethnocultural empathy) at both time points, which is desirable.

Another limitation is a lack of a diverse population (80.2% of the sample was white). Although the proportions of ethnicities present in the sample mirror that of the scale developers (Wang et al., 2003), it would be useful to have a more diverse population and/or environment to which students were exposed. A limitation, therefore, is the lack of diversity on campus; students may not have been exposed to many people from different ethnic backgrounds, or not realized it if they had.

Future Research

This research opens the door to many other possible studies for the SEE. The present study represents a follow-up to a previous study which altered the scale based on misfit in the structure and an examination of the substantive items (Gerstner, 2011). This

study replicated the good fit of this revised model. However, as the SEE and its measurement model have been altered from what Wang et al. (2003) developed, further validity evidence is needed to ensure that these factors and the scale as a whole are truly addressing ethnocultural empathy as intended. Therefore, future areas of research could relate these factors to external variables to determine if they relate as expected by theory.

Furthermore, although the SEE was determined to be invariant across time, which would typically indicate that observed scores would do an adequate job representing change in the latent construct, observed scores are inappropriate to utilize given the factor structure. This study demonstrated, consistent with previous research, that the negative wording factor is explaining a large amount of variance in the items that are negatively worded on the scale ($\lambda_{stand} \approx .30$). If observed scores are used to measure change over time, they will represent more than a person's level of ethnocultural empathy for all but the EA factor, which has no negatively worded items. That is, observed scores will consist of variability due to the negative wording of items, prohibiting the use of observed scores as "clean" measures of the empathy dimensions. Because of the presence of a negative wording factor, ethnocultural empathy levels at a single time as well as their change over time can *only* be examined in a SEM framework. In other words, a high level of methodological sophistication is needed to appropriately use the SEE.

Because of this, it would also be interesting to attempt to reword the negatively-worded items on the SEE. As mentioned, currently observed scores cannot be accurately used to measure the empathy dimensions on the SEE given the negatively-worded nature of some items. If future research reworded the negatively-worded items to be positively-worded, observed scores may be able to be used. Altering the negatively-worded items

would make the scale more accessible to the general population, because structural equation modeling would not be necessary for scoring purposes.

Finally, it would be interesting to examine change in ethnocultural empathy (and invariance) for a larger time period or across multiple time points. This study examined a restricted time span of two years. It would be interesting to add another time point to examine change perhaps from the start of college to when students graduate. Another area that would be interesting would be examining change from before to after a study abroad program, or another experience directly targeted at adjusting levels of ethnocultural empathy.

An important concluding point involves both a limitation of sorts and an area for future research. The SEE was found to be invariant across time for this sample. However, this does not mean the SEE will always be invariant across time. Our results could be specific to the time or to the sample. Moreover, no invariance across groups was examined. This could be something to explore in future research to examine if the SEE is an equivalent measure of ethnocultural empathy across various ethnic groups, for example.

Conclusions

Ethnocultural empathy is a construct that relates to many different fields (e.g., medical, educational, counseling). Currently, the SEE is the only published instrument purporting to measure this construct. This study showed support for the revised structure of the scale and evidence that the measurement properties of the scale remain invariant across time. However, in terms of practical use of the instrument, observed scores will be wrought with measurement error and be heavily influenced by the negative wording factor. As this is the only measure available and ethnocultural empathy is strongly relevant to these areas, researchers can utilize the SEE, although proceeding with caution. That is, the measurement properties of the SEE held across time in this study; however, when using the

instrument to measure change over time outside of the SEM framework, the scores will include variability associated with systematic variance attributed to the negative wording factor and not solely the construct of interest (i.e., ethnocultural empathy).

