2018

Measuring Undergraduates' Global Perspective Development: Examining the Construct and Cross-Cultural Validity of the Global Perspective Inventory Across Ethnoracial Groups

Lisa Davidson

LOYOLA UNIVERSITY CHICAGO

MEASURING UNDERGRADUATES' GLOBAL PERSPECTIVE DEVELOPMENT: EXAMINING THE CONSTRUCT AND CROSS-CULTURAL VALIDITY OF THE GLOBAL PERSPECTIVE INVENTORY ACROSS ETHNORACIAL GROUPS

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE GRADUATE SCHOOL

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

PROGRAM IN HIGHER EDUCATION

BY

LISA M. DAVIDSON

CHICAGO, IL

AUGUST 2018

ACKNOWLEDGEMENTS

My experience as a doctoral student transformed my thinking—*and person*—in profound ways. I have many to thank for their role in this. I have completed this project and this experience with *and* because of you.

To my advisor, Dr. Mark Engberg, I am incredibly thankful to have worked with you over these five years. I have learned so much from you and am so appreciative of the many opportunities you have extended to me. I have appreciated your belief in me, namely during times when I did not believe in myself. I have also appreciated that you continuously model excellence and drive in so many ways. You have helped me understand myself as a scholar by encouraging me to wrestle with big ideas and questions. I am a better teacher, practitioner, and scholar because of you. And I look forward to our continued thought partnership.

Next, I want to thank my dissertation committee, Drs. Mark Engberg, Ken Fujimoto, and Robert Reason, for making this project possible. I have learned from each of you and am incredibly appreciative of your guidance, support, and expertise. You three challenged me to think about this work in ways that enabled more learning than I thought possible. Truly, thank you. I also want to thank Dr. Joshua Mitchell, Project Manager for the Global Perspective Inventory, for working with me at the outset of this study and for assistance in providing the survey data used for this study. I want to thank the Research Institute for Studies in Education at Iowa State University for its support of this project and interest in using the study's findings.

I can only begin to thank my Loyola Ph.D. cohort for their role in this journey. Natasha Turman, Ester Sihite, Mark Torrez, and Jim Neumeister, I cannot imagine this experience without each of you. From the very beginning, our friendship and unconditional support of one another has meant everything. We celebrated the highs and processed the uncertain and difficult times associated with this experience. Joining you four as learners in this process transformed how I understood our field, my research, and my person. Megan Segoshi, my friend (and adopted cohort mate!), I also thank you for sharing your journey with me and for all the space you created for me to reflect on my experiences and ask important questions. Your friendship has meant so much. I love you all.

I also want to acknowledge the following individuals who taught my courses in LUC's higher education program. Thank you, Drs. Mark Engberg, John Dugan, OiYan Poon, Sunny Nakae, and Terry Williams, for cultivating profound learning experiences, both in and out of the classroom. I also want to particularly acknowledge two other LUC faculty, Dr. Fred Bryant, psychologist, and Dr. Ken Fujimoto, research methodologist, whose support and expertise I relied on in and out of the classroom. I have been so appreciative of your methodological expertise. And I am especially grateful to have received training from you both. Your empowering approaches help your students believe they can contribute meaningfully. I hope you both never forget the importance of that.

I also want to thank Dr. Bridget Turner Kelly for her leadership of our higher education program during a very transitional time. In particular, I thank you, Bridget, for the opportunity to teach within our program and for the personal *and* logistical support during the last leg of this project. Drs. Darren Pierre, Blanca Torres-Olave, and Eilene Edejer, while we did not meet in the classroom, I am so glad our paths eventually crossed at Loyola. Your support and guidance

as colleagues in the School of Education have meant so much to me. Thank you for your support, collaboration, interest in my work, and uplifting conversations.

I have been so fortunate to have incredibly brilliant and supportive colleagues who—at various stages during this project—helped me process the ups and downs, listened to my ideas, and created spaces for me to reflect on what I was learning. Dr. J.T. Snipes, thank you for caring about both my research *and* my person during this process. I have learned so much from you and have appreciated that our conversations always encourage me to reflect deeply on my research and our field. Dr. Janett Cordovés, thank you for sharing your own dissertation journey with me and for your support during mine. You instilled so much hope and always reminded me to find time to love and laugh. Dr. Ben Correia-Harker, my fellow SEM-er, thank you for talking with me about important nuances related to my study, your overall understanding and encouragement during this process, and your support of my next professional steps. Kathleen, Stacey, Kayla, Henry, and Kris, thank you all for valuing the direct application of my work within institutional research and for your support, insights, and interest in my work along the way.

I am grateful to my family as I conclude this process. I want to thank my brother-in-law, Matthew, for taking such an interest in the *process* of my work as it unfolded. Even during the busiest of times, you helped me reflect on my work's purpose and talked with me about the deeply personal aspects of putting one's work out there for others. To my brother and fellow Ph.D. student, A.J., your brilliance continues to inspire me; thank you for reminding me to always strive for greatness. To my dad, Glenn, thank you for your questions and encouragement along the way. To my mom, Carol, an educator herself who instilled in me a love of learning, thank you. And thank you, Jim, Malinda, Paige, and David for your support throughout this

journey as well. I also want to acknowledge Frankie, my canine companion, who lovingly kept me company until the very end of this project. I will always remember your unconditional love during my doctoral studies.

Finally, I am eternally grateful to my partner, Mark, for all that he has done and who he has been throughout this experience. You helped me believe in my ability to pursue this degree before I began my studies. During the most difficult experiences, you lovingly listened and reminded me of my purpose. You insisted on celebrating all of my successes (big and small). You understood what was, much of the time, the totalizing nature of this experience. Thank you for making this journey—and all the other ones—so meaningful. I love you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

This study examined the construct and cross-cultural validity of the Global Perspective Inventory's (GPI's) 32 global perspective development items, which measure six related dimensions of development spanning cognitive, intrapersonal, and interpersonal domains. The study's sample consisted of 7,092 undergraduates who completed the GPI General Form between 2015-2017. The GPI's hypothesized hierarchical factor structure was used for all confirmatory factor analyses and multiple-group confirmatory factor analyses for this study. The data from the GPI were ordinal in nature, presenting important considerations. This study makes several contributions to survey validation efforts. First, it provides a nuanced definition of the concept of validity and presents the processes of cross-cultural and construct validation in accessible ways. The study carefully outlines analytical procedures for measurement invariance testing using hierarchical factor structures and ordinal data. Second, examining the GPI's hierarchical factor structure as well as convergent and discriminant validity of its six developmental scales revealed important evidence. The measurement invariance results suggest that the GPI's developmental constructs are theorized, understood, and measured equivalently across African American/Black, Asian/Pacific Islander/Native Hawaiian, Hispanic, and white undergraduates. The construct validity evidence illuminated specific opportunities for both scale- and item-level refinement opportunities. This study provides a roadmap for informed refinement and subsequent validation of the GPI and discusses future inquiry related to this instrument and measurement issues in cross-cultural survey research more generally.

CHAPTER ONE

INTRODUCTION

Postsecondary education in the United States (U.S.) has long stood as a social institution shaped by society's needs (Gumport, 2000). Indeed, the notion of a public or collective good has served as a central element of the charter between postsecondary education and the societies in which this is situated (Kezar, Chambers, & Burkhardt, 2004). The emergence of particular postsecondary institutional types (i.e., land-grant institutions, community colleges), the GI bill, federal investment in financial aid, and the academy's knowledge production and dissemination functions illustrate ways in which U.S. higher education has responded to society's needs during various eras (National Center for Public Policy and Higher Education, 2008).

In the 21st century, globalization (i.e., the interaction between the world's social, political, economic, business, and physical domains; Riveras & Harrison, 2016) presents new realities that U.S. postsecondary education must address. For instance, given the challenges embedded in such an interrelated society (e.g., ensuring that the benefits of globalization are equally distributed, that global competition does not erode individual and environmental rights; Carnegie Council for Ethics in International Affairs, 2017), a postsecondary education must help its students answer in what ways they access information about and consider diverse global issues. How do students understand themselves in relation to an interconnected globe? In what ways do they interact across difference and consider the needs of others? Both the internationalization and diversification of postsecondary education complicate examining these questions. Posing

such questions necessitates *inclusive* theoretical and methodological considerations. To what extent is postsecondary education measuring the type of learning necessary in today's globalized society, and to what extent are these methods cross-culturally appropriate?

Internationalization within U.S. Postsecondary Education

Our increasingly interconnected global contexts serve as the most influential drivers of the internationalization of postsecondary education around the world (Altbach & Knight, 2007; International Association of Universities [IAU], 2014). The American Council on Education (ACE, 2012) defines internationalization as institutional-level efforts that cultivate the competencies needed in a globalized society by "incorporating global perspectives into teaching, learning, and research; building international and intercultural competence among students, faculty, and staff; and establishing relationships and collaborations with people and institutions abroad" (p. 3). Internationalization as a comprehensive strategy in postsecondary education has only emerged within the last two decades (de Wit, 2013) and has since fueled much of the international focus and strategies evident on U.S. college campuses (Whitehead, 2016). Internationalization has become a nearly ubiquitous theme. From its most recent Global Survey administration, the IAU (2014) reports that of the 1,336 postsecondary institutions surveyed—spanning 131 countries in every world region—75% already had or were developing institutional-level internationalization strategies or policies.

Over the last two decades, several associations (e.g., the Association of American Colleges and Universities (AAC&U), ACE and its Center for Internationalization and Global Engagement, and Council for the Advancement of Standards in Higher Education) have examined the U.S. postsecondary education system relative to the globalized contexts in which it

is situated. These efforts have sought to shape policies and practice that underscore the import of preparing students to engage in an increasingly globalized society. Stemming from these efforts, recommendations have foregrounded the development of students' intercultural competencies and global perspective, international programs, and curricular and co-curricular efforts focused on global citizenship and engagement (Whitehead, 2016). For instance, ACE's Center for Internationalization and Global Engagement advanced a model for comprehensive internationalization that explains where such internationalization efforts unfold within an institution (ACE, 2012). These areas span structural - (i.e., institutional commitment, international recruitment, and global partnerships), staff and faculty - (i.e., curricular and co-curricular learning outcomes, faculty development and research), and student-level (i.e., study abroad and other programmatic participation) components (ACE, 2012). These efforts have also manifested in locating global learning centrally within a majority of institutional strategic plans (ACE, 2012) and mission statements (Whitehead, 2016) and the proliferation of global student enrollments (Institute of International Education [IIE], 2012), study-away programs (Sobania, 2015; Sobania & Braskamp, 2009), service learning (Thomson, Smith-Tolken, Naidoo, & Bringle, 2011), and cross-border institutional partnerships (Olcott, 2009). The advancement of undergraduate learning outcomes including local and global civic engagement, intercultural knowledge, humanitarianism, and global learning (AAC&U, 2007; Council for the Advancement of Standards in Higher Education [CAS], 2015) and the assessment of such learning (ACE, 2012; Green, 2013) also reflect this focus.

Increasingly Diverse Postsecondary Educational Contexts

Assumed within the aforementioned outputs of internationalization are intercultural interaction and competencies (Stier, 2006). Diversifying postsecondary educational contexts to allow for such interaction across difference has become a priority for many institutions (Brown, 2004; Smith, 2009) and an outcome of an ample body of social science research that has demonstrated the educational benefits of diversity within higher education (e.g., Cabrera, Nora, Terenzini, Pascarella, & Hagedorn, 1999; Gurin, Dey, Hurtado, & Gurin, 2002). The racial and ethnic composition of U.S. college campuses has diversified over time. The U.S. Department of Education (USDE, 2015a) reports that of all undergraduate enrollments in degree-granting U.S. postsecondary institutions, students identify as white (55%), Hispanic (17%), black (14%), Asian/Pacific Islander (6%), multiracial (3%), and American Indian/Alaska Native (< 1%). While on an aggregate level, U.S. postsecondary education is still predominantly white, enrollments by students of color have increased over the last 40 years (Aud, Fox, & KewalRamani, 2010).

Additionally, a record number of international students attended U.S. postsecondary institutions in 2015-16; over one million international students enrolled (or approximately 5% of all U.S. postsecondary enrollments; IIE, 2016). These international students travel largely from China (31%), India (14%), South Korea (7%), and Saudi Arabia (6%). Over a third (42%, or nearly 371,000 students) of these international students are undergraduates, while the rest are pursuing graduate or professional studies. These demographic shifts illuminate the need to understand student engagement, learning, and development across different ethnoracial groups *and* to ensure that any such study is theorized in ways that are cross-culturally relevant.

Diversity involves myriad identities. However, the centrality of race in the organization of U.S. society and its postsecondary educational institutions underscores the import of examining students' experiences relative to the racialized contexts in which these unfold (Harper, 2012). In particular, these racialized contexts influence the study of educational outcomes involving interaction across cultural differences, such as those related to institutions' internationalization efforts. For instance, ethnoracially minoritized groups' location within predominantly white postsecondary settings necessitates that they regularly negotiate intercultural difference in interfacing with dominant groups, ideology, and normative practices. The novelty of such interracial interactions, emotional expense of such interactions (Sorenson, Nagda, Gurin, & Maxwell, 2009; Stephan & Stephan, 2001), and differential outcomes related to interracial interactions (Harper, 2012) situate minoritized and majority students differently relative to negotiating difference.

Given these findings, it may be expected that ethnoracially minoritized and white students understand and experience interacting across difference, intercultural knowledge, sensitivity toward difference, and their own ethnoracial identities differently. These expectations are especially salient relative to conceptions of global learning given that this type of learning and development often requires intercultural exchange. If curricular and co-curricular interactions across difference are understood differently across ethnoracial groups, implications abound related to examining the dimensions of development required for global learning.

Global Learning as an Educational Priority

While the aforementioned internationalization efforts within postsecondary education span various levels of the academy, the measurement of students' global learning is the focus of

the present study. Global learning, in its broadest sense, is an expected output of many internationalization efforts; this concept represents "the full scope and substance of engagement with learning in and about the world" (Whitehead, 2015, p. 9). However, the IAU (2014) statement on internationalization calls postsecondary institutions to continuously examine underlying values and intentions associated with their internationalization efforts, explicitly naming intercultural learning as a dimension to examine. As institution-wide internationalization efforts increasingly unfold, postsecondary education must align the type of global learning expected of students with educational opportunities that cultivate such development (AAC&U, 2007). Further, institutions must move beyond the identification of particular curricular and co-curricular global learning opportunities and also assess the extent to which these learning outcomes are actualized (Green, 2013). However, the broad concept of global learning is understood differently among various postsecondary contexts, presenting conceptual and methodological challenges for institutions committed to developing global citizens (Green, 2013) and ensuring that such preparation is conceptualized inclusively (IAU, 2014).

Dimensions of Global Learning

Conceptually, global learning spans an array of developmental dimensions. For instance, the AAC&U (2007) advanced a multidimensional definition of global learning that involves conceptually distinct dimensions of students' learning and development (i.e., global self-awareness, perspective taking, cultural diversity, personal and social responsibility, and global systems; Whitehead, 2016). CAS (2015) articulates humanitarianism and civic engagement as a key developmental domain that is comprised of four distinct global learning dimensions (i.e., understanding and appreciation of cultural and human differences, global perspective, social

responsibility, and sense of civic responsibility). Related to these concepts, the present study's focus is on the global perspective development construct measured by the Global Perspective Inventory (GPI; Iowa State University, 2015). Based on the instrument's conceptualization, the development of a global perspective is defined to include "the acquisition of knowledge, attitudes, and skills important to intercultural communication, as well as the development of more complex epistemological processes, identities, and interpersonal relations" (Engberg & Fox, 2011, pp. 86-87). Table 1 includes the GPI's six global perspective development scales and items.

Table 1. The GPI's 32 Global Perspective Development Survey Items

SCALE: COGNITIVE KNOWING	**CR = .59[a]**
COGEP01 - When I notice cultural differences, my culture tends to have the better approach.(r)	
COGEP06 - Some people have culture and others do not.(r)	
COGEP07 - In different settings what is right and wrong is simple to determine.(r)	
COGEP16 - I take into account different perspectives before drawing conclusions about the world around me.	
COGEP19 - I consider different cultural perspectives when evaluating global problems.	
COGEP20 - I rely primarily on authorities to determine what is true in the world.(r)	
COGEP30 - I rarely question what I have been taught about the world around me.(r)	
SCALE: COGNITIVE KNOWLEDGE	**CR = .81**
COGKNW08 - I am informed of current issues that impact international relations.	
COGKNW13 - I understand the reasons and causes of conflict among nations of different cultures.	
COGKNW17 - I understand how various cultures of this world interact socially.	
COGKNW21 - I know how to analyze the basic characteristics of a culture.	
COGKNW27- I can discuss cultural differences from an informed perspective.	
SCALE: INTRAPERSONAL IDENTITY	**CR = .80**
IDENT02 - I have a definite purpose in my life.	
IDENT03 - I can explain my own personal values to people who are different from me.	
IDENT09 - I know who I am as a person.	