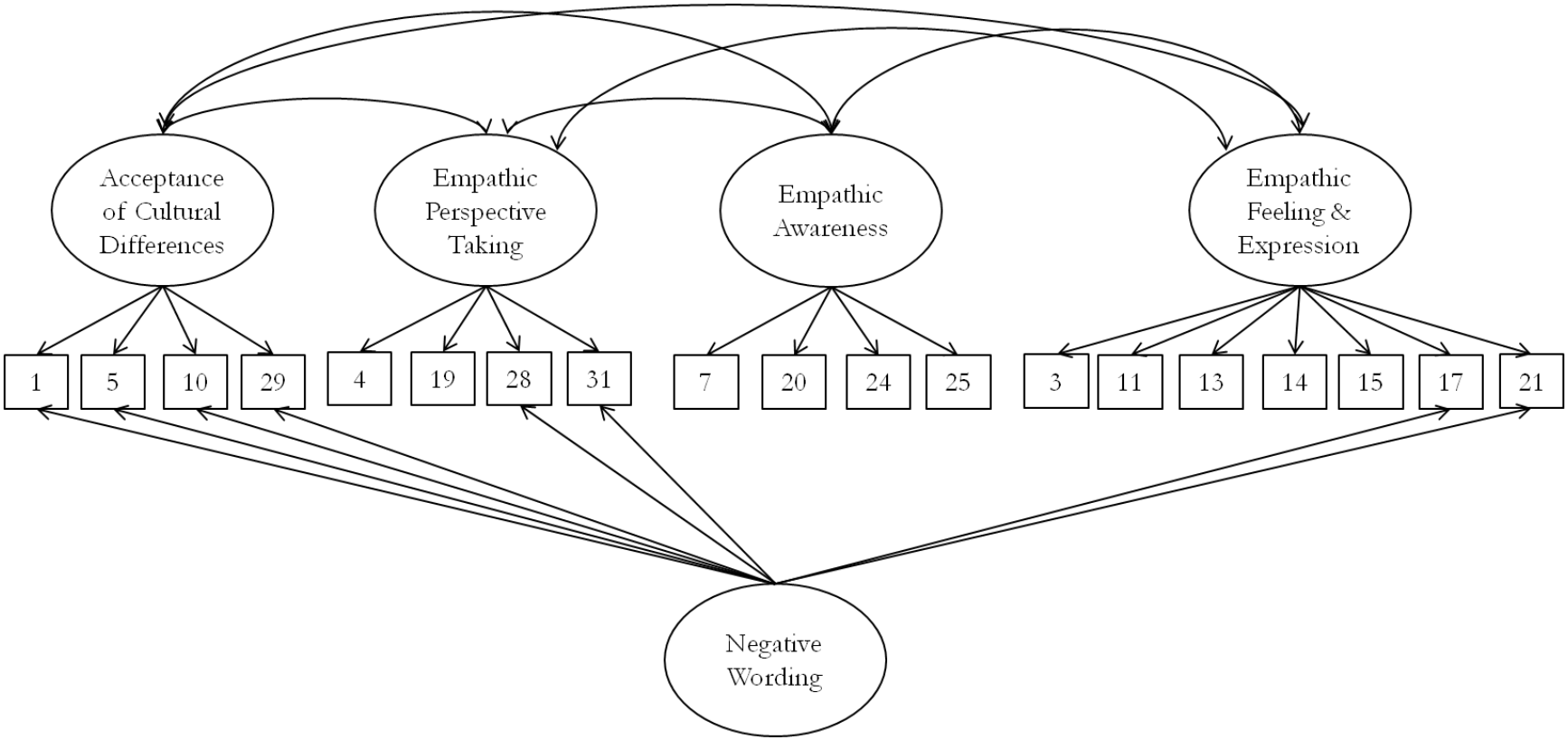

Figure 1. This presents the revised bifactor model (Gerstner, 2011). Error variances are not depicted in the figure for simplicity of presentation. Variances of all factors were set to a value of 1.0 in order to scale the factor. Item numbers correspond to Wang et al.'s (2003) original scale, as presented in Table 1 (p. 225).

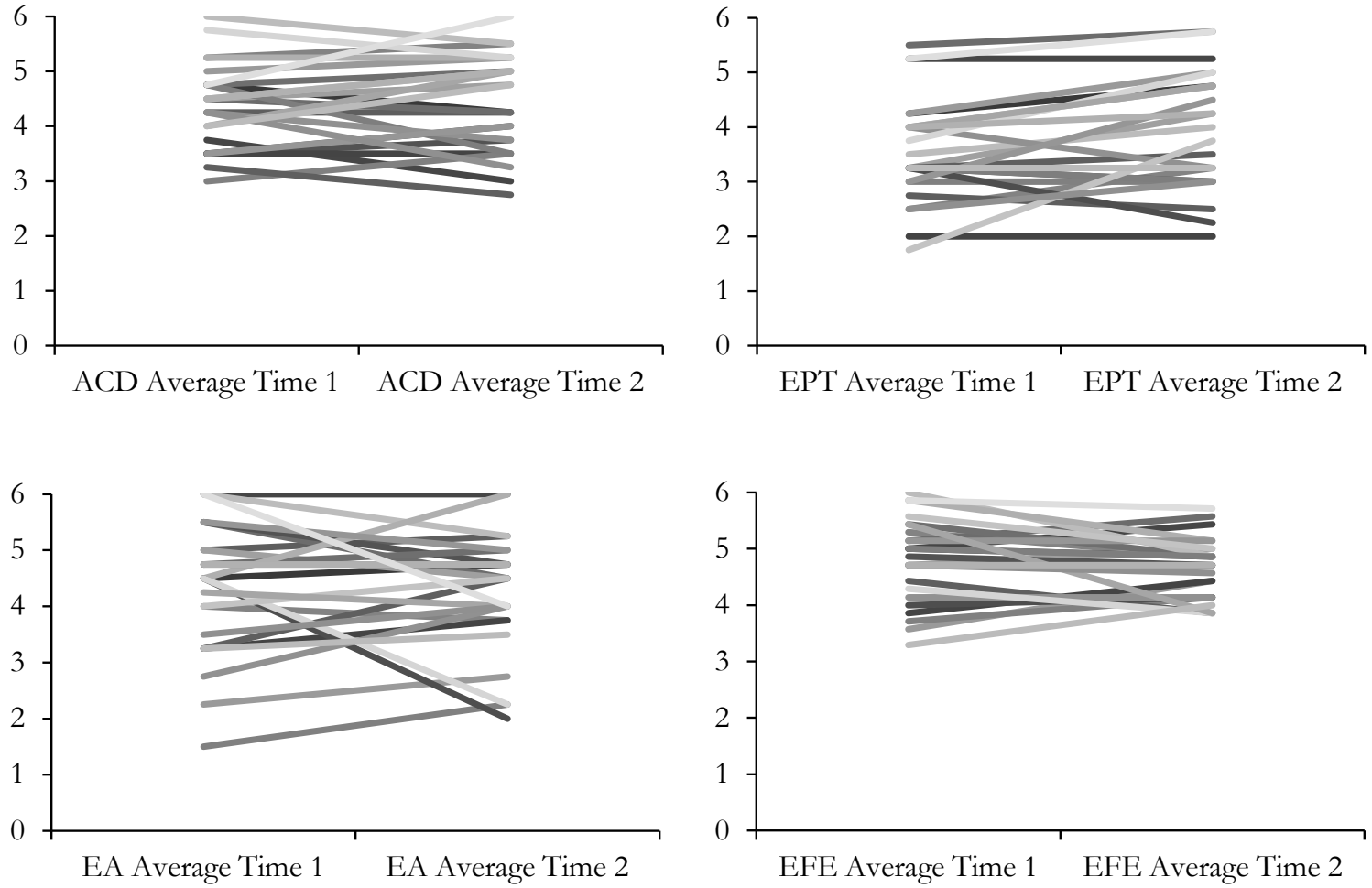

Figure 2. Examining change over time in a random sample of 25 students.

Appendix

The Revised Scale of Ethnocultural Empathy[7]

1. I feel annoyed when people do not speak standard English.*
3. I am touched by movies or books about discrimination issues faced by racial or ethnic groups other than my own.
4. I know what it feels like to be the only person of a certain race or ethnicity in a group of people.
5. I get impatient when communicating with people from other racial or ethnic backgrounds, regardless of how well they speak English.*
7. I am aware of institutional barriers (e.g., restricted opportunities for job promotion) that discriminate against racial or ethnic groups other than my own.
10. I feel irritated when people of different racial or ethnic backgrounds speak their language around me.*
11. When I know my friends are treated unfairly because of their racial or ethnic backgrounds, I speak up for them.
13. When I interact with people from other racial or ethnic backgrounds, I show my appreciation of their cultural norms.
14. I feel supportive of people of other racial or ethnic groups, if I feel they are being taken advantage of.
15. I get disturbed when other people experience misfortunes due to their racial or ethnic backgrounds.
17. I am not likely to participate in events that promote equal rights for people of all racial and ethnic backgrounds.*
19. It is easy for me to understand what it would feel like to be a person of another racial or ethnic background other than my own.
20. I can see how other racial or ethnic groups are systematically oppressed in our society.
21. I don't care if people make racist statements against other racial or ethnic groups.*
24. I recognize that the media often portrays people based on racial or ethnic stereotypes.
25. I am aware of how society differentially treats racial or ethnic groups other than my own.
28. It is difficult for me to put myself in the shoes of someone who is racially and/or ethnically different from me.*
29. I feel uncomfortable when I am around a significant number of people who are racially/ethnically different than me.*
31. It is difficult for me to relate to stories in which people talk about racial or ethnic discrimination they experience in their day to day lives.*

[7] All items are reprinted from Wang et al. (2003), p. 225. The items presented are those retained following the CFA study by Gerstner (2011). Asterisks denote the item is reverse-scored.