IDENT12 - I am willing to defend my views when they differ from others.
IDENT18 - I put my beliefs into action by standing up for my principles.
IDENT 28 - I am developing a meaningful philosophy of life.

SCALE: INTRAPERSONAL AFFECT **CR = .80**

AFFECT22 - I am sensitive to those who are discriminated against.
AFFECT23 - I do not feel threatened emotionally when presented with multiple perspectives.
AFFECT25 - I am accepting of people with different religious and spiritual traditions.
AFFECT31 - I enjoy when my friends from other cultures teach me about our cultural differences.
AFFECT33 - I am open to people who strive to live lives very different from my own life style.

SCALE: INTERPERSONAL SOCIAL RESPONSIBILITY **CR = .78**

SOCRES05 - I think of my life in terms of giving back to society.
SOCRES14 - I work for the rights of others.
SOCRES26 - I put the needs of others above my own personal wants.
SOCRES32 - I consciously behave in terms of making a difference.
SOCRES34 - Volunteering is not an important priority in my life.(r)

SCALE: INTERPERSONAL SOCIAL INTERACTION **CR = .77**

SOCINT04 - Most of my friends are from my own ethnic background.(r)
SOCINT24 - I frequently interact with people from a race/ethnic group different from my own.
SOCINT29 - I intentionally involve people from many cultural backgrounds in my life.
SOCINT35 - I frequently interact with people from a country different from my own.

Note. (r) Indicates a reverse-worded item; these items were recoded so that a high mean score signifies more positive levels related to the specific dimension of development. [a] Composite reliability (CR) is the SEM approach for estimating scale reliability using the items' standardized factor loadings, error variance, and item R^2 (Raykov, 1997).

In alignment with the IAU's (2014) charge to critically examine conceptions of intercultural learning, emerging work has explicitly emphasized the uneven distribution of power embedded in conceptualizations of global learning. Andreotti (2010), for instance, advanced the concept of critical global citizenship, which "requires an acknowledgement that contemporary societies are complex, diverse, changing, uncertain and deeply unequal;" such a reality requires an educational focus on decolonization so that learners can understand the nature of such global inequities and "tools to negotiate a future that could be 'otherwise'" (p. 234). Internationalization

can drive ideological convergence (i.e., imposing "cultural conformity" or blurring once-held geographical identities; Stier, 2006, p. 4). Roberts and Komives (2016) argue that the west and north influences on international postsecondary educational efforts illustrate pronounced power imbalances. Although the internationalization of postsecondary education affords opportunities for students to learn in geographically diverse areas, cross-border educational partnerships have at times imposed culturally ineffective understandings and practices on host countries.

In addition to the potential for culturally irrelevant practices Roberts and Komives (2016) discuss, the measurement of relationships between these practices and particular global learning outcomes—and the conceptualization of the dimensions actually measured through this process—also risk cross-cultural irrelevance (Dowd, Sawatzky, & Korn, 2011). Such distinct conceptualizations of global learning complicate institutional efforts to articulate the particular global competencies with which their students should emerge (i.e., how one understands and prioritizes these notions), develop specific and measureable global learning outcomes, and identify particular learning opportunities that align with such student learning. This complexity extends beyond postsecondary practice and also complicates scholars' understanding of global learning and the subsequent conceptual and empirical work to which they apply such an understanding. In turn, this illuminates a key challenge institutions and scholars currently face in demonstrating whether global learning outcomes have been actualized. Without conceptual precision, methodological rigor in the measurement of students' global learning outcomes suffers.

The Measurement of Global Learning

Green (2012) argues that internationalization must be considered a *means* to particular goals rather than the end goal itself; it is a strategy for actualizing institutional outcomes, including the preparation of globally competent graduates. While part of this strategy must entail identifying global outcomes and related learning opportunities, these processes must also involve articulating specific criteria and methods to assess whether these outcomes have been actualized (Green, 2013). The measurement of global learning, then, must become a priority if postsecondary education is to prepare global citizens in ways contemporary educational and political rhetoric suggest it should. In considering the measurement of global learning, the specific dimensions measured are certainly of import. But beyond these conceptual concerns identified earlier, *who* is measuring global learning, and *how* is global learning measured?

Situating the measurement of global learning within the various contexts in which this happens serves to clarify its different purposes, methods, and uses. For instance, measuring global learning is often of interest for practitioners charged with overseeing institutional effectiveness, including those in institutional research, strategic planning, and curricular and co-curricular assessment, as well as other institutional areas whose explicit focus involves global learning (i.e., global study and engagement programs, community-based learning, curricular and co-curricular intercultural efforts). However, global learning scholars also have a stake in the measurement of global learning, as they both conceptualize and utilize global learning constructs in their research. So, both practitioners and scholars who examine global learning are therefore involved in considering what to measure and how they will do so.

Though global learning concepts are applied to both qualitative and quantitative inquiry and assessment, the present study focuses on quantitative approaches to the measurement of global learning—and, in particular, within the domain of survey methodology—given the wide use of large-scale, self-report surveys to measure a variety of postsecondary educational outcomes (Kuh & Ikenberry, 2009; Porter, 2011). Further, a key aim of quantitative inquiry is the generalizability of findings to the populations under study. Postsecondary accreditation criteria that charge institutions to assess their missions' support of and their students' contributions to diverse, global societies (e.g., Higher Learning Commission, 2017) and seemingly ubiquitous national and institutional discourses surrounding the improvement of student learning more broadly (Green, 2013) often require the type of large-scale, generalizable measurement efforts that survey methods afford.

A variety of survey instruments currently exists to measure distinct dimensions of global learning including cross-cultural adaptability (Kelley & Myers, 1995), cross-cultural leadership competencies (Kozai Group, 2010), cross-cultural world mindedness (Der-Kerabetian & Metzer, 1993), diversity awareness (Mendez-Russell, Wilderson, Jr., Tolbert, 2003; Stinson, 2007), global perspective development (Iowa State University, 2015), intercultural conflict (Hammer, 2005), intercultural effectiveness (Portalla & Chen, 2010), intercultural readiness (Brinkmann & van Weerdenburg, 2014), and intercultural sensitivity (Bhawuk & Brislin, 1992; Hammer, Bennett, & Wiseman, 2003). These various dimensions—and the array of instruments used to measure them—illustrate both the conceptual scope and complexity embedded in different notions of global learning. Such variety among global learning measures underscores the need for ongoing evaluations of these instruments. Given the types of conclusions drawn from this

variety of global learning data (e.g., determining the achievement of specific global learning outcomes or the effectiveness of particular global learning efforts), the *quality* of the information provided by surveys is a primary consideration (Cone & Foster, 1991). Therefore, examining and improving an instrument's quality should be a primary aim of survey methodologists and a major consideration for consumers of data from survey-based research and assessment efforts (Cizek, Bowen, & Church, 2010).

Examining the Validity of Postsecondary Educational Surveys

Instrument development entails the precise conceptualization of the constructs—or theoretical concepts—under study, but also the operationalization of these constructs into survey items. Instrument development also entails carefully examining an instrument's measurement model, or the relationships between observed indicators (i.e., participants' survey item responses, test scores, or behavioral observation ratings) and latent variables (i.e., constructs, factors, or scales; Brown, 2015). Such examination empirically tests whether the instrument measures what it is theorized to measure within the population under study. The increased use of large-scale surveys in postsecondary institutional assessment and research efforts necessitates more commitment to ensuring valid survey results for both scholars and practitioners (Porter, 2011). Scholars have emphasized the import of examining the factor structures (e.g., Campbell & Cabrera, 2011; LaNasa, Cabrera, & Tangsrud, 2009; Porter, 2011), hierarchical factor structures (e.g., Davidson & Engberg, under review; Nelson Laird, Shoup, & Kuh, 2005; NSSE, 2010), and the development of latent constructs (Sharkness, DeAngelo, & Pryor, 2010) that comprise large-scale, postsecondary surveys. Other scholars have underscored the import of examining the validity of global learning measures in particular (e.g., Hammer, 2011; Hammer et al., 2003;

Paige, Jacobs-Cassuto, Yershova, & DeJaeghere, 2003). The present study adds to the extant literature that has emphasized the value of these types of psychometric evaluation.

Cross-cultural Validation of Survey Instruments

In addition to examining an instrument's measurement model, survey constructs should be understood and measured equivalently across the various groups within the population under study (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 2014). Determining whether this is the case involves examining the equivalency of various aspects of an instrument's measurement model across groups, or examining *measurement invariance* (Dimitrov, 2010). Survey constructs (or scales) represent a group of survey items that measures complex, difficult-to-measure aspects of students' experiences (i.e., their perceptions, attitudes, and behaviors; Brown, 2015). Scales operate with the assumption that respondents' different levels of the underlying construct (respondents' different levels of intercultural knowledge, as an example) relate to their selection of particular responses to the items the scale measures (Bowen & Masa, 2015). When survey scale scores are compared across different ethnoracial groups, this is often done with the assumption that the various groups understand the survey constructs identically and that relationships between the scales' items and the underlying constructs are identical across groups when this may not be the case (Bowen & Masa, 2015). Only when measurement invariance is empirically determined can differences relative to means and regression coefficients be validly compared across groups (van de Schoot, Lugtig, & Hox, 2012). However, when measurement invariance across groups cannot be empirically determined, the groups' different responses may be understood in terms of item bias or constructs operating differently

across groups (Brown, 2015). This is particularly important given that cross-cultural scholars have argued that group comparisons often assume measurement model and response bias invariance, despite the heterogeneity often present within populations under study (Steinmetz, Schmidt, Tina-Booh, Wieczorek, & Schwartz, 2009).

The conceptualization of survey constructs is inherently reductionist and undergirded by assumptions, values, and context (Tanaka, 2002). To include particular aspects in one's construction of a concept serves to exclude other components from such an understanding. Survey items—operationalized from these conceptualizations—are then subject to a response process involving distinct cognitive strategies survey respondents employ (Tourangeau, Rips, & Rasinski, 2000). Students use their cultural experiences in making meaning and responding to survey items (Warnecke, Johnson, Chávez, Sudman, O'Rourke, Lacey, & Horm, 1997); words are subjectively interpreted, "bound by culture and context" (Gonyea, 2005, p. 76). Such arguments illustrate the need to expand psychometric evaluation—including cross-cultural validation studies in particular—for educational surveys.

Because survey respondents can understand identical items in markedly different ways, the examination of whether and why this is the case is imperative in survey research (King, Murray, Salomon, & Tandon, 2004). Cross-cultural scholars discuss this in terms of viewing this through an applicability paradigm, or a consideration of the applicability of an instrument's factors outside of the context in which they were developed (Marsh, 2007). Put differently, etic and emic constructs are differentiated in their research. Etic constructs are considered universally understood across all cultural groups, while emic constructs are culturally specific,

assume particular meaning in different cultural groups, and can be differently or not understood across groups (Warnecke et al., 1997).

> In multicultural societies such as the United States, a so-called "standard question" presumed to be etic may in fact be emic and, hence, answered differently by respondents of varying educational, racial, or ethnic backgrounds. When constructs that are emic are treated as etic, a category fallacy results, the practical significance of which may constitute a problem in generalizing data across respondents and/or a failure of the respondent to answer the question being asked by the investigator. (Warnecke et al., 1997, p. 335)

The potential for divergent understandings relative to intercultural exchange between diverse groups of students frames the need for cross-cultural validation of global learning measures. Further, changing demographics within U.S. postsecondary education (USDE, 2015a; IIE, 2016) warrant the need for inclusive assessment and research practices that also necessitate this type of inquiry. Previous research using the GPI has suggested different relationships between particular types of on- and off-campus engagement and the various developmental outcomes measured by the instrument across ethnoracial (e.g., Engberg, Davidson, Manderino, & Jourian, 2016) and international (e.g., Engberg et al., 2016; Glass, Buus, & Braskamp, 2013) student groups. However, what remains unclear is whether these observed differences are resultant from differences in how global perspective development is measured across groups rather than simply different levels of development across these groups.

Statement of the Problem

The implications of failing to examine measurement invariance across groups involve generalizing or comparing findings from measures presumed universally applicable across groups. Further, validity studies require the examination of cultural bias and interrogating "value assumptions" attached to the theorized constructs under study (Dowd et al., 2011, p. 23). Failure

to examine instruments in these ways poses the risk of reifying a construct developed for a particular cultural group that has been applied to other cultural groups without establishing validity or coherence (Kleinman, 1988). The implications of this relate to how undergraduates' development and developmental interventions are theorized and the extent to which these are culturally appropriate. As a hypothetical example, if evidence suggests that any of the GPI's constructs are understood differently across groups, this could suggest that the theorization of those constructs is not cross-culturally relevant. Even if the constructs are understood equivalently across groups, ensuring that items do not function differently (i.e., there is no systematic measurement bias conditional on one's group membership) is critical in making the claim that particular measures are suited for diverse student populations. If educational opportunities—and the evaluation of them—are aligned with dimensions of development that are either not understood or measured the same across groups, the educational benefits of these efforts do not extend to all students. Given its increasingly ethnoracially diverse settings (USDE, 2015a), commitment to diversifying its institutions (Smith, 2009), and prioritization of global learning outcomes (AAC&U, 2007), those within postsecondary education examining global learning must carefully execute this area of study.

To date, the GPI has undergone psychometric evaluation (i.e., Braskamp, Braskamp, & Engberg, 2014; Davidson & Engberg, under review) but no study yet exists that has tested for measurement invariance across ethnoracial groups. Given the instrument's wide use—over 120,000 individuals have completed the GPI since 2008 to measure various dimensions of their global perspective development (Braskamp et al., 2014)—this type of inquiry is warranted. Prior research using the GPI has used the findings from the instrument to—among other analytic

strategies—compare means or model coefficients across ethnoracial groups. Further, as scholars use the GPI's conceptualization of global perspective development in their work, testing the cross-cultural validity of such a conceptualization is essential. Finally, as practitioners use GPI data in a variety of institutional research or curricular and co-curricular assessment efforts, the validity of such work warrants an examination of whether the instrument functions the same across ethnoracial groups.

Purpose and Research Questions

The domains that comprise the GPI's global perspective development construct may not be universally theorized, understood, or measured across ethnoracial groups. There exists a need to examine the validity of the GPI's measurement model across ethnoracial groups for two key reasons. First, doing so will provide evidence of whether the developmental constructs operate equivalently across a diverse group of undergraduates. Such evidence relates to the conceptualization of aspects of undergraduates' development and—if understanding such development stands to inform developmental opportunities—interventions aimed at promoting aspects of their development. Second, examining the cross-cultural validity of the GPI will also provide important evidence for practitioners and researchers that informs their ability to compare students' global perspective development across ethnoracial groups, which has historically been of interest to those using the GPI to understand nuances related to undergraduates' development.

To investigate this, multiple-group confirmatory factor analyses (MGCFAs) will be conducted to examine measurement invariance of the GPI across four ethnoracial groups: Asian/Native Hawaiian/other Pacific Islander, black/African American, Hispanic, and white. Examining measurement invariance via MGCFAs requires testing the cross-group equivalency

of various measurement model parameters in a particular order. Evidence for specific types of invariance are required in order to determine whether particular model parameters differ across the groups under study (Steenkamp & Baumgartner, 1998). While an expanded discussion on each of these types of invariance to be tested (i.e., definitions of each and what such evidence suggests) is provided in Chapter Three, I mention each here to connect these to my research questions. First, evidence of both equal form (i.e., equivalent factor structures) and factor loadings/coefficients (i.e., the strength of relationships between the items and scales/higher- and lower-order constructs) across groups is required to determine whether the groups understand the measured constructs equivalently. Second, evidence of equal item thresholds (i.e., correspondence between levels of the underlying constructs and response choices on the scales' items) across groups is required in order to compare GPI mean scores and regression coefficients across ethnoracial groups. To test these types of invariance, this study first seeks to answer the following three research questions in this sequence:

Research Q1: Is *equal form* observed across ethnoracial groups for the GPI's items?

Research Q2: Are *equal first-order factor loadings and second-order coefficients* observed across ethnoracial groups?

Research Q3: Are the GPI's *item thresholds* invariant across ethnoracial groups?

If evidence related to these first three research questions suggests invariance across those model parameters, stricter forms of invariance will be tested. To examine this, I will test the equality of unexplained variance in both the first-order factors (i.e., equal factor disturbances) and in each GPI item (i.e., equal item error variances) to understand whether the latent constructs are measured equivalently across groups. I will answer the following two research questions:

Research Q4: Are *equal disturbances* of the first-order factors observed across ethnoracial groups?