References

Bandalos, D. L., & Finney, S. J. (2010). Factor analysis: Confirmatory and exploratory. In G. R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 93-114). New York: Routledge.

Bandalos, D. L., & Finney, S. J. (2001). Item parceling issues in structural equation modeling. In G. Marcoulides & R. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 269-296). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Batson, C. D. (1990). How social an animal? The human capacity for caring. *American Psychologist, 45*, 336-346.

Bellini, L. M., & Shea, J. A. (2005). Mood change and empathy decline persist during three years of internal medicine training. *Academic Medicine, 80*, 164-167.

Benson, J. (1998). Developing a strong program of construct validation: A test anxiety example. *Educational Measurement: Issues and Practice,17*, 10-17.

Bentler, P. M., & Wu, E. J. C. (2003). *EQS for Windows User's Guide*. Encino, CA: Multivariate Software, Inc.

Bowman, N. A. (2011). Validity of college self-reported gains at diverse institutions. *Educational Researcher, 40*, 22-24.

Brown, A. R., & Finney, S. J. (2011). Low-stakes testing and psychological reactance: Using the Hong Psychological Reactance Scale to better understand compliant and non-compliant examinees. *International Journal of Testing, 11*, 248-270.

Brown, E. L. (2004). What precipitates change in cultural diversity awareness during a multicultural course: The message or the method? *Journal of Teacher Education, 55*(4), 325-340.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research* (1st ed.). New York: Guilford Press.

Chan, F. H., Blalock, K., Cardoso, E., Steven, P. R., & Eun-Jeong, L. (2007). *Confirmatory factory analysis of the Ethnocultural Empathy Scale*. Presented at the annual meeting of the American Psychological Association, San Francisco, CA.

Chen, D., Lew, R., Hershman, W., & Orlander, J. (2007). A cross-sectional measurement of medical student empathy. *Journal of General Internal Medicine, 22*, 1434-8.

Chung, R. C., & Bemak, F. (2002). The relationship of cultural and empathy in cross-cultural counseling. *Journal of Counseling & Development, 80*, 154-159.

Cockrell, K. S., Placier, P. L., Cockrell, D. H., & Middleton, J. N. (1999). Coming to terms with "diversity" and "multiculturalism" in teacher education: Learning about our students, changing our practice. *Teaching and Teacher Education, 15*, 351-366.

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology, 44*, 113-126.

Davis, M. H., & Franzoi, S. L. (1991). Stability and change in adolescent self-consciousness and empathy. *Journal of Research in Personality, 25*, 70-87.

DeCarlo, L. (1997). On the meaning and use of kurtosis. *Psychological Methods, 2*, 292 – 307.

Duan, C., & Hill, C. E. (1996). The current state of empathy research. *Journal of Counseling Psychology, 43*, 261-274.

Eisenberg, N., & Miller, P. A. (1987). The relation of empathy to prosocial and related behaviors. *Psychological Bulletin, 101*, 91-119.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*, 272-299.

Forster, N. (2006). Expatriates and the impact of cross-cultural training. *Human Resource Management Journal, 10*, 63-78.

Gay, G., & Kirkland, K. (2003). Developing cultural critical consciousness and self-reflection in preservice teacher education. *Theory Into Practice, 42*, 181-187.

Gerstner, J. J. (2011). *Evaluating the psychometric properties of the Scale of Ethnocultural Empathy: A confirmatory factor analytic approach*. Manuscript in preparation.

Gerstner, J. J., & Pastor, D. A. (2011b, October). *A second look at the structural validity of the Scale of Ethnocultural Empathy: A replication*. Paper presented at the annual meeting of the Northeastern Educational Research Association, Rocky Hill, CT.

Gerstner, J. J., & Pastor, D. A. (2011a, May). *A factor analytic study of the Scale of Ethnocultural Empathy*. Poster presented at the annual meeting of the Association for Psychological Science, Washington, D.C.

Gladstein, G. A. (1983). Understanding empathy: Integrating counseling, developmental and social psychology perspectives. *Journal of Counseling Psychology, 30*, 467-482.

Golembiewski, R. T., Billingsley, K., & Yeager, S. (1976). Measuring change and persistence in human affairs: Types of change generated by OD designs. *Journal of Applied Behavioral Science, 12,* 133-157.

Green, S. B., & Yang, Y. (2009). Commentary of coefficient alpha: A cautionary tale. *Psychometrika, 7*(1), 121-135.

Grief, E. B., & Hogan, R. (1973). The theory and measurement of empathy. *Journal of Counseling Psychology, 30*, 280-284.

Hays, P. A. (1996). Addressing the complexities of culture and gender in counseling. *Journal of Counseling & Development, 74*, 332-338.

Hogan, R. (1969). Development of an empathy scale. *Journal of Counseling and Clinical Psychology, 33*, 307-316.

Hojat, M., Mangione, S., Kane, G. C., & Gonnella, J. S. (2005). Relationships between scores of the Jefferson Scale of Physician Empathy (JSPE) and the Interpersonal Reactivity Index (IRI). *Medical Teacher, 27*, 625-628.

Hojat, M., Mangione, S., Nasca, T. J., Cohen, M. J., Gonnella, J. S., Erdmann, J. B., Veloski, J., & Magee, M. (2001). The Jefferson Scale of Physician Empathy: Development and preliminary psychometric data. *Educational and Psychological Measurement, 61*, 349-365.

Hojat, M., Mangione, S., Nasca, T. J., Rattner, S., Erdmann, J. B., Gonnella, J. S., & Magee, M. (2004). An empirical study of decline in empathy in medical school. *Medical Education, 38*, 934-941.

Hojat, M., Vergare, M. J., Maxwell, K., Brainard, G., Herrine, S. K., Isenberg, G. A., Veloski, J., & Gonnella, J. S. (2009). The devil is in the third year: A longitudinal study of erosion of empathy in medical school. *Academic Medicine, 84*, 1182-1191.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

Ivey, A. E., Ivey, M. B., & Simek-Morgan, L. (1993). The empathic attitude: Individual, family and culture. In A. Ivey, M. Ivey, & L. Simek-Morgan (Eds.), *Counseling and psychotherapy: A multicultural perspective* (3rd ed., pp. 21-44). Boston: Allyn & Bacon.

Johnston, M. M., & Finney, S. J. (2010). Measuring basic needs satisfaction: Evaluating previous research and conducting new psychometric evaluations of the Basic Needs Satisfaction in General Scale. *Contemporary Educational Psychology, 35,* 280-296.

Jolliffe, D., & Farrington, D. P. (2004). Empathy and offending: A systematic review and meta-analysis. *Aggression and Violent Behavior, 9*, 441-476.

Jolliffe, D., & Farrington, D. P. (2006). Development and validation of the Basic Empathy Scale. *Journal of Adolescence, 29,* 589-611.

Kitsantas, A. (2004). Studying abroad: The role of college students' goals on the development of cross-cultural skills and global understanding. *College Student Journal, 38*, 441-452.

Little, T. D., Card, N. A., Slegers, D. W., & Ledford, E. C. (2007). Representing contextual effects in multiple-group MACS models. In T. D. Little, J. A. Bovaird & N. A Card (Eds.), *Modeling contextual effects in longitudinal studies* (pp. 121-147). Mahwah, N. J.: Lawrence Erlbaum Associates.

Long, E. C., Angera, J. T., Carter, S. J., Nakamoto, M., & Kalso, M. (1999). Understanding the one you love: A longitudinal assessment of an empathy training program for couples in romantic relationships. *Family Relations, 48*, 245-242.

Loudin, J. L., Loukas, A., & Robinson, S. (2003). Relational aggression in college students: Examining the roles of social anxiety and empathy. *Aggressive Behavior, 29*, 430-9.

Mead, G. H. (1934). *Mind, self and society*. Chicago: University of Chicago Press.

Mehrabian, A., & Epstein, N. (1972). A measure of emotional empathy. *Journal of Personality, 40*, 525-543.

Miville, M. L., Gelso, C. J., Pannu, R., Liu, W., Touradji, P., Holloway, P., & Fuertes, J. (1999). Appreciating similarities and valuing differences: The Miville-Guzman Universality-Diversity Scale. *Journal of Counseling Psychology, 46*, 291-307.

Motl, R. W., & DiStefano, C. (2002). Longitudinal invariance of self-esteem and method effects associated with negatively worded items. *Structural Equation Modeling, 9*, 562-578.

Munroe, A., & Pearson, C. (2006). The Munroe Multicultural Attitude Scale Questionnaire: A new instrument for multicultural studies. *Educational and Psychological Measurement, 66*, 819-834.