Research Q5: Are *equal item error variances* observed across ethnoracial groups?

Finally, there are two aspects related to the GPI that have not yet been examined. First, an earlier study (Davidson & Engberg, under review) empirically validated the GPI's hierarchical factor structure. If the present study also provides evidence for this factor structure, one can use the higher-order factor mean to understand students' average level of overall global perspective development. Using MGCFAs to examine the equivalence of the GPI's second-order factor mean across groups provides evidence whether students' level of global perspective development differs across groups. The following research question examines this:

Research Q6: Do significant cross-group differences relative to the GPI's second-order global perspective development factor mean exist?

Second, given the validity evidence obtained from the study's analyses, this study also seeks to inform any refinement efforts of the GPI. In particular, I will examine validity evidence related to the GPI's hierarchical factor structure and the convergent and discriminant validity of the GPI's scales. The purpose of the validation study by Davidson and Engberg (under review) involved testing various factor structures for the GPI. Their results—using two independent samples of undergraduates—suggested that a hierarchical factor structure provided the best model fit; this provided important empirical evidence to support the GPI's theorization. However, these results also illuminated that while the hierarchical model fit the data well, particular model parameters (i.e., the relationship between the GPI's Cognitive Knowing factor

and the higher-order global perspective development factor, GPI item error variances) could benefit from more extensive psychometric study.

Together, convergent and discriminant validity comprise important aspects of construct validity that allow instrument developers to understand the extent that an instrument has been operationalized in ways that accurately reflect its underlying theories. Specifically, examining convergent and discriminant validity evidence illuminates more/less effective measures of an instrument's underlying constructs. Such evidence is useful in identifying specific item or scale refinement opportunities. The study's final research question examines these aspects:

Research Q7: Does evidence support the hierarchical factor structure of the GPI and convergent and discriminant validity of the GPI's scales?

Scope of the Data for the Study

Data for this study were provided by Iowa State University's Research Institute for Studies in Education (RISE), which administers the GPI to college students, faculty, and administrators at colleges and universities across the globe. This study used data from the GPI General Form administered during the 2015-16 and 2016-17 academic years and includes data from only undergraduate students. The undergraduate student sample used for the present study is a multi-institutional sample that includes both domestic (defined on the instrument as American students attending an American college/university) and international students (defined on the instrument as non-American students attending an American college/university).

Theoretical Grounding

Employing confirmatory factor analyses (CFAs) for validation purposes requires a solid theoretical basis to guide model specification and evaluation (Brown, 2015). In the case of

cross-cultural validation studies using a CFA framework (i.e., comparing CFA models across groups using MGCFA), theory undergirds any hypotheses concerning differences between model parameters across the particular groups under study (Steinmetz et al., 2009). Kirkhart (1995) explains that in multicultural validation efforts, theoretical foundations and methodologies must be examined to understand the extent that they are culturally appropriate. Extending from this, two theories ground the present cross-cultural validity investigation. This broader theoretical rationale grounds the need to examine measurement invariance more generally by situating the present study within literature that illuminates the subjective, culturally-bound nature of both survey constructs and the survey response process. The concept of *construct underrepresentation* (Messick, 1994) offers an explanation of both the culturally-bound nature of survey constructs and implications for under-theorized notions of such developmental constructs. The *cognitive model of the survey response process* outlined by Tourangeau et al. (2000) explains the complex, culturally-bound nature of one's interaction with survey instruments. Both of these concepts are discussed in detail in Chapter Two. In addition to the broader theoretical rationale surrounding multicultural validation efforts, I also ground this study in the extensive literature on cultural variability related to each of the GPI's six developmental dimensions, namely as such variability relates to students' ethnoracial identities. In particular, this expanded discussion in Chapter Two provides in-depth theoretical rationale to justify examining measurement invariance of the GPI across ethnoracial groups.

Key Terms

Before proceeding, it is useful to define particular concepts as they appear throughout this study for conceptual clarity. I use the following concepts drawn from scholars' multidisciplinary

work. But my intentional use of particular terms also arises from my own roots in the quantitative criticalism paradigm, which informs the theorization and integration of analytical strategies in ways that place misrepresented and subordinated groups and systems of oppression at the center of inquiry, aiming to disrupt normative quantitative research (Stage, 2007).

Ethnorace and *ethnoracial identity* are used throughout instead of race/ethnicity and racial/ethnic identity. While still an emerging concept, my use of it aligns with the development of this concept to apply to ethnic groups who possess both ethnic and racialized characteristics distinct from others with whom they share a racial category (Alcoff, 2009) and to acknowledge "what were once called either 'races' or 'ethnic groups,' recognizing the blurred, contingent, and constructed character of the relevant boundaries" (Hollinger, 2003, p. 1378). Alcoff (2009) explains that "clear-cut distinctions between race and ethnicity do not always hold" and that capturing the meaning of a group's identity sometimes necessitates including both, as neither race nor ethnicity—"binary concepts imagined as independent"—by themselves adequately captures the racialized complexity of particular groups (pp. 122-123). Although Alcoff originally developed the concept to more accurately represent the racialized identities of Latinx individuals, others have applied the concept of ethnorace to those of Asian descent to examine the diversity of identity meaning and status within such a heterogeneous group (e.g., Jiménez & Horowitz, 2013). I use the concept of ethnorace in the present study to acknowledge this hybrid identity (Alcoff, 2009) given the available racial and ethnic identity categories on the GPI. The potential ethnic diversity within the GPI's Asian/Native Hawaiian/other Pacific Islander and Hispanic identity categories, in particular, necessitates careful consideration of the interaction

between students' ethnicities and the potentially different ways these operate in the racialized contexts of U.S. postsecondary education.

Appropriately theorizing both the context and process of racial identification is necessary in understanding the meaning survey respondents give to *race* in responding about their racial backgrounds (Johnston, Ozaki, Pizzolato, & Chaudhari, 2014). For instance, Rockquemore, Brunsma, and Delgado (2009) differentiate between three constructions of how one understands one's race in responding to survey items about racial identification; individuals consider their racial identity (1) as their own self-understanding or identification, (2) in terms of racial ascription, or how others understand and racially categorize them, and (3) relative to the available racial identity categories in a particular situation. Though these constructions are interrelated, they are distinct; when survey respondents "are asked about their 'race' they may enact one or many of these meanings in formulating their answers" (Johnston et al., 2014, p. 58).

On the GPI, respondents are asked to indicate in a single item their racial or ethnic backgrounds by selecting *all that apply* from the following six categories: American Indian or Alaska Native, Asian, Black or African American, Hispanic (of any race), Native Hawaiian or other Pacific Islander, and white. The GPI does not currently include options in its racial or ethnic background item that allow respondents to indicate their racial or ethnic background as something other than the categories provided (either as a stand-alone category or as an open-field response option), though respondents can choose to not answer this item. The complexity and variability in individuals' response processes about their race relate to the validity of data on respondents' race but also to the utilization of findings using such data (Rockquemore et al., 2009). Respondents' own racial identities, how their racialized selves are understood by others,

and the available racial categories from which to select may not align for a variety of reasons (Rockquemore et al., 2009). The present study bears in mind the theoretical, methodological, and utilization implications of this.

Finally, *minoritized* is intentionally used throughout this study instead of minorities. This strategy aligns with the socially constructed nature of racial subordination; minority status is not objective, fixed, or inherent in every context, so the act of assigning persons to an ethnoracially minoritized status reflects the dominating force of whiteness in those particular contexts (Gillborn, 2005; Harper, 2012).

Contributions of the Study

The present study informs postsecondary educational practice and scholarship in several ways. The findings add to the large-scale postsecondary educational survey validation and cross-cultural validation literature related to global learning and intercultural competency instruments. The study also informs an emerging body of work examining widely accepted notions of global learning. First, the general increase in the use of surveys in both educational practice and scholarship (Porter, 2011) coupled with an increased focus on improving, and therefore assessing, students' global learning (Green, 2013) necessitates rigorous survey validation study. Though validity as a concept (i.e., the precise components that should comprise this overall definition) is still widely debated, it stands as the gold standard for determining an instrument's quality (Cizek et al., 2010) and an expected standard practice in educational and psychological testing (AERA et al., 2014). However, though theoretically dense, validation work has a "long history of rather anemic practice" (Cizek et al., 2010, p. 1). Given the GPI's use in both institutional assessment and scholarly efforts, the present validation study is timely.

The validity evidence collected from this study can illuminate any necessary instrument refinement opportunities. In terms of cross-cultural validation, if there are particular parameters (i.e., factor loadings, item thresholds) that do not operate the same across groups, this will be important to address by revisiting the constructs' theorization and operationalization as appropriate. In addition, convergent and discriminant validity evidence provide very specific information about items that are more/less effective measures. If measurement invariance across the four groups is determined, for instance, this additional evidence provides insight into any items that would benefit from refinement.

In terms of cross-cultural validation of global learning instruments, the present study adds to emerging inquiry in this area. As U.S. postsecondary education continues to diversify relative to the ethnoracial and international student composition of its campuses, and as the internationalization of higher education continues to drive institutional efforts including global learning, the assessment of such efforts will continue to grow, demanding valid and inclusive methods for doing so. Given the possibility of construct underrepresentation in the field's major surveys (Porter, 2011) and the cultural variation inherent in particular components of the survey response process (Tourangeau et al., 2000; Warnecke et al., 1997), cross-cultural validation work is necessary.

There is a paucity of rigorous validity studies of intercultural competency measures (Schnabel, Kelava, van de Vijver, & Seifert, 2015). There are numerous examples of cross-cultural validation studies of instruments similar in scope to the GPI. However, these earlier validation efforts have largely not employed the methods necessary to claim that the theorization and measurement of their constructs actually operate equivalently across groups. For instance,

Hammer (2011) tested the cross-cultural validity of the Intercultural Development Inventory (IDI)—a widely used and cited measure of intercultural competencies—with 11 cross-cultural groups (i.e., employees and students across nine countries). Though Hammer used confirmatory factor analyses (CFAs) to validate the IDI's measurement model, he concluded cross-group validity using only the evidence from the groups' separate baseline models that suggested adequate fit. By not employing MGCFAs, Hammer's validation study does not provide evidence of the equivalency of the various parameters outlined in the present study's research questions. Additionally, Hammer (2005) asserted that examining the effects of gender, education, and previously living in another culture using *t* tests and ANOVAs provided validity evidence suggesting that no differences exist across these groups relative to what the Intercultural Conflict Style Inventory measures. Studies involving the Global Competencies Inventory (Stevens, Bird, Mendenhall, & Oddou, 2014) and the Intercultural Readiness Check (Van der Zee & Brinkmann, 2004) examined only group mean differences, not underlying factor structure equivalencies or differential item functioning. None of these studies employed MGCFAs to test for measurement invariance across the groups under study. Without examining measurement invariance in concluding cross-group validity, these previous findings *assume* the underlying theories on which those surveys are based are cross-group appropriate and that the same constructs are measured in the same way across groups without evidence to support such an assumption. The present study helps address this serious issue and significant gap related to global learning-related instrumentation.

Finally, findings from this study will aid in the examination of widely accepted notions of global learning and their applicability across ethnoracial groups. While the theoretical

underpinnings of the GPI span multidimensional models of cognitive, intrapersonal, and interpersonal development, ensuring that such concepts apply universally is paramount. Such work carries implications for the operationalization of global learning concepts into surveys. These findings can also inform how institutions and scholars understand these global learning concepts for other purposes as well (i.e., conceptual precision relative to identifying institutional efforts and determining assessment criteria related to global learning and qualitative inquiry).

Organization of Remaining Chapters

This study's remaining chapters are organized as follows. Chapter Two provides theoretical support for investigating cross-cultural validity, a review of the theoretical underpinnings of the GPI, and a review of the literature on cultural variability in relation to each of the GPI's developmental dimensions. Chapter Three includes a detailed discussion on the study's research questions, methodologies, and limitations. Chapter Four presents the findings to answer each of my research questions and summarizes the study's overall findings. Finally, Chapter Five includes an expanded discussion of the study's findings, implications for research and practice, and ways in which these findings can inform future inquiry.

CHAPTER TWO

LITERATURE REVIEW

In survey research, cross-cultural validation studies require a sound theoretical rationale relative to investigating particular differences across groups (Steinmetz et al., 2009). As such, this chapter is structured as follows. Given that this study seeks to contribute toward survey validation efforts, I first review the validity framework as explained in the current *Standards for Psychological and Educational Testing* (AERA et al., 2014) with particular attention to the evaluation of validity and its relation to culturally-relevant measurement. Because this study seeks to examine measurement invariance, I discuss the broader theoretical rationale for doing so given the inherently subjective nature of both survey constructs and the survey response process. Next, I discuss the GPI's theoretical foundations with attention to the overarching theories that inform the interrelatedness of the developmental domains included in the instrument as well as the theories that inform the six dimensions that comprise the GPI's global perspective development construct. Extending from that discussion, I review the literature on cultural variability in relation to each of the GPI's six developmental dimensions to provide precise theoretical justification for examining whether the instrument's global perspective development items are universally understood across ethnoracial groups. I conclude the chapter with the study's conceptual framework.

Examining the Validity of Survey Instruments

The notion of validity is too often assumed or implied in the utilization of surveys in postsecondary educational research and assessment efforts (Porter, 2011). Even when validity is explicitly examined, its understanding and evidence vary immensely across studies (Cizek et al., 2010; Kane, 2001; Sireci, 2007). Though the centrality of validity evidence in psychometric evaluation has been widely accepted by researchers, Kane (2013) posits that a debate continues about which type of validity is of central importance to researchers, noting that "general discussions of validity can become quite complicated, and validation can seem daunting" (p. 448).

Underscoring the culturally-bound process of validation work, Kirkhart's (1995) concept of multicultural validity locates culture at the center of validation efforts. Examining multicultural validity involves centering the consideration of how cultural influences relate to diverse understandings and meanings; this spans epistemological, theoretical, and methodological dimensions as well as professional practice (Kirkart, 1995). The present study seeks to contribute to the cross-cultural validation body of educational survey research. In doing so, my aim is to contribute toward culturally-relevant assessment, evaluation, and research efforts with the precision and rigor necessary to advance this type of inquiry. As such, the following discussion situates this study within the current understanding of validity and describes the particular contributions that extend from doing so.

Because the GPI is used for educational assessment, evaluation, and research purposes, standards for (1) evaluating validity and (2) considering fairness in testing as explained in the current *Standards* (AERA et al., 2014) guide the current study. Importantly, the *Standards*

prioritize examining the extent to which educational measurement considers the culturally diverse theoretical, methodological, and consequential aspects encapsulated within Kirkhart's (1995) multicultural validity concept. I review the *Standards'* concepts of validity and validation and apply these to the present study here given its cross-cultural validation purpose.

With reference to the concept of validity, the *Standards* explain this in terms of what is known as the argument-based approach to validation. Kane (2013) explains the argument-based approach to measurement validation as constructing a validity *argument* instead of conducting validity research. The argument-based approach to validation involves two steps: first, one must clearly specify the claims that support the particular interpretation and use of test scores (i.e., referred to as the interpretation/use argument), then one must provide rigorous empirical evidence to evaluate such claims (i.e., referred to as the validity argument). Newton and Shaw (2013) explain the utility of such an approach:

> Argument provides evaluators with a methodology for subdividing the big question of validity into manageable chunks. It clarifies where to begin (with the intended interpretation and use of test scores), how to proceed (by making explicit the claims that would support that interpretation and use in the form of an argument, and by testing their assumptions), and when to stop (when the argument is judged to be coherent and complete, and when its inferences and assumptions are judged to be plausible). The argument-based approach better equips evaluators to identify the full range of issues that need to be addressed, as well as the issues that require most attention: the weakest links in the argument chain. (p. 23)

In explaining its associations' position on validation, the *Standards* remind educational researchers and evaluators that one evaluates the interpretations of test scores for particular uses, not the test itself. The *Standards* make clear that each intended interpretation of test scores must be articulated, and validity evidence that supports each intended interpretation must be provided. Concerning the GPI, the survey is explained to potential users as "an appropriate assessment for

aggregate data to inform self-studies, program assessment or evaluation, institutional effectiveness, and accreditation" (RISE, 2017a, p. 3). The present study seeks to examine whether additional validity evidence exists that supports the disaggregation of GPI data to compare GPI item and scale scores across ethnoracial groups. Such evidence is critical in scholars' and institutions' efforts related to understanding potentially different levels of global perspective development during the undergraduate years and in advancing interventions that address particular developmental needs. Per the argument-based approach to validation—and therefore, per the validity principles outlined in the *Standards*—if the GPI is to be used in ways that report subgroup item or scale scores or differences, this intended use must be articulated by test users and validity evidence should support this use. The present study aims to contribute toward such validity evidence.

The *Standards* explain the principles for establishing validity in terms of three thematic clusters: establishing intended uses and interpretations, issues regarding samples and settings used in validation, and specific forms of validity evidence. Salient to the current study, the *Standards* outline seven initial standards related to establishing the intended uses and interpretations of a test. Among other aspects, these particular standards explain that the population(s) for whom a test is intended must be clearly stated and that the construct(s) measured by the test should be clearly described. These standards also explain that if evidence for the validity of an interpretation for a particular use does not yet exist, this should be made clear to potential users, and so should the recommendation to not make unsupported interpretations based on the test. The *Standards* also explain specific types of validity evidence that should be used to support the intended interpretation of a test given a particular use. While

there are numerous standards that outline various aspects of validity evidence, there is one standard of particular importance to the present study. The *Standards* explain validity evidence based on the internal structure of an instrument.

In addition to offering the argument-based approach to validation as a guiding framework, the *Standards* also clearly explain that the consideration of fairness is central to the study of validity and outline guiding principles that span all phases of test development and use. With reference to the notion of fairness, the *Standards* define this as "responsiveness to individual characteristics and testing contexts so that test scores will yield valid interpretations for intended uses" (p. 50). Of particular significance to the present study, the *Standards* cite measurement bias as an important threat to ensuring fairness in testing. As mentioned in the guiding principles related to validity, the *Standards* state in the principles that guide fairness in testing that the construct being measured by an instrument must have equivalent meaning across the groups within the population under study, explaining that "this is especially important when the assessment crosses international borders and cultures" (p. 52).

The current study seeks to provide evidence related to the cross-cultural applicability of the GPI's global perspective development construct. Subsumed underneath the cross-cultural validation purpose of the present study, it also addresses the *Standards'* charge to consider fairness in testing. Providing this particular evidence relates to specific intended interpretations and uses of the GPI's scale scores across the ethnoracial groups under study. The present study will examine measurement invariance of the GPI across ethnoracial groups. The following section outlines the overarching theoretical rationale surrounding the current study's validation efforts. This broader rationale grounds the need to examine measurement invariance more

generally by situating the present study within literature that illuminates the subjective, culturally-bound nature of both survey constructs and the survey response process.

Theoretical Grounding for Examining Measurement Invariance

Kirkhart (1995) posits that in multicultural validation efforts, theoretical foundations and methodologies must be examined to understand the extent that they are culturally appropriate. Extending from this, two theories ground the present investigation to examine whether the GPI's global perspective development constructs are universally understood across an ethnoracially diverse sample of undergraduates. Construct underrepresentation (Messick, 1994) offers an explanation of both the culturally-bound nature of survey constructs and implications for under-theorized notions of such developmental constructs. The cognitive model of the survey response process outlined by Tourangeau et al. (2000) explains the complex, culturally-bound nature of one's interaction with survey instruments. Both theories are discussed below, with particular attention to cultural variability.

Construct Underrepresentation

In survey research, construct underrepresentation poses a major threat to instrument validity (AERA et al., 2014; Hubley & Zumbo, 2011; Messick, 1994). Construct underrepresentation is observed when "the assessment is too narrow and fails to include important dimensions or facets of the construct" (Messick, 1994, p. 4). Obtaining empirical evidence that supports or negates construct underrepresentation of a given instrument is essential in determining the extent of cultural bias within an assessment (Goodwin & Leech, 2003). Further, a primary measurement concern relative to findings generated from a particular instrument involves more consequential considerations. Potential adverse social consequences

for any given group must not emerge from any type of instrument invalidity, including construct underrepresentation (Hubley & Zumbo, 2011; Messick, 1994).

Dowd et al. (2011) explain the seriousness of potential construct underrepresentation in large-scale postsecondary educational surveys, namely as this relates to neglect on the part of survey developers to adequately and inclusively conceptualize the experiences of marginalized student groups in the constructs under study. The current study is also informed by Tanaka's (2002) position that charges those in survey research "to renew the underlying theoretical bases for these modern instruments by recontextualizing them within today's shifting, overlapping, polycultural terrain" (p. 279). In terms of the GPI, the concept of construct underrepresentation drives the current study's methodology since—as the forthcoming literature on the culturally variable nature of student development will illuminate—it is plausible for students of color and white students to understand interacting across difference, sensitivity toward difference, and their own identities (i.e., constructs the GPI measures) markedly differently.

The Survey Response Process

Tourangeau et al. (2000) advanced a four-stage theory of decision making and the cognitive strategies respondents employ when answering survey items. The model includes four cognitive strategies survey respondents employ: comprehension, retrieval, judgment, and response. Understanding the psychology of the survey response process is essential in considering the validity of self-report, attitudinal survey items (Gonyea, 2005). Such items comprise a majority of the large-scale surveys such as the GPI that higher education professionals utilize to measure a range of student engagement, learning, and development.

Effects that influence survey responses are possible in each of the four model components outlined by Tourangeau et al. (2000) and are discussed below.

Item comprehension. First, the comprehension component of the model explains how respondents attend to items and instructions, assign meaning to an item, determine the item's purpose (i.e., the specific information sought in the response), and employ reasoning to understand survey language. Issues arise in various aspects related to respondents' comprehension, but at the very least, items may be understood differently across respondents (Tourangeau et al., 2000). If this happens, a respondent could answer an entirely different question than intended (Tourangeau et al., 2000).

Porter (2011) further underscores the import of this particular component in the argument to attend more to issues surrounding survey validity. He argues that researchers using college student surveys often neglect to consider students' comprehension of the words and phrases embedded in surveys (e.g., that educational jargon or overly abstract concepts are often used) and whether any comprehension differs across students. Tourangeau et al. (2000) illustrate a paradox of sorts related to survey item comprehension. Relative to any given construct a survey item measures, both the researcher who developed the item and the respondent have their own understanding of this, although these two likely differ in terms of the knowledge they bring to the survey process relative to that construct. The item developer typically offers a more theoretically-based, nuanced understanding of the construct than a respondent; as such, Tourangeau et al. (2000) question how their interpretations could *ever* be the same. Implications arise related to constructs that are operationalized from a dominant postsecondary educational

discourse (Harper, 2012) and instruments developed by a majority culture (Johnson, Cho, Holbrook, O'Rourke, Warnecke, & Chavez, 2006).

Scholars argue that substantial cultural variation exists relative to the comprehension of many survey items (e.g., Johnson et al., 2006; Tourangeau et al., 2000; Warnecke et al., 1997). Tourangeau et al. (2000) outline several aspects of survey items that could pose difficulty for respondents. These include grammatical and syntax issues (i.e., ambiguity, complexity, or unfamiliarity), semantic effects including false presuppositions (i.e., assuming particular characteristics apply to respondents when they do not) and leading questions (i.e., items containing a false presupposition about something), and vague concepts and/or quantifiers. In particular, Johnson et al. (2006) found respondents' race or ethnicity moderated the effects of item response format, reading level, and item length on item comprehension, even when controlling for other relevant demographic variables such as age and education. Johnson et al. explain that "simplifying question length, reading level, and response formats would appear to improve overall question comprehension at the cost of enhancing cross-cultural disparities" (p. 667). They argue that the body of research on survey design has historically been conducted mostly with white, non-Hispanic respondents; as such, it could be expected that these findings do not hold across cultures, as they found in their study.

Retrieval. The retrieval component of the Tourangeau et al. (2000) model is also salient to the present study. This second component explains how respondents retrieve information and/or infer missing information to answer survey items. Tourangeau et al. explain that these cognitive strategies differ depending on whether a given item is factual or attitudinal. A cognitive theory is necessary to understand this component relative to attitudinal items because

there is "a continuum corresponding to how well articulated a respondent's attitude is. At the more articulated end, the respondent has a preformed opinion just waiting to be offered…; at the less articulated end, the respondent has no opinion" (p. 12).

Most of the GPI's global perspective development items are attitudinal in nature (the exception involves three of the four items on the Social Interaction scale that ask for behavioral information about the extent of interaction across difference). Attitudes are defined as "object evaluations stored in memory" (Judd, Drake, Downing, & Krosnick, 1991, p. 193). In searching for information to answer attitudinal items, respondents must either retrieve an existing evaluation (or attitude) from memory or search for attitude-relevant information they can use to develop an attitudinal response at that moment (Judd et al., 1991). During this stage, memories are retrieved either episodically (i.e., through specific memories) or semantically (i.e., through generalizations or schemas that are more non-specific). Semantic memory is the representation of knowledge about one's world; its content is generalized from one's experiences and is therefore conceptual without particular event-specific references (Binder & Desai, 2011). The *constructed* nature of culture arises from conceptual knowledge, so the extent to which humans use concepts from their semantic memory in daily life is vast (Binder & Desai, 2011). Given this context, it is expected that semantic memories, in particular, are culturally influenced (Wernecke et al., 1997).

While some respondents retrieve information from existing memory to answer items, some respondents rely on other strategies during this stage. Tourangeau et al. (2000) discuss three other strategies respondents use in answering attitudinal items; the strategy used is most dependent on the accessibility of information. First, respondents can rely on their general

impressions or stereotypes about the target instead of more specific information in answering an item, though this is often conditional on time afforded for the survey and the perceived import of respondents' answers. Second, respondents can rely on their general attitudes or values in answering items. Broader ideological predispositions and values often reflect deeply held principles and undergird respondents' attitudes toward specific issues (Tourangeau et al., 2000), including the types of developmental constructs embedded in the GPI. Tourangeau et al. explain that respondents often do not have static views about the very specific issues sometimes asked about in surveys. Rather, they often deduce from more generalized values in formulating a response, which carries implications for item phrasing and context (i.e., subtle changes in wording or ordering of items can elicit dramatically different responses; Tourangeau et al., 2000). Finally, respondents can rely on their specific beliefs or feelings about what is being asked in answering an item more inductively (i.e., essentially the opposite process as just described). With this strategy, respondents recall and use specific information related to the issue in responding. This largely occurs when respondents are answering an overall evaluative item and base such a rating on specific dimensions or events they recall as they answer the item.

Judgment. The judgment component of the Tourangeau et al. (2000) model explains how respondents rate, estimate frequencies, consider their confidence in a given response, agree or disagree with a particular position, or evaluate the import of conflicting information from the prior retrieval processes in order to form a judgment. Tourangeau et al. argue that "the judgments called for by attitude questions are rarely absolute but are typically made in relation to some standard, generally an implicit one" (p. 197). As such, attitudinal items are particularly subject to instability of responses across time and across contexts (Tourangeau et al., 2000).

If memory retrieval yields an exact answer, no further judgment is required by a respondent. However, as just discussed, sometimes a need arises to render a judgment about information retrieved from memory; this is a complex task that Wernecke et al. (1997) argue is easily affected by respondents' racial and cultural backgrounds. Specifically, Wernecke et al. outline three potential influences in respondents' judgment formation processes that are subject to cultural variation. First, the more a respondent has thought about a given survey topic, the more accessible information becomes for rendering a judgment about related survey items. Second, cues within survey items may be culturally influenced (e.g., avoiding extreme responses, qualifying responses, subject to conversational norms with varying emphases relative to social interactions). Third, there is cultural variation in respondents' probabilistic thinking (i.e., the level of uncertainty tolerated by a respondent).

Survey responses. Finally, the response component outlined by Tourangeau et al. (2000) explains how respondents map their judgment onto available response categories and edit their responses for consistency, acceptability, or other criteria. Tourangeau et al. argue that sensitive items are of special concern here (i.e., understanding the perceived risks or benefits to respondents in answering truthfully) and that, accordingly, there are conscious and unconscious processes that relate to this component. For instance, social desirability and social or cultural distance influence response editing (i.e., overreporting and underreporting; Wernecke et al., 1997). Definitions of social desirability are culturally variable and contextual, influenced by social interaction and contextual norms, and contingent on what is measured and the social location of the participants (Gonyea, 2005; Wernecke et al., 1997). Students perceive particular norms within their college contexts, including the value placed on their learning and

development in a general sense (Bowman & Hill, 2011) but also the particular types of learning and development measured by the GPI.

Considered together, subjectivity relative to both the theorization of survey constructs and the survey response process underscore the need to examine whether diverse groups share an understanding of developmental survey measures such as the GPI. If evidence suggests that diverse groups' understandings of the developmental constructs under study do not vary, cross-group comparisons related to those measures may be made. One may then proceed with using such cross-group comparisons in understanding diverse groups' development and using such an understanding to inform developmental efforts. However, if evidence instead suggests that diverse groups understand the developmental constructs under study differently, serious consequences emerge. Such evidence likely suggests that for particular students, their development may not have been theorized accurately. Extending from this, interventions that aim to facilitate a limited understanding of development may not be conceptualized appropriately for particular groups. The effectiveness of these interventions cannot be gauged accurately if the measures to determine this are not valid across groups. Thus, this broader theoretical rationale to examine the cross-cultural validity of the GPI is compelling as institutions increasingly seek to understand and report their students' intercultural competencies (Green, 2013) and use the GPI to do so. Extending from the aforementioned broader rationale to conduct this study, the following discussions first outline the theoretical foundations of the GPI and subsequently review the literature on the culturally variable nature of the types of developmental constructs measured by the GPI.

Theoretical Foundations of the GPI

The GPI's global perspective development construct is rooted in holistic human development. Such a conceptualization implies a multidimensional understanding of the particular developmental domains that underlie the instrument's survey items (RISE, 2017a). Cultural development and intercultural communication theoretical perspectives inform the three developmental domains that the GPI's items measure. The three domains relate to how individuals think (cognitive development), feel (intrapersonal development), and relate (interpersonal development) throughout the lifespan (RISE, 2017a). There are six GPI scales that span these cognitive, intrapersonal, and interpersonal domains of development with two scales per developmental domain. In each developmental domain, one scale is derived from cultural development theories, while the other scale is derived from intercultural communication theory (RISE, 2017a).

Cultural Development

Two main cultural development theoretical perspectives—Kegan's (1994) concept of self-authorship and King and Baxter Magolda's (2005) concept of intercultural maturity—undergird three of the GPI's scales (i.e., the Cognitive Knowing, Intrapersonal Identity, and Interpersonal Social Responsibility scales). Kegan (1994) advanced a constructive-developmental way of understanding the increasingly complex ways that individuals construct meaning in their lives, a concept he referred to as self-authorship. The concept is rooted in constructivism (i.e., how individuals make meaning of their experiences and construct their realities) and developmentalism (i.e., how individuals evolve through distinct, increasingly complex phases over the lifespan). Kegan argued that self-authorship is a key developmental

process involving meaning making across three domains of human development: thinking, feeling, and relating to others. Individuals develop an integrated understanding of the socially constructed existence of knowledge (cognitive development), an internally defined sense of self (intrapersonal development), and an approach to relationships that is reciprocal and supportive (interpersonal development; Kegan, 1994). Reflecting these three developmental dimensions of self-authorship, three questions underlie the type of developmental considerations related to this concept: How do I know? (cognitive), Who am I? (intrapersonal), and How am I in relationships with others? (interpersonal; RISE, 2017a).

Dissonance is the key catalyst for self-authorship development, whereby one's meaning making becomes increasingly more complex (Kegan, 1994). However, students' cultural background matters relative to how they know, develop a sense of self, and understand relationships with others (Shweder, Goodnow, Hatano, LeVine, Markus, & Miller, 1998). As such, "there is a need to further specify the current understandings of the catalysts and processes involved in self-authorship development" (Pizzolato, Nguyen, Johnston, & Wang, 2012, p. 656). How cross-culturally relevant are these? This has implications for the cross-cultural applicability of self-authorship development (i.e., relationships between the three developmental domains and cultural variability relative to catalysts for development) and requires an examination of the role of context and relationships in self-authorship development (Pizzolato et al., 2012).

King and Baxter Magolda (2005) applied Kegan's (1994) conceptualization to college students' social-cultural development. Considering the globalized contexts of U.S. postsecondary education, they advanced the notion of intercultural maturity, a multidimensional framework that describes the processes whereby students develop intercultural awareness in

understanding others, themselves, and in their interactions. Such a notion advances a holistic explanation for how college students progress toward intercultural competency outcomes. Intercultural maturity extends Kegan's theory in that the concept of *maturity* underscores the developmental progression of students across cognitive, intrapersonal, and interpersonal developmental domains. Specifically, development moves from awareness toward application in different contexts, where the application of one's development serves to illustrate the achievement of particular intercultural outcomes (King & Baxter Magolda, 2005).

Intercultural Communication

The concept of intercultural communication competence (Chen & Starosta, 1996) also serves as a key theoretical underpinning of the GPI's items. This concept informs the three remaining scales of the GPI (i.e., the Cognitive Knowledge, Intrapersonal Affect, and Interpersonal Social Interaction scales). Chen and Starosta advanced the triangular model of intercultural communication competence, which involves intercultural awareness (cognitive), intercultural sensitivity (affective), and intercultural adroitness (interpersonal) dimensions. Thinking, feeling, and relating are essential dimensions of communicating across cultures. The intercultural communication competence model explains that those with intercultural competence "must possess the capacities of knowing their own and their counterparts' cultural conventions, demonstrating a positive feeling of acknowledging, respecting, and even accepting cultural differences, and acting appropriately and effectively in the process of intercultural interaction" (Chen, 2014, p. 19).

Theoretical Interrelatedness of the GPI's Developmental Dimensions

The cultural development and intercultural communication theoretical perspectives that inform the GPI explicitly describe distinct yet theoretically related cognitive, intrapersonal, and interpersonal domains of development. In particular, both theoretical perspectives explain that development in one domain often spurs development in the other domains. For instance, King and Baxter Magolda (20005) explain that within U.S. postsecondary education

> the developmental complexity that allows a learner to understand and accept the general idea of difference from self without feeling threat to self enables a person to offer positive regard to others across many types of difference, such as race, ethnicity, social class, gender, sexual orientation, and religion. Without this foundation, students may be able to learn about cultural differences; however, the [intercultural maturity] model suggests that they will find it difficult if not impossible to use this knowledge in an intercultural interaction. In other words, less complex levels of cognitive and intrapersonal (identity) development may hinder one's ability to use one's intercultural skills. (pp. 572-573)

Chen and Starosta (1996) also underscore that the three developmental dimensions of their intercultural communication competence model are not developed in isolation; rather—and in alignment with the tenets of intercultural maturity and self-authorship—development in one area relates to development in others. The interplay between these developmental domains is explained by Chen (1997) in terms of intercultural awareness (cognitive dimension) representing the foundation of intercultural sensitivity (affective dimension), which in turn promotes intercultural competence (behavioral, or interpersonal dimension).

Considered together, both overarching theoretical perspectives suggest interrelatedness between the GPI's developmental dimensions. Earlier empirical work has examined the extent of interrelatedness between the GPI's cognitive, intrapersonal, and interpersonal developmental domains. In particular, recent findings suggested that the GPI's six developmental dimensions

are related to each other as theoretically expected *and* that these six dimensions relate to a single, higher-order global perspective development construct (Davidson & Engberg, under review). Each of the GPI's six developmental dimensions is described in the following section with particular attention to its theoretical relatedness to the other areas of development.

The Six Developmental Dimensions of the GPI

Cognitive Development

The GPI includes two cognitive developmental dimensions. The Cognitive Knowing scale is rooted in cultural development theories and measures individuals' ability to evaluate sources of knowledge and ascribe value to these within different cultural contexts. The Cognitive Knowledge scale is rooted in intercultural communication theory and measures individuals' understanding of different cultures and how these are situated in society.

Cognitive Knowing scale. Theoretically, this particular GPI dimension relates to the ways in which individuals *think about* issues related to difference and the complexity of their thinking. Kegan (1994) suggests that one's knowing involves "meaning-constructive or meaning-organizational capacities" that are inherently connected to one's affective and relational spheres of life (p. 29). Kegan argues that the broader concept of *ways of knowing* is derived from a constructivist perspective. Importantly, Kegan underscores the increasingly complex ways that individuals construe their experiences over their lifespan. The construction of one's experiences involves cognitive, affective, and interpersonal considerations. Analyses of such constructions emphasize examining what Kegan refers to as the form or complexity of the organization, not the actual content of thoughts, feelings, and relating (i.e., the focus is on *how* one thinks or feels as opposed to *what* one thinks or feels).

King and Baxter Magolda (2005) argue that "since there is cognitive complexity in the presence of diverse worldviews, accepting ambiguity and understanding the basis of differing worldviews require complex thinking skills" (p. 577). They draw from various theorists' work in explaining increasingly complex cognitive and epistemological development within their intercultural maturity model. The cognitive dimension of the intercultural maturity model relates to the way students think about and understand issues related to diversity (King & Baxter Magolda, 2005). The initial phase of cognitive development within this model is observed when knowledge is viewed as certain and when the right/wrong evaluation of knowledge prevents one's learning about or acceptance of different perspectives (King & Baxter Magolda, 2005). The intermediate phase of development within this model is observed when the view of knowledge as certain shifts to a view that eventually realizes the uncertainty involved in any knowledge claim. Such uncertainty influences one's openness toward different perspectives, and as one makes sense of knowledge claims, one acknowledges the legitimacy of differing views across different individuals (King & Baxter Magolda, 2005). Finally, the mature phase of cognitive development within the intercultural maturity model is observed when one views knowledge as constructed and contextual.

Challenges in measuring students' epistemological development. The theoretical scope of this particular GPI scale relates to the documented difficulty of measuring students' epistemic beliefs via self-report surveys (see DeBacker, Crowson, Beesley, Thoma, & Hestevold, 2008 for a review). Scholars have examined the conceptual challenges and dimensionality (Buehl & Alexander, 2001; Duell & Schommer-Aikins, 2001) and psychometric properties that fail to validate theorized notions of students' epistemological development, highlighting

particular challenges related to differentiating epistemic beliefs from beliefs about learning (DeBacker et al., 2008).

In addition to these conceptual considerations relative to epistemological development, there are also methodological considerations that pertain to the present study. Five of the seven items on the GPI's Cognitive Knowing scale are negatively worded (and are thus reverse scored). The practice of using such survey items has typically been used to mitigate respondents' acquiescence response bias (i.e., an inclination to agree with all item statements regardless of their content) and to encourage respondents to pay careful attention to the survey items when motivation to provide thoughtful responses may be compromised (Barnette, 2000). However, earlier work has suggested that negatively worded survey items lowered internal consistency and mean scores (Schriesheim & Hill, 1981), resulted in different factor structures (Knight, Chisholm, Marsh, & Godfrey, 1988; Pilotte & Gable, 1990), and were related to more inconsistent responses for adults with lower levels of educational attainment (Barnette, 1996; Melnick & Gable, 1990). The compromised reliability and validity of scores from instruments that use mixed (i.e., directly and negatively worded) items emerges because the negatively worded items are often not understood as the actual opposite of directly worded items (Barnette, 2000). This particular issue may help explain the historically and consistently lower internal consistency level observed for the GPI's Cognitive Knowing scale and the results related to the relationship between the GPI's cognitive factors when imposing a hierarchical factor structure on the GPI's global perspective development items (see Davidson & Engberg, under review).

Cognitive Knowledge scale. Theoretically, this particular GPI dimension is rooted in intercultural communication theory and relates to individuals' intercultural awareness. The

concept of intercultural awareness involves individuals' ability to understand intercultural similarities and differences (Fritz, Möllenberg, & Chen, 2001). Hanvey (1987) explains three levels of intercultural awareness: partial or superficial awareness of stereotypical cultural traits, understanding that others' cultural traits can differ substantially from our own, and awareness of how other cultures feel from their own perspectives.

As intercultural awareness develops, cultural conflict—realizing others' cultural norms differ from our own—surfaces, which relates to a host of negative emotional responses (i.e., anxiety, hostility, helplessness, disorientation, and withdrawal; Chen & Starosta, 1998). However, if individuals can work through these emotional responses, they can eventually make understand cultural differences in ways that engender positive emotional responses (i.e., respect, appreciation, and sensitivity toward difference; Chen & Starosta, 1998). Such a process illuminates the interrelatedness between cognitive and affective dimensions of development. As new information (knowledge) elicits emotional responses, one's ability to negotiate such affective responses also relates to different, more nuanced intercultural understandings.

Intercultural awareness involves two components, self-awareness and cultural awareness (Fritz et al., 2001). Hunter, White, and Godbey (2006) argue that "the most critical step in becoming globally competent is for a person to develop a keen understanding of his or her own cultural norms and expectations" (p. 279). Part of intercultural awareness development requires individuals' self-awareness that they are "cultural beings" and to apply such a notion to understanding others' cultural traits (Chen & Starosta, 1998, p. 30). It is one's understanding of one's own culture that serves as a foundation that allows for the recognition and eventual understanding of different cultures (Bennett, 2009).

Intrapersonal Development

The GPI's global perspective development construct also includes two intrapersonal developmental dimensions. The Intrapersonal Identity scale is rooted in cultural development theories and measures awareness and acceptance of one's identity during intercultural exchange. The Intrapersonal Affect scale is rooted in intercultural communication theory and measures the extent to which individuals respect and accept cultural differences and their emotional awareness surrounding cultural difference.

Intrapersonal Identity scale. Theoretically, this GPI dimension relates to individuals' awareness and acceptance of their identities as they interact across cultural differences. Kegan (1994) argues that intercultural competence is only possible when an *internally defined* sense of self exists to mitigate the emotional threat of interacting across difference. The intrapersonal dimension of the intercultural maturity model also addresses the integration of students' values and beliefs into how they live their lives, the ways in which students understand their social identities, and the extent to which they rely on others for self-definition (King & Baxter Magolda, 2005). This developmental dimension emerged from a wide body of general (e.g., Chickering & Reisser, 1993; Kegan, 1994) and particular (e.g., racial and ethnic identity development models, including Cross, 1991; Helms, 1995; Phinney; 1990) identity development research. Considered together, these theories largely posit that individuals transition from a lack of awareness about their identities to "a complex, internally defined perspective on how one's [various identities] are integrated into one's view of oneself and the world" (King & Baxter Magolda, 2005, p. 578).

In the intercultural maturity model, the initial phase of intrapersonal development is described as a lack of awareness about one's social identities, an externally defined identity, an unexamined endorsement of cultural beliefs, values, or behaviors, and perceived threat from cultural values or social identities different from one's own (King & Baxter Magolda, 2005). The intermediate phase of development within this model is observed as a tension between internally and externally derived senses of self. This phase involves students simultaneously attempting to understand their own experiences within their immediate cultural contexts and also how their immediate cultural contexts are situated in broader social contexts (King & Baxter Magolda, 2005). The mature phase of intrapersonal development within the intercultural maturity model is described as a sense of self that involves an integration of the different aspects of one's identity; such an integration arises from a variety of diverse perspectives and experiences that inherently involve considerations around cultural sensitivity and related to intercultural exchange (King & Baxter Magolda, 2005). During this phase, students seek opportunities that challenge their perspectives; they are no longer threatened by divergent views or beliefs. Identity development parallel to this stage has been linked to a variety of student outcomes, including an increased ability to analyze social dynamics in coursework (Howard-Hamilton, 2000) and more confidence in acting as a social justice ally since the emotional threat of representing the interests of minoritized groups is no longer present in this stage (Broido, 2000).

Intrapersonal Affect scale. Theoretically, this GPI dimension relates to individuals' respect for and appreciation of difference and their emotional awareness in interacting across difference. The affective dimension of intercultural communication is often explained in terms

of intercultural sensitivity, which involves individuals' "desire to motivate themselves to understand, appreciate, and accept differences among cultures" (Chen & Starosta, 1998, p. 231). While the concept of intercultural sensitivity is related to cognitive, intrapersonal, and interpersonal developmental domains within intercultural contexts, the affective domain is of primary focus (Chen, 1997). In particular, to develop positive emotional responses toward different cultures, interculturally sensitive individuals must possess self-esteem, self-monitoring, open-mindedness, empathy, interaction involvement, and non-judgment, underscoring the multidimensionality of this concept (Chen, 1997). These dimensions are discussed in more detail below.

The self-worth and confidence related to higher levels of one's self-esteem relate to intercultural sensitivity because they equip an individual with tools to more effectively negotiate the alienation and stress sometimes involved in intercultural communication efforts (Chen & Starosta, 2000). During intercultural interaction, those who self-monitor at higher levels (i.e., detect situational aspects and align their behaviors accordingly) tend to employ more sensitivity toward others and adjust their communication style accordingly (Chen & Starosta, 2000). Open-mindedness involves a willingness to identify, accept, and appreciate divergent viewpoints, which cultivates the consideration (or sensitivity) of others' needs and differences (Bennett, 1986; Chen & Starosta, 2000). Several scholars (e.g., Bennett, 1986; Chen & Starosta, 1997; Gudykunst, 1993) argue that empathy comprises the core component of intercultural sensitivity. Higher levels of concern for others' feelings and experiences, accuracy in understanding another's internal state, and affective orientation and understanding relate to higher levels of intercultural sensitivity (Chen & Starosta, 2000). Interaction involvement—an individual's

sensitivity during interactions—is explained in terms of responsiveness, attentiveness, and perceptiveness. The higher the intercultural sensitivity, the more an individual can understand and behave in intercultural exchanges in ways that facilitate positive interactions. Finally, non-judgment relates to individuals' ability to listen to culturally different information without prematurely concluding meaning; such a quality facilitates intercultural exchange and relationships (Chen & Starosta, 2000).

Interpersonal Development

The GPI's global perspective development construct also includes two interpersonal developmental dimensions. The Interpersonal Social Responsibility scale is rooted in cultural development theories and measures one's inclination to operate interdependently and with a concern for others. The Interpersonal Social Interaction scale is rooted in intercultural communication theory and measures one's proclivity to interact across difference.

Interpersonal Social Responsibility scale. Theoretically, this GPI dimension relates to one's ability to relate to others "in terms of moving from dependency to independence to interdependence, which is considered as the most mature perspective in effectively living in a global society" (RISE, 2017a). The interpersonal developmental dimension of the intercultural maturity model addresses students' ability to effectively and collaboratively interact across difference and negotiate the need for others' approval in relating across difference. The eventual goal is for students to develop the ability to relate to others in ways that acknowledge an understanding and respect for different perspectives but that also incorporate their own beliefs and values (King & Baxter Magolda, 2005). Students move from an individualistic, to a relativistic, to an eventual orientation that acknowledges the role of social systems in

intercultural relations, which is necessary in understanding the realities of minoritized groups (King & Baxter Magolda, 2005).

During the initial stage of interpersonal development in the intercultural maturity model, relationships are largely rooted in students' primary social identity groups and often involve "egocentric standards to judge cultural differences" (King & Baxter Magolda, 2005, p. 580). There is little attribution of social or structural explanations for phenomena or more abstract thinking about the causes and consequences of intercultural relations. The intermediate phase of this model involves an increased capacity to understand the causes and consequences of intergroup differences and to interact across difference. An understanding of social systems—and their constructed nature—emerges at this stage; however, openness to different perspectives is attenuated due to reliance on others' approval relative to how one perceives appropriate beliefs and actions (King & Baxter Magolda, 2005). Finally, the mature phase of interpersonal development within the intercultural maturity model is observed when interaction across difference is viewed as educative and additive to students' understanding of themselves and their place in a broader society instead of as a threat to their identity (King & Baxter Magolda, 2005).

Interpersonal Social Interaction scale. Theoretically, this GPI dimension relates to one's tendency to interact across cultures. Chen and Starosta (1996) advanced the behavioral component of their intercultural competence model by explaining that interculturally competent individuals learn to interact effectively across cultural differences. More specifically, this dimension involves intercultural adroitness, or one's ability to execute particular tasks or achieve certain communication goals across cultural differences (e.g., linguistic skills, managing

interactions, appropriate levels of self-disclosure, flexibility in interacting, and social skills such as empathy and perspective taking; Chen, 2014).

Chen and Starosta's (1996) concept of intercultural adroitness was influenced by the earlier behavioral approach of Ruben (1976), who sought to examine how knowledge of competent intercultural behaviors translated into observable behaviors in intercultural situations. Ruben advanced seven dimensions of intercultural competence: display of respect, interaction posture (i.e., one's ability to respond to another in a descriptive, non-threatening way), orientation to knowledge (i.e., understanding the relativity of knowledge), empathy, self-oriented role behavior (i.e., balancing requests for information with group dynamics), interaction management (i.e., ability to take turns in discussion based on need), and tolerance for ambiguity (i.e., ability to react to new or unpredictable situations without discomfort).

Extending from this theoretical explanation of the GPI's six developmental dimensions, the following discussion reviews the literature on cultural variability related to each of these six areas of development. In particular, the following discussion provides in-depth theoretical rationale to justify examining the GPI's measurement invariance across ethnoracial groups.

Cultural Variability Related to the GPI's Developmental Domains

Milem, Chang, and Antonio (2005) explain that the racialized contexts of U.S. postsecondary education influence the study of educational outcomes involving interaction across difference. Extending from this, it is necessary to comprehend the ways in which ethnoracially minoritized and white students *understand* interacting across difference, intercultural knowledge, sensitivity toward difference, and their own identities differently. The

following discussion reviews the literature on the culturally variable nature of the particular types of student development the GPI is intended to measure.

Cultural Variability Relative to Epistemological Development

The theoretical scope of the GPI's Cognitive Knowing scale is quite complex. Broadly, this scale is theorized to measure individuals' ability to recognize the role of cultural context in how they know (i.e., emphasizing the complexity of how they know; RISE, 2017a). As such, I review the literature on epistemological development and its cultural variability in this section. Epistemological beliefs entail "how individuals come to know, the theories and beliefs they hold about knowing, and the manner in which such epistemological premises are a part of and an influence on the cognitive processes of thinking and reasoning" (Hofer & Pintrich, 1997, p. 88). However, individuals are nested within broader social contexts; as such, epistemology is also understood as a "*system* of knowing" shaped by individuals' worldviews that are constructed given their contexts (Delgado Bernal, 2002, p. 106, emphasis added). Scholars argue that notions of epistemological development, as well as the processes through which such development unfolds, are culturally variable (Broido & Schreiber, 2016). Hofer (2008) argues for an expanded understanding of developed epistemologies across cultures and suggests that additional dimensions may be necessary to more precisely conceptualize what more sophisticated stages of reasoning resemble in different cultural contexts.

For the sake of precision for this discussion, a close examination of this GPI scale's seven items suggests a multidimensional understanding of epistemological development that involves (1) ethnocentrism, (2) the sources and justification of knowledge dimension of epistemological development, and (3) the development of a pluralistic orientation. Considered together, the

theoretical scope of this particular GPI scale relates to the particular cultural variability issues discussed here.

Ethnocentrism. Ethnocentrism is a judgment of other cultures using one's own culture as the ideal or reference point. Ethnocentrism prevents individuals from understanding cultural differences. In addition to constructing our own cultural norms, we construct the *difference* between ours and others' norms (Kegan, 1994). Ethnocentrics make attributions about others based on difference rather than examining the difference between their norms and others' (Kegan, 1994). Though the GPI is not theoretically derived from the developmental model of intercultural sensitivity (DMIS; Bennett, 1986), this particular model aligns with the theoretical underpinnings of the items on the GPI Cognitive Knowing scale that relate to ethnocentrism. For instance, the DMIS—also rooted in constructivism—explains how individuals make meaning of cultural difference. The model represents increasingly more complex worldview structures that undergird individuals' intercultural attitudes and behaviors. The DMIS includes six worldview orientations. The first three worldviews of denial, defense, and minimization are theorized as ethnocentric; they explain individuals' strategies for avoiding cultural difference (i.e., by denying that difference exists, raising defenses against such difference, or by minimizing the importance of the difference). The last three DMIS orientations of acceptance, adaptation, and integration are theorized as ethnorelative; these explain individuals' inclination to seek cultural difference (i.e., by accepting the importance of difference, adapting their perspectives to consider difference, or by integrating the concept of difference into how they define identity; Hammer et al., 2003).

Hammer et al. (2003) argue that two of the DMIS's ethnocentric worldviews are particularly subject to cultural variation. The second ethnocentric orientation of the DMIS—defense against difference—suggests difference is negotiated in terms of one's *defense* against it (i.e., denigrating another's culture, uncritically viewing one's own culture as superior, and/or a reversal of the defense, where others' cultures are viewed as superior while one's own is denigrated). Within this orientation, clear boundaries that define one's own and others' cultures are drawn. Such boundaries promote an understanding of a cultural hierarchy and enable one to place one's own culture as the epitome of all cultures (Bennett, 1986). Individuals from dominant cultures are inclined to experience the defense orientation as "an attack on their values (often perceived by others as privileges)," while those within non-dominant cultures are more inclined to experience this same orientation with the purpose of "discovering and solidifying a separate cultural identity in contrast to the dominant group" (Hammer et al., 2003, p. 424). The third ethnocentric orientation of the DMIS—minimization of difference—involves recognizing commonalities across all individuals and avoidance of recognizing nuance relative to one's culture or others' realities. Relative to the third DMIS orientation of minimization, for those from dominant cultures, this worldview tends to obscure recognition of their own culture (e.g., ethnicity) and the inherent systematic privilege extended to those in the group (Hammer et al., 2003).

An alternative theoretical explanation for ethnoracial variability relative to the items measuring ethnocentrism (i.e., *When I notice cultural differences, my culture tends to have the better approach)* involves the development of an oppositional social identity by ethnoracially minoritized individuals. Ogbu and Simons (1998) explain such development:

> In responding to their forced incorporation into U.S. society and their subsequent mistreatment, [ethnoracially minoritized individuals] develop a collective identity defined to a great extent by its difference from and opposition to white American identity. Given this interpretation, some individuals feel that if they learn white American ways or 'white talk' they will lose their minority identity. For them, adopting white ways and language is a subtractive or replacement process that threatens minority identity and therefore is resisted. (p. 175)

It is possible for a respondent to reply to this item in a manner that reflects their endorsement of their own culture as familiar and therefore normative (Kegan, 1994), which represents a lower level of cognitive development. However, instead of reflecting one's orientation toward ethnocentrism, it is also quite possible for a marginalized respondent to reply in a way that reflects their endorsement of a protective, developmental orientation toward opposing the oppressive forces faced in everyday life by minoritized individuals (e.g., *my culture tends to have the better approach*). These two hypothetical respondents' divergent understandings reflect markedly different levels of complexity relative to interacting across difference. The former's response would indicate a lower level of complexity (i.e., an unawareness, denial, or apathy toward difference), while the latter's response would suggest a much higher level of development (i.e., the development of an oppositional social identity *requires* an increasing awareness of one's experiences with oppressive forces; Fordham & Ogbu, 1986).

Sources and justification of knowledge. This dimension of epistemological development is subject to cultural variability due to divergent socialization around the involvement of authorities in one's life, collectivist orientations, and the invalidation of knowledge based on social location. Roberts and Komives (2016) explain that socialization within cultures where "deference to authorities, personal humility, and affiliation with the group is the norm" runs counter to widely accepted Western pedagogical practices (p. 17). Chan and

Elliot (2004) argue that the notion of *authority* in Chinese culture, for instance, connotes a sense of reverence and admiration toward elders and experts; students in such a cultural context would certainly internalize a particular way of structuring their personal epistemic beliefs.

Schommer-Aiken's (2004) cultural-relational work involves the role of self-construal (i.e., how one defines one's self and whether one is embedded within an interdependent or independent cultural context) and its influence on personal epistemologies. Schommer-Aikens argued that students' natural tendencies in social interactions likely parallel the way they approach knowledge acquisition and sources of knowledge (i.e., student-instructor relationships, the value of knowledge handed down from authorities). Tasaki's findings (2001) also underscored the relationship between students' independent or interdependent self-construal and the type of epistemological beliefs with which they resonate. For instance, Tasaki found significant relationships between interdependent self-construal and epistemological beliefs that measured omniscient authority, certainty of knowledge, rigid learning, and innate ability. And conversely, Tasaki found that students with an independent self-construal were more likely to view knowledge as uncertain and dynamic and less likely to endorse beliefs in omniscient authority. Tasaki's findings suggest that the epistemological beliefs advanced through so much of Western postsecondary education (i.e., understood as nested within a primarily *independent* cultural context) may be biased against cultures that define their epistemological beliefs within a more interdependent context. In particular, in examining individualistic and collectivist cultures, Youn (2000) also suggested that self-beliefs relative to learning develop differently in these two cultures. Wong and Chai (2010) explain that various studies of the epistemological development of Asians have suggested differences in the structure of epistemic beliefs across various ethnic

groups. Scholars have argued that the roles of economic development, historical evolution, and cultural diversity (Chai, Hong, & Teo, 2009) as well as social conservatism and religion (Wong & Chai, 2010) relative to individuals' personal epistemological beliefs must be considered in this research.

In addition to the individualistic or collectivist cultural context, students' social location within postsecondary education also figures prominently into this dimension of epistemological development. Many students of color report experiencing an invalidation of their knowledge based on their social location. For instance, students of color can feel that their experiences are misunderstood because they do not represent "an acceptable source of knowledge from which to draw on in academic settings" (Delgado Bernal, 2002, p. 106). Put differently, people of color must negotiate assimilation or acculturation into a white world, where dominant ways of knowing are often imposed on them (Goldberger, 1996). They are silenced by systemic oppression and inequities (Bing & Reid, 1996) and can struggle to develop their voice given the complexities of their identities (Hurtado, 1996). Given these findings, it can be expected that one's social location—in particular, within the ethnoracial hierarchy in the U.S. and its institutions of higher education—influences one's understanding of the GPI items that measure this dimension of development. It is theoretically possible that an ethnoracially minoritized respondent could comprehend this item in a way that reflects their internalized silence relative to sources of their knowledge.

Pluralistic orientation. Two items on this GPI scale ask students about the extent to which they consider different perspectives in evaluating their broader contexts. Engberg and Hurtado (2011) define pluralistic orientation as "higher levels of complex thinking that enable

students to engage in cooperative behaviors, manage controversial issues, and develop a high regard for others' perspectives, beliefs, and backgrounds" (p. 417). Several scholars have reported differences in undergraduates' pluralistic orientations across ethnoracial groups. For instance, in a longitudinal study, Engberg and Hurtado found that undergraduates reported similar self-assessments of their pluralistic orientations across ethnoracial groups when entering college, but after two years, the Asian students in their sample reported lower self-assessments compared to their peers. Engberg and Hurtado attributed this particular finding to a possible accentuation effect. However, they also discussed earlier work that suggests that relative to what their pluralistic orientation measure involved (i.e., conflict, challenge, and difference), Asian students have demonstrated lower levels of assertiveness and self-efficacy in novel interactions (Zane, Sue, Hu, & Kwon, 1991) and preferences for collectivism, respect for authority, and conformity to social norms (Kim, Atkinson, & Yang, 1999). This is particularly interesting given that Engberg and Hurtado (2011) obtained strong evidence for measurement invariance for their pluralistic orientation construct (as well as the other latent constructs in their models) across the four ethnoracial groups in their sample, suggesting that students in their sample understood their pluralistic orientation measure similarly.

Using latent class analysis to classify groups of entering undergraduates based on their pluralistic orientation, Denson and Ing (2014) found that the probability of being classified into one of the four pluralistic orientation groups in their model was dependent on a student's race or ethnicity, with Asian and Hispanic students reporting different likelihoods than white students to be classified into particular groups. Engberg, Meader, and Hurtado (2003) examined a structural model of pluralistic orientation that revealed differences between white and non-white

undergraduates. These differences were most notably related to these students' peer interactions, which the authors reported as related to both direct and indirect effects on students' pluralistic orientation scores. These findings relate to previous research that has demonstrated that while involvement in structured diversity experiences (e.g., courses, workshops, cultural events) generally relates to positive outcomes for all participants, differences exist across ethnoracial groups (Engberg, 2004; Spanierman, Neville, Liao, Hammer, & Wang, 2008). Finally, Engberg (2007) found that the ethnoracial diversity of the institutions in his sample had an indirect effect on students' second-year pluralistic orientation that was attributed to positive intercultural interactions. Considered together, these findings suggest that undergraduates' pluralistic orientation—part of what this GPI scale measures—differs across groups of students. But the mechanisms to explain such variation seem to also operate differently across ethnoracial groups.

Cultural Variability Relative to Intercultural Understanding and Awareness

Intercultural awareness involves the development of both self-awareness and a broader awareness of other cultures (Fritz et al., 2001). Several elements relate to the development of both types of awareness, and these elements are experienced differently across ethnoracial groups. The following discussion outlines elements that theoretically relate to the development of intercultural understanding and awareness. In particular, I discuss the role of mono- and intercultural socialization and intergroup interaction and how these vary across ethnoracial groups. Considered together, these findings provide a rationale for examining the extent to which this particular developmental dimension could be understood differently across ethnoracially diverse undergraduates.

Socialization and cultural worldview development. To achieve intercultural understanding, one must first be able to employ a high level of self-awareness as this relates to one's own cultural background (Yan & Wong, 2005). This involves cultivating an awareness of one's own cultural values, biases, and stereotypes (Sue & Sue, 1990) as well as cultural identification and cultural influences (Lum, 1999). Extending from this cultural self-awareness, the development of intercultural awareness involves increasingly more nuanced understandings of aspects of other cultures (i.e., histories, values, politics, economics, customs; AAC&U, 2017). A cultural worldview is understood as the lens through which individuals perceive these cultural elements and how they construe meaning relative to them (AAC&U, 2017). The extent to which individuals are socialized around cultural difference influences their ability to cultivate self- or intercultural awareness, and therefore to *understand* such difference. Hammer et al. (2003) describe this as follows:

> Individuals who have received largely monocultural socialization normally have access only to their own cultural worldview, so they are unable to construe (and thus are unable to experience) the difference between their own perception and that of people who are culturally different. The crux of the development of intercultural sensitivity is attaining the ability to construe (and thus to experience) cultural difference in more complex ways. (p. 423)

Undergraduates' mono- or intercultural socialization prior to college relates to how they comprehend the process of understanding cultural difference. Members of ethnoracially minoritized groups in K-12 education often are required to balance the customs, beliefs, and behaviors associated with two different cultures, to seek success in schools dominated by another culture, and to negotiate the realities of their minoritized status and racism (Hamm & Coleman, 2001; Spencer & Dornbush, 1990). In contrasting the experiences of what he terms traditional

(e.g., white, Euro-American, middle-class, male, heterosexual) and nontraditional students (e.g., students of color, women, international students, gay, lesbian, bisexual), Sedlacek (2003) explains that "nontraditional people must learn to be 'multicultural' and examine issues from different perspectives" (p. 265). By contrast, many white adolescents—by nature of representing the majority in many educational contexts—are not required or encouraged to develop strategies for interethnic peer interactions; this task is often understood to be the required work solely of ethnoracially minoritized students (Fine, Weis, Powell, & Wong, 1997). Further, academic tracking in high schools that often segregates students by race and ethnicity (Phelen, Yu, & Davidson, 1994) and many white parents' lack of intercultural socialization strategies with their children, namely in predominantly white social contexts (Frankenberg, 1993), often relates to white students' general unawareness and avoidance of peers from other ethnic groups (Hamm & Coleman, 2001).

In considering the role of socialization relative to the development of intercultural awareness, the concept of ethnocentrism once again figures prominently into this discussion. In particular, Sue (2004) describes ethnocentric monoculturalism on the part of whites as "the invisible veil of a worldview that keeps White Euro Americans from recognizing the ethnocentric basis of their beliefs, values, and assumptions" (p. 764). Embedded within a monocultural worldview, individuals' unchallenged meanings ascribed to aspects of life become taken-for-granted realities.

Ethnocentric biases are culturally conditioned (Neuliep, Chaudoir, & McCroskey, 2001). In describing the "invisibility of ethnocentric monoculturalism," Sue (2004) explains that whiteness "represents institutional normality, and White people are taught to think of their lives

as morally neutral, average, and ideal" (p. 764). Given that this GPI scale asks about, among other things, one's understanding of how various cultures interact socially and one's ability to discuss cultural differences from an informed perspective, the concept of ethnocentric monoculturalism suggests that many white students (i.e., those socialized in monocultural contexts) may initially deny cultural difference and fail to understand the consequences of doing so (Sue, 2004).

Intergroup interaction. Intergroup contact theory (Allport, 1954) explains that under the appropriate conditions, contact with others culturally different from one's self is effective in reducing intergroup prejudice. This is so because intergroup communication allows for an *understanding*—even appreciation—of diverse perspectives. Key in Allport's hypothesis was that intergroup contact creates opportunities to learn about another cultural group, reduces anxiety and fear of out-group members, and increases one's ability to take the perspective of and empathize with those culturally different from one's self.

Postsecondary educational research has suggested that as structural diversity (i.e., the numerical composition of diverse groups of students) increases, so do opportunities for students to interact with diverse peers (Chang, 1999; Chang, Astin, & Kim, 2004; Gurin, 1999), though the type and quality of those interactions as well as the campus climate are important considerations (Milem et al., 2005). Salient to this particular discussion, undergraduates' cross-racial interactions have positive effects on their development of *cultural awareness* (Astin, 1993; Chang, 2001; Hurtado, 2003; Saenz, Ngai, & Hurtado, 2007). Interactions with others who are culturally different allow one to observe different cultural conventions and contrast those with one's own; such cognitive engagement encourages reflection on otherwise taken-for-granted

elements and a deeper examination of one's beliefs and assumptions (King, Perez, & Shim, 2013). In turn, such thinking promotes the development of more nuanced intercultural awareness. However, King et al. found that students' perceived sense of safety within their learning contexts mediated their willingness to engage and learn from cultural differences. But the notion of a "safe" learning context is variable across individuals. For instance, a wide body of literature discusses that compared to their white peers, many ethnoracially minoritized students report feeling more isolation, distress, mistrust, harassment, discrimination, and racial microaggressions in both academic and social postsecondary contexts (e.g., Jay & D'augelli, 1991; Sedlacek, 1999; Solórzano, Ceja, & Yosso, 2000).

Even in structurally diverse secondary educational contexts, racism, discrimination, and pronounced intergroup inequities often offset students' interracial interactions prior to college (Wells, Holme, Rivilla, & Atanda, 2009). Institutional-level factors within secondary educational contexts (i.e., determinants of students' academic placement, racialized academic or extracurricular activities, maintenance of whiteness through staff hiring and student recruitment, structural diversity of the student body) negatively relate to students' ability to cross cultural boundaries, or what Carter (2010) refers to as cultural flexibility. Students with cultural flexibility effectively navigate diverse social settings because they seek multiple viewpoints in constructing their understandings of themselves and their contexts, and they prioritize inclusive ways of operating within their contexts (Carter, 2010). Educational research suggests that positive intergroup interaction promotes intercultural awareness through the acquisition of increasingly complex knowledge (i.e., awareness), skills, (i.e., empathy, communication), and attitudes (i.e., openness, curiosity; AAC&U, 2017). However, the body of research cited above

also illuminates that students across ethnoracial groups do not interact across difference or construe their interaction across difference identically.

Cultural Variability Relative to Identity Awareness and Integration

Theoretically, the GPI's Intrapersonal Identity scale relates to individuals' awareness and integration of their identities in their interactions across difference. This scale does not measure particular types of identity development (e.g., racial, ethnic, gender, or other particular social identities). Rather, this scale measures students' sense of their identities more broadly and their perceived ability to align their identities with particular types of insights (i.e., their sense of purpose, their self-understanding) and actions (i.e., explaining their values to others, defending their beliefs). As such, the following discussion focuses on variability related to these dimensions of identity.

Scholars have explained that although there is generally an international recognition of the need to understand myriad aspects of college students' identity development, the ways in which such development has been theorized have not been executed cross-culturally (Broido & Schreiber, 2016). For instance, the structure of identity development has been shown to differ across cultures (Schwartz, Zamboanga, Meca, & Ritchie, 2012). Though this GPI scale does not measure identity development, these points are relevant because the dimensions of identity contained within these developmental models relate to how one's identity is understood and enacted (i.e., what this GPI scale *does* intend to measure).

Two discussions that follow explain potential cultural variability related to students' awareness and integration of their identities in their interactions across difference. First, related to students' sense of their identities, previous work suggests that individuals in minoritized

groups tend to identify with their group more strongly (i.e., given the threat to their self-concept, they are more motivated to perceive in-group similarity, legitimacy, and solidarity in order to maintain a positive social identity; Simon & Brown, 1987) and attribute more importance to their identities (Phinney & Alipuria, 1990) compared to majority group members. Second—and related to the alignment of students' identities with their actions—the systematic silencing of marginalized groups complicates their ability to explain and defend their principles and values, namely when these diverge from majority views (Applebaum, 2003). Both reasons are discussed below.

Sense of one's identities. In order to understand one's sense of identity, it is useful to ground this discussion in social identity theory. This social psychological theory, advanced by Tajfel (1978), explains social identity as a component of one's self-concept that is derived from membership within particular social groups. Tajfel's concept of social identity also involves the perceived value and emotional attributions of belonging to particular social groups. Since social identities undergird one's self-concept—and individuals aim for a positive self-concept—they are motivated to maintain a positive evaluation of their own social group (i.e., a positive social identity relative to their in-group). This need drives an inherent social categorization process where in- and out-groups are established. The need for positive group distinctiveness—to preserve a positive social identity—results in social comparison between one's in-group and salient out-groups and an understanding of one's in-group as more favorable (i.e., cultivating in-group bias), regardless of whether there exists intergroup conflict.

A key component of social identity theory is an individual's intrinsic need for positive group distinctiveness, requiring both a differentiation between one's group and others and a

positive evaluation of one's group. However, positive group distinctiveness operates differently for ethnoracially minoritized individuals than for whites in several key ways (Dovidio, Gaertner, & Saguy, 2009; Molix & Bettencourt, 2010; Simon & Brown, 1987). First, those who have a stronger group identification tend to demonstrate more motivation to maintain positive group distinctiveness (Mummendey, Klink, & Brown, 2001; Simon, Kulla, & Zobel, 1995). Extant literature suggests that individuals belonging to minoritized groups tend to value the qualities that distinguish their identity group (i.e., the salient dimensions that characterize the social categorization; Tajfel, 1978) more than individuals belonging to majority groups, especially when minoritized individuals perceive status hierarchies and group boundaries as permanent and any resultant group disparities arising from illegitimate reasons (Bettencourt, Dorr, Charlton, & Hume, 2001). Put more simply, members of minoritized groups tend to identify with their group more strongly compared to majority group members (Dovidio et al., 2009; Simon & Brown, 1987).

Second, due to the social stigma and resulting prejudice and discrimination faced by minoritized individuals, they engage in a coping process to preserve (Molix & Bettencourt, 2010) or accentuate (Simon & Brown, 1987) their positive social identity. Whites in the U.S. typically are not stigmatized because of their racial identity; as such, they usually do not need to develop a comparable coping strategy or contend with the same threat to this social identity to maintain a positive self-concept (Simon & Brown, 1987). Finally, those belonging to majority groups are typically motivated to *protect* their collective identity, while those belonging to minoritized groups are typically motivated to *enhance* the group's collective identity (Scheepers, Spears, Doosje, & Manstead, 2006). In summary, these findings explain three mechanisms that

relate to the variability of one's sense of identity across ethnoracial groups: stronger group affiliation for minoritized individuals (i.e., what Simon & Brown (1987) referenced as the perceived enhanced entitativity or groupness experienced by minoritized individuals), different threats to one's positive self-concept based on upheld social stigmatization of particular groups, and different motivations for engaging one's identity based on the social location of that identity.

Recall that this particular GPI scale is conceptually grounded in particular cultural development theories that emphasize the ways in which individuals develop an understanding of their social identities, integrate their values and beliefs into their lives, and rely on others for self-definition (i.e., Kegan, 1994; King & Baxter Magolda, 2005). Using the same theoretical models as King and Baxter Magolda used in advancing the intrapersonal dimension of their intercultural maturity framework, it is clear that cultural variability exists in terms of this developmental dimension. For instance—and of particular import to the present study—ethnic identity is important to most ethnoracially minoritized individuals and has historically lacked the same significance for most whites (Phinney, 1989). This reflects the invisibility of white privilege; developmentally, such privilege is initially perceived as a norm by whites (McIntosh, 2003). Many white students in the U.S. fail to understand whiteness as a racial identity, and they often think racism does not affect them because they are not students of color (Bonilla-Silva & Forman, 2000; McIntosh, 2003; Tatum, 2007). In broader terms, this aligns with pronounced differences observed across various minority and majority identity development models (e.g., Atkinson, Morten, and Sue's (1989) racial/cultural identity development model, Helms' (1995) white racial identity development model). For instance, the initial stage of minority identity development (i.e., Atkinson et al.'s [1989] conformity stage) illuminates the socialization of

minoritized individuals to think of their context and selves in terms of non- or anti-minority with a devaluation of their development of a minority identity; there is an immediate and *acute awareness* of one's minoritized status. Conversely, the initial stage of majority identity development (i.e., Helms' [1995] contact status) illuminates that for majority group members, there exists *obliviousness* to multicultural contexts and to any acknowledgement of their dominant group status or its implications within social systems.

Silenced identities. Applebaum (2003) reminds us that "classrooms and schools represent a 'culture of power' to the extent that they mirror unjust social relations that exist in the larger society," silencing particular groups of students on the margins (p. 151). Part of this GPI scale relates to how individuals perceive their ability to enact parts of their identity. In considering the extent to which dominant and marginalized groups understand their agency relative to enacting parts of their identities, understanding *silencing forces* as a key socializing structure and experience is particularly relevant to this discussion. Silencing forces arise from cultural norms that expect silence on the part of some identities or that altogether fail to acknowledge identities perceived as non-normative (Shapiro, Rios, & Stewart, 2010). Although Shapiro et al. applied this socialization structure to lesbian identity development, I argue that there are parallels to marginalized identities more generally, given that the construct of identity serves as an intermediary between broader social structures and individuals' behavior within those contexts (Hogg, Terry, & White, 1995). Shapiro et al. (2010) explain that silencing is context-dependent, spanning cultural, institutional, and personal levels and is understood as both internal (i.e., one feeling silenced because of one's sense of danger or discomfort) and external to the self (i.e., the social invisibility of particular groups).

McLean (2008) argues that society silences particular narratives and entire groups by explaining this phenomenon as such:

> Silencing refers to the explicit or implicit message that one's stories, and consequently, one's self, are not acceptable, interesting, or relevant, thus rendering one's voice unheard. Therefore, the canonical narrative in a given culture is given privilege and authority over the non-canonical narrative. *Voice* is given to those people who have personal narratives that match the canonical narrative, as their experiences are both socially accepted and assumed. Conversely, those people who cannot identify with the canonical narrative have experiences that are *silenced*. (p. 1695; emphases in original)

In considering the extent to which ethnoracially minoritized individuals feel agentic in explaining deeply held aspects of themselves (i.e., personal values, beliefs, and views) to others who are either culturally different or who hold divergent views, McLean's (2008) argument underscores that this is not a neutral process.

Cultural Variability Relative to the Affective Dimension of Intercultural Exchange

Theoretically, the GPI's Intrapersonal Affect scale relates to feelings of respect and acceptance toward those culturally different from one's self and the extent of one's emotional self-awareness in intercultural exchange. This GPI scale is informed by the theoretical underpinnings of the intercultural sensitivity construct previously explained (Chen & Starosta, 2000). However, evidence of the cross-cultural applicability of Chen and Starosta's conceptualization of intercultural sensitivity has been inconclusive.

For instance, while the theorized five-factor structure of Chen and Starosta's (2000) Intercultural Sensitivity Scale initially demonstrated acceptable fit for both U.S. and German samples (Fritz, Möllenberg, & Chen, 2002, 2003), subsequent cross-cultural validation studies using both U.S. and German samples failed to reproduce these results (Fritz, Graf, Hentze, Möllenberg, & Chen, 2005). And while Peng (2006) examined the reliability of the Intercultural

Sensitivity Scale among a Chinese sample, and Peng, Rangsipahat, and Thaipakdee (2005) used the instrument to compare differences in intercultural sensitivity among Chinese and Thai samples, neither study assessed the cross-cultural validity of this measure. Extending from the inconclusiveness of these earlier findings, Tamam (2010) investigated the factor structure of the Intercultural Sensitivity Scale using a Malaysian sample; these findings suggested a three-factor structure, compared to the theorized five-factor structure. Tamam argues that Chen and Starosta's (2000) conceptualization of intercultural sensitivity "is not a generic model that is culture-free" (p. 181). Tamam (2010) suggests the possibility of conceptual overlap of particular intercultural sensitivity factors in various cultural contexts given cultural values including individuals' baseline intercultural sensitivity (i.e., how individuals are socialized in an incredibly culturally diverse society), communication style (i.e., preferences for indirect or non-confrontational interactions), and collectivist orientation (i.e., group harmony and mutuality in relationships). Considered together, these findings underscore the need to examine theoretical and methodological considerations related to intercultural sensitivity within diverse cultural contexts.

Cultural Variability Relative to Social Responsibility

The GPI's Social Responsibility scale relates to individuals' sense of interdependency and their broader consideration of others' needs. The items on this GPI scale may relate to culturally variable interpretations because cultural value orientations that emphasize collectivism or individualism underlie one's understanding and integration of what this scale measures. Cultural value orientations "shape and justify individual and group beliefs, actions, and goals. Institutional arrangements and policies, norms, and everyday practices express underlying

cultural value emphases in societies" (Schwartz, 2006, p. 139). Individuals' differing orientations around the concept of interdependence and its relation to advancing social change may also explain potential cultural variability relative to the concept of social responsibility. Both of these reasons are discussed below.

Collectivist and individualist cultural value orientations. Earlier work has suggested that cultural value orientations that emphasize collectivism as opposed to individualism relate to observed differences in undergraduates' proclivity to engage in work on behalf of others. For instance, in their examination of undergraduates' socially responsible leadership development, Dugan, Komives, and Segar (2008) observed ethnoracial group differences relative to the citizenship construct used in their study, which measured students' proclivity to work for positive change on behalf of others (i.e., this was one of several dimensions relating to an overarching socially responsible leadership construct). Dugan et al. speculated that observing white students' lower scores on the citizenship scale could represent these students' more individualistic cultural value orientation. Other scholars argue that minoritized groups—because of their marginalized status in many contexts and the resultant uneven distribution of power within these contexts—often must cultivate more relational, collaborative styles of leadership (Madden, 2005; Sanchez-Hucles & Sanchez, 2007). On college campuses, ethnoracially minoritized students often internalize the emphasis within "mainstream" student organizations on "the singular leader" instead of a more collective approach to leading others and the racialized contexts of these student organizations (i.e., as white spaces; Harper & Quaye, 2007, p. 131), which often represent clashes with their cultural values. Socialization around collectivist or individualist cultural values would certainly be expected to shape students' *responses* to these

particular items; this is reflected in observed group differences across ethnoracial categories for these variables (e.g., Dugan et al., 2008). However, such socialization would also be expected to shape their *understanding* of these items. Tourangeau et al. (2000) argue that broader ideologies and values are often reflected in individuals' deeply held principles and, as such, undergird survey respondents' attitudes toward specific issues.

Advancing social change. People of color in leadership roles often explain that their reason for pursuing leadership is to effect social change (Arminio et al., 2000; Harper & Quaye, 2007). This parallels interracial contact research that suggests that marginalized groups often approach intercultural exchange with the aim of changing power structures, while advantaged groups often wish to emphasize intergroup similarities (e.g., Saguy, Dovidio, & Pratto, 2008). For instance, Harper and Quaye (2007) found that each of the African American males in their sample explicitly discussed their commitments to uplifting the African American community (both on campus and beyond), to combatting racial stereotypes, removing barriers, and creating new opportunities for other African Americans on campus. One's inclination to advance social change—and the realities that make this inclination variable across groups—is a serious consideration in understanding how diverse groups prioritize socially responsible practices.

Cultural Variability Relative to Intercultural Interaction

The GPI's Social Interaction scale relates to one's tendency to interact across cultures. Indeed, it is this scale that represents a key behavioral dimension of a given campus's climate (Milem et al., 2005). While a large body of evidence suggests numerous cognitive, psychosocial, and interpersonal benefits of intercultural contact (see Antonio et al., 2004; Gurin et al., 2002 for comprehensive reviews), the effort required to engage across cultures and the

learning and development resultant from intercultural contact are not experienced the same across ethnoracial groups (Tanaka, 2002). As such, it could be expected that ethnoracially diverse samples of undergraduates could understand these particular GPI items—and therefore the particular dimension of student development they relate to—differently.

The methods by which students' social interactions are measured in survey research too often emerge from a dominant discourse that views interpersonal interactions as categorically developmental and nested within culturally neutral contexts (Tanaka, 2002). In particular, Tanaka argues that many of the developmental constructs used in large-scale postsecondary surveys fail to consider "the impact of student racial demographic shifts, issues of power, or the cultural norms that may impact student development" (p. 266). In considering students' interactions across difference, the following discussion outlines key differences related to how students perceive these.

Interracial contact. In reviewing the vast interracial contact literature, Bowman and Park (2014) argue that "cross-racial interaction comes with the opportunity for both substantive engagement and conflict" (p. 662). First, the interracial contact literature suggests that ethnoracially minoritized and white students interact across groups at markedly different rates within predominantly white institutions. For instance, daily diary studies where undergraduates recorded their interracial contact suggest that most white students have minimal or no daily interracial contact on their predominantly white campus, whereas nearly half of black undergraduates reported daily contact on campus with those from a racial out-group (Nezlek, 2007).

Second, the interracial contact literature also suggests differential outcomes across ethnoracial groups from interracial/interethnic contact. Several scholars have reported that at its outset, interracial contact often engenders anxiety on the part of both white (i.e., due to actual negative racial attitudes or fear of being perceived as racially prejudiced; Dunton & Fazio, 1997; Plant & Devine, 1998; Stephan & Stephan, 2001) and ethnoracially minoritized individuals (i.e., due to anxiety about being targets of racial prejudice; Clark, Anderson, Clark, & Williams, 1999; Major & O'Brien, 2005; Mendoza-Denton, Downey, Purdie, Davis, & Pietrzak, 2002; Tropp, 2003). Other scholars have reported the differential experiences of white and minoritized individuals during interracial contact, namely when such contact involves race-related content (e.g., Trawalter & Richeson, 2008). Illustrative of this is a particular finding from Harper and Quaye's (2007) examination of African American male students' experiences on campus:

> The [African American] students usually chose the phrase "interact with" when they spoke more generally about their exchanges with peers from different cultural and racial/ethnic minority backgrounds, but used "deal with" when referring specifically to their interactions with White people. This semantic difference is noteworthy, as it indicates variable levels of comfort and authenticity in cross-racial interactions and relationships. (p. 138)

Considered together, these particular findings suggest pronounced differences in the numbers and quality of students' interracial exchanges. However, the following discussion illuminates reasons why students could understand such exchanges markedly differently.

Desire for intercultural contact. The cognitive and psychological effort required by many ethnoracially minoritized students to interact across difference relate to variability in students' perceived desirability for intercultural interactions (Harper, 2012; Sorensen et al., 2009;

Tanaka, 2002). Within predominantly white institutions, ethnoracially minoritized students often internalize perceptions of both their campus racial climate (i.e., as more racist, alienating, discriminatory, insensitive and as less accepting; D'Augelli & Hershberger, 1993; Nora & Cabrera, 1996; Rankin & Reason, 2005) *and* their interactions nested within this context (i.e., that they are intellectually inferior, do not belong, and are culturally misunderstood; Hurtado, Carter, & Spuler, 1996; Hurtado & Carter, 1997). Such perceptions increase minoritized students' experiences of racial and ethnic tension on campus (Hurtado, 1994). Further, during intergroup contact, marginalized groups often approach these interactions with the explicit goal of discussing power imbalances and changing power structures, while advantaged groups often wish to focus on similarities between groups, effectively minimizing important identity-related differences (Saguy et al., 2008). Diversity literature suggests that white undergraduates often reap more benefits from interaction across ethnoracial difference compared to their minoritized peers. This well documented finding is often explained in terms of the novelty of such interactions for many white undergraduates (i.e., they often stand to gain more because of their pre-college lack of exposure to diverse groups) instead of in terms of the costs to minoritized students who often must interact across difference with peers substantially less versed in this and who must do so within racist campus climates (Harper, 2012). Considering such findings, the significant costs of interaction across difference—and their relation to students seeking such exchange—cannot be ignored.

Affirming counter spaces. The aforementioned costs of intercultural interaction for many minoritized students often necessitate affirmative cultural enclaves on campus. This is of particular import here since this GPI scale measures intercultural *interaction*. While some can

view group-specific enclaves as segregated from the larger campus—indeed, the perceived self-segregation of students by race and ethnicity is a key criticism of the aim to maintain the ethnoracial diversification of U.S. college campuses as a priority (Duster, 1991)—many students of color view such spaces as invaluable to their successful navigation of campus (Ancis, Sedlacek, & Mohr, 2000). The role of academic, social, and identity-affirming counter spaces (Carter, 2007; Solórzano & Villalpando, 1998; Solórzano et al., 2000) in coping with and resisting the negative effects of racism in predominantly white settings is of particular import here. Solórzano et al. advanced a model to explain how racial microaggressions influence the campus racial climate for African American undergraduates. Racial microaggressions in both academic and social settings negatively influence the academic and social experiences of students (i.e., leading to students' self-doubt, frustration, isolation, and lack of academic progress), which in turn relates to students forming academic and social counter spaces to cope with these negative effects. The counter spaces discussed by the students in the Solórzano et al. study "serve as sites where deficit notions of people of color can be challenged and where a positive collegiate racial climate can be established and maintained," providing important educational, psychological, and cultural supports (p. 70). Academically, counter spaces provide African Americans a supportive learning context where their experiences and knowledge sources are legitimized (Solórzano & Villalpando, 1998), while socially, counter spaces provide an opportunity for students to process frustration and bond or sympathize with others who can relate to being targets of racial microaggressions (Solórzano et al., 2000). Harper and Hurtado (2007) extended this notion across an ethnoracially diverse sample of undergraduates in discussing the pervasiveness of whiteness relative to all aspects of campus life. Beyond particular ethnic and

multicultural centers on campus, "Asian American, Black, Latino, and Native American students found it difficult to identify other spaces on campus in which they felt shared cultural ownership. White interests were thought to be privileged over others…" (p. 18).

Homophily bias. The concept of homophily bias, or individuals' tendency to interact with others similar to themselves (McPherson, Smith-Lovin, & Cook, 2001), is relevant here. Homophily bias operates because—as discussed earlier—individuals categorize themselves and others in automatic, unconscious ways, and the categories used to organize everyday life often serve to delineate similarity or difference between individuals (Roth, 2004). Interpersonal similarities often drive interpersonal exchange because such similarities—and the shared interests that often underlie these—facilitate easier communication (McPherson et al., 2001). While ascribed interpersonal similarities or differences span various dimensions, "shared cultural schemas provide ready classifications of self and others along dimensions that are immediately apparent, such as gender and race" (Roth, 2004, p. 192). Earlier work has suggested that ethnoracially minoritized individuals demonstrate more racially homophilous networks compared to whites, even when ethnoracially minoritized individuals have fewer same-race peers from which to select relationships (Mollica, Gray, & Treviño, 2003). Aligned with the aforementioned discussion on affirming counter spaces, this could be because of the documented positive effect for minoritized individuals of same-race relationships on their psychosocial wellbeing (Thomas & Gabarro, 1999) and in maintaining their racial identity in predominantly white contexts (Ibarra, 1993).

Interactions between U.S.-based and international students. The extant literature on international students' postsecondary outcomes suggests that—compared to students from the

U.S.—these students report more dissatisfaction with college (Glass et al., 2013), feeling more isolated from meaningful peer interaction (Glass & Braskamp, 2012), and ethnocentric tendencies on the part of faculty and their peers (Lee, 2010). Further, Engberg et al. (2016) found that relationships between curricular and co-curricular intercultural engagement and the GPI's Social Interaction scale were conditional on students' international/U.S. student status, suggesting that these groups of students experience intercultural exchange differently. Students from the U.S. are also less inclined to form relationships with international students on campus (Glass et al., 2013; Roberts & Komives, 2016).

In summary, this discussion illuminates key ways in which students' perceptions of intercultural interaction differ. In their framework to understand the campus climate for racial and ethnic diversity, Milem et al. (2005) describe the behavioral component of a campus as including the type, amount, and quality of interactions across ethnoracial groups. This component of their model is especially salient to the present study for several reasons that the preceding discussion outlined. The amount of and desire for intercultural contact vary across student groups, as does the learning and development resultant from such interaction. Given the costs related to intercultural interaction for some students, the role of affirming counter spaces and homophily bias in buffering such costs are critical to understand. Considered together, these differences complicate ways to understand and measure students' intercultural interactions.

Conceptual Framework

The concept of multicultural validity (Kirkhart, 1995) and the *Standards'* framework for validity (AERA et al., 2014) both focus on culturally-relevant approaches to validation related to epistemological, theoretical, and methodological considerations of measurement. As such, the

framework for the present study is informed by the literature reviewed in this chapter that spans three areas: (1) construct underrepresentation, (2) the cognitive model of the survey response process, and (3) culturally variable understandings of the dimensions that comprise the GPI's global perspective development construct. Extending from the need to engage in culturally-relevant validation work, the first two theoretical areas provide broader explanation around the culturally-bound nature of survey research from the perspectives of both survey developers and survey respondents. These first two theoretical bases explain *that* one should conduct measurement invariance testing when using instrumentation to measure diverse populations. In the case of the GPI, the concept of construct underrepresentation drives the need to test whether the instrument's constructs are theorized inclusively. In terms of how students interact with the survey instrument itself, this chapter reviewed the literature on the culturally variable nature of the survey response process. That literature explained that students could understand and respond to the GPI items in ways that are systematically related to their ethnoracial identity. This drives the need to examine measurement invariance across these groups.

Finally, the third theoretical area explains specifically *why* one should conduct measurement invariance testing across ethnoracially diverse groups of students. The review of the literature on culturally variable dimensions of undergraduates' development provides particular theoretical rationale that explains how the developmental dimensions measured by the GPI could be understood or measured differently across diverse ethnoracial groups. Importantly, this nuanced theoretical rationale supports Kirkhart's (1995) assertion that multicultural validity must involve the consideration of the extent to which what is measured reflects participants' lived experiences. Informed by these bodies of literature, this study seeks to empirically test the

validity of the GPI's global perspective development items across an ethnoracially diverse sample of undergraduates. The next chapter will outline the study's methodology, including the study's research questions, analytical approaches, and limitations.

CHAPTER THREE

METHODS

This study investigates the validity of the GPI's global perspective development constructs and whether these are theorized, understood, and measured differently across ethnoracial groups. To execute this, I examine measurement invariance, or the invariance of the GPI's measurement model, across four ethnoracial groups. This study's methodology is informed by the analytical approaches outlined by Chen, Sousa, and West (2005) and Dimitrov (2010) relative to examining measurement invariance of second-order factor models in particular. Since an earlier study validated a hierarchical factor structure for the GPI's global perspective development items (Davidson & Engberg, under review), attending to the additional aspects involved in examining measurement invariance of hierarchical CFA models is critical. In addition, this study's methodology is informed by Bowen and Masa's (2015) application of invariance testing to ordinal datasets. The GPI includes ordinal measures, which require specific considerations relative to the particular model parameters investigated.

Methods Overview

This chapter outlines the study's research questions and methodological approaches and considerations in answering them. As such, I have structured this chapter as follows. I begin by explaining the study's research questions with particular attention to the sequence in which these are asked and what the evidence related to each question suggests. I then describe the study's research context and participants. I describe the GPI, including its development and subsequent

validation as well as the scale and item properties I will examine for the current study. I then overview the analytical approaches I will take to answer my research questions and include in this discussion key considerations related to data characteristics of the GPI (i.e., ordinal, non-normal, and missing data) as well as the criteria selected for examining model fit. Next, I describe the specific analytical procedures that will be used to examine measurement invariance, the second-order factor mean equivalence, and the evidence related to the GPI's factor structure as well as convergent and discriminant validity. I outline the initial procedures as well as the specific models to be tested. I conclude the chapter with a discussion of the study's limitations.

Research Questions

To examine both measurement invariance of the GPI items across ethnoracial groups and the additional validity evidence to inform any instrument refinement efforts, this study will proceed in several steps. I will test various estimates of model parameters in the particular order described here. Each of these model parameters corresponds to a different research question—and, therefore to different aspects of invariance testing—and requires evidence from preceding models to continue.

I will first test for configural invariance, which involves examining whether there are equal factor structures across the ethnoracial groups (i.e., whether there are an equal number of factors and an identical pattern of indictor-factor loadings across groups); this type of invariance is also known as form invariance or equal form and is required to first establish that a baseline model exists across the groups under study (Dimitrov, 2010). Once configural invariance is established, testing for measurement invariance involves examining invariance across different estimated model parameters; this includes examining the equivalence of first-order factor

loadings and second-order coefficients (i.e., metric invariance) and item thresholds (i.e., scalar invariance; Dimitrov, 2010). Metric invariance indicates the equivalence of the magnitude of factor loadings and second-order coefficients, therefore addressing whether respondents answer the items in the same manner (i.e., whether the strength of relationships between indicators and the factors to which they relate are the same). Evidence of metric invariance suggests the different groups investigated understand the measured constructs equivalently. Since GPI data are ordinal in nature, scalar invariance relates to the equivalence of the items' thresholds across groups (i.e., not to the cross-group equivalence of the items' intercepts, as is the case with testing scalar invariance with continuous data; Sass, 2011). Item thresholds are the boundaries (or cut points) where, on average, respondents vary between two distinct response options; item thresholds divide the distribution of responses into distinct categories and equal the total number of item response categories minus one (i.e., the GPI uses a five-point response scale so includes four item thresholds for each item; Sass, 2011). As will be discussed shortly, the estimation method used in this study analyzes a polychoric correlation matrix. This matrix involves computed correlations between each pair of ordinal variables based on the theoretical assumption that a continuous latent variable that is normally distributed underlies the obtained frequencies of ordinal responses for each variable. Following this, each observed ordinal response value corresponds to a range of values between thresholds (i.e., category cutoff points) on the underlying latent variable (Bowen & Masa, 2015). Threshold (scalar) invariance suggests that students' true levels of the latent constructs correspond to the same probability of selecting the same response choices on items related to those latent constructs and that this probability does not differ as a condition of group membership. Evidence of configural, metric, *and* scalar

invariance are required in order to compare GPI mean scores and regression coefficients across ethnoracial groups. To test for configural, metric, and scalar invariance, the study first seeks to answer the following three research questions in this sequence:

Research Q1: Is *equal form* observed across ethnoracial groups for the GPI's items?

Research Q2: Are *equal first-order factor loadings and second-order coefficients* observed across ethnoracial groups?

Research Q3: Are the GPI's *item thresholds* invariant across ethnoracial groups?

Answering the study's next two research questions first requires an empirical determination of configural and at least partial metric and scalar invariance across ethnoracial groups. If evidence related to the first three research questions suggests at least partial invariance across those model parameters, stricter forms of invariance can be tested. The GPI's hierarchical factor structure requires an examination of the equality of disturbances of the first-order factors (i.e., the residual variance in first-order factors not explained by the second-order factor; Chen et al., 2005). This invariance indicates that the amount of any unmeasured causes relative to the first-order factor (as well as random or measurement error) is equivalent across groups. Invariance related to item error variance indicates that the explained variance for each item is equivalent across groups, suggesting that the latent constructs are measured identically across groups. To examine this, I will answer the following two research questions:

Research Q4: Are *equal disturbances* of the first-order factors observed across ethnoracial groups?

Research Q5: Are *equal item error variances* observed across ethnoracial groups?

Chen et al. (2005) explain that one can also use MGCFAs to estimate any cross-group differences in structural parameters (i.e., factor means and any applicable factor covariances). Examining the equivalence of the GPI's second-order factor mean across groups provides evidence to determine whether one can compare students' level of global perspective development across groups. Because I will test the GPI's *hierarchical* factor structure across groups, it is important to note that the hypothesized single second-order factor accounts for the GPI's six first-order factors (i.e., the second-order factor explains all the variance and covariance associated with the first-order factors). As such, the GPI's six first-order factor means are contingent on the single second-order factor mean and cannot be directly compared in this hierarchical model. Only the single second-order factor mean can be compared across groups, and this can only happen if evidence suggests configural and at least partial metric and scalar invariance.

The particular significance of the sixth research question also addresses the import of comparing group means through an MGCFA framework instead of through an analysis of variance (ANOVA) typically used to compare means across three or more groups. Testing for second-order factor mean equivalency corrects for and estimates measurement error within factors and—of particular importance in drawing valid cross-group comparisons—estimates whether the measurement or structural models are equivalent across groups, whereas ANOVAs do not (Sass, 2011). Further, using MGCFAs to test group means is also more appropriate for ordinal data since ANOVAs assume data are continuous and particular MGCFA estimation methods account for ordinal data. The following research question examines this:

Research Q6: Do significant cross-group differences relative to the GPI's second-order global perspective development factor mean exist?

Finally, given the validity evidence obtained from the study's analyses, this study also seeks to inform any refinement efforts of the GPI. In particular, I will examine validity evidence related to the GPI's hierarchical factor structure and the convergent and discriminant validity of the GPI's scales. Earlier results (Davidson & Engberg, under review) suggested that while the GPI's hierarchical factor structure fit the data well, particular model parameters (i.e., the relationship between the GPI's Cognitive Knowing factor and the higher-order global perspective development factor, GPI item error variances) could benefit from more extensive psychometric study. Additionally, convergent and discriminant validity comprise important aspects of construct validity that allow instrument developers to understand the extent that an instrument has been operationalized in ways that accurately reflect its underlying theories. Specifically, examining convergent and discriminant validity evidence illuminates more/less effective measures of an instrument's underlying constructs. Such evidence is useful in identifying specific item or scale refinement opportunities. The study's final research question examines these aspects:

Research Q7: Does evidence support the hierarchical factor structure of the GPI and convergent and discriminant validity of the GPI's scales?

Research Context and Participants

Data Collection

Data for this study were provided by Iowa State University's Research Institute for Studies in Education (RISE), which administered the GPI General Form to students attending

postsecondary educational institutions across the 2015-16 and 2016-17 academic years. Though participating institutions can elect to receive their GPI data in one of two forms (i.e., through identifiable or anonymous data collection procedures), the dataset provided by RISE for the present study included only de-identified responses to protect individual respondents.

Participants

The present study's sample represents a total of 7,092 undergraduates who completed the GPI General Form between 2015-2017. Table 2 includes the sample's demographic composition at the time of survey completion. The sample includes both domestic (n = 6,465, 91.2%), defined on the instrument as American students attending an American college/university) and international undergraduates (n = 403, 5.7%), defined on the instrument as non-American students attending an American college/university). Note that 224 students (3.2% of the sample) indicated "other" student status. Because these students also identified as undergraduates attending a postsecondary institution, they were retained in the study's sample.

The sample includes students who identified as Asian/Native Hawaiian/other Pacific Islander (n = 