Newton, B. W., Barber, L., Clardy, J., Cleveland, E., & O'Sullivan, P. (2008). Is there a hardening of heart during medical school? *Academic Medicine, 83*, 244-9.

Nguyen, P. (2003). *Social context, ethnic identity, and ethnocultural empathy* (Master's thesis). Retrieved from http://csus-dspace.calstate.edu/handle/10211.9/84

Olsson, U. H., Foss, T., Troye, S. V., & Howell, R. D. (2000). The performance of ML, GLS, & WLS estimation in structural equation modeling under conditions of misspecification and nonnormality. *Structural Equation Modeling, 7*, 557-595.

Parson, E. R. (1993). Ethnotherapeutic empathy (EthE)–Part II: Techniques in interpersonal cognition and vicarious experiencing across cultures. *Journal of Contemporary Psychology, 23*, 171-182.

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology, 46*, 598-609.

Pratto, F., Sidanius, J., Stallworth, L.M., & Malle, B.F. (1994). Social dominance orientation: A personality variable predicting social and political attitudes. *Journal of Personality and Social Psychology, 67*, 741-763.

Rasoal, C. (2009). *Ethnocultural empathy: Measurement, psychometric properties, and differences between students in health care education programmes* (Doctoral thesis, Linköping University, Linköping, Sweden). Retrieved from http://liu.diva-portal.org/smash/get/diva2:278188/FULLTEXT01

Rasoal, C., Eklund, J., & Hansen, E. M. (2011). Toward a conceptualization of ethnocultural empathy. *Journal of Social, Evolutionary, and Cultural Psychology, 5*, 1-13.

Rensvold, R. B., & Cheung, G. W. (2001). Testing for metric invariance using structural equation models: Solving the standardization problem. In C. A. Schriesheim & L. L. Neider (Eds.), *Research in Management* (pp. 25-50). Greenwich, CT: Information Age Publishing.

Ridley, C. R., & Lingle, D. W. (1996). Cultural empathy in multicultural counseling: A multidimensional process model. In Pedersen, P. B., Draguns, J. G., Lonner, W. J., & Trimble, J. E. (Eds.), *Counseling across cultures*(4th ed.), (pp. 21-46). Thousand Oaks, CA: Sage Publications, Inc.

Riordan, C. M., Richardson, H. A., Schaffer, B. S., & Vandenberg, R. J. (2001). Alpha, beta, and gamma change: A review of past research with recommendations for new directions. In C. A. Schriesheim & L. L. Neider (Eds.), *Research in Management* (pp. 51-81). Greenwich, CT: Information Age Publishing.

Rogers, C. R. (1975). Empathic: An unappreciated way of being. *The Counseling Psychologist, 5*(2), 2-10.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors on covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent Variables Analysis* (pp. 399-419). Thousand Oaks, CA: Sage.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistics for moment structure analysis. *Psychometrika, 66*(4), 507-514.

Scott, N. E., & Borodovsky, L. G. (1990). Effective use of cultural role taking. *Professional Psychology: Research and Practice, 21*, 167-170.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*, 147-177.

Steenkamp, J. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research, 25*, 78-90.

Struch, N., & Schwartz, S.H. (1989). Intergroup aggression: Its predictors and distinctness from in-group bias. *Journal of Personality and Social Psychology, 56*, 364-373.

Sue, S., Fujino, D. C., Hu, L., Takeuchi, D. T., & Zane, N. W. S. (1991). Community mental health services for ethnic minority groups: A test of the cultural responsiveness hypothesis. *Journal of Consulting and Clinical Psychology, 59*, 533-540.

Sue, S., & Zane, N. (1987). The role of culture and cultural techniques in psychotherapy: A critique and reformulation. *American Psychologist, 42*(1), 37-45.

Titchener, E. (1924). *A textbook of psychology*. New York: Macmillan.

U.S. Census Bureau. (2010). 2010 Census Results [Demographic map]. Retrieved from http://2010.census.gov/2010census/data/index.php

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4-69.

Wang, Y., Davidson, M. M., Yakushko, O.F., Bielstein, H. B., Tan, J. A., & Bleier, J.K. (2003). The scale of ethnocultural empathy: Development, validation, and reliability. *Journal of Counseling Psychology, 50*, 221-234.

Yu, C., & Muthén, B. (2002, April). *Evaluation of model fit indices for latent variable models with categorical and continuous outcomes*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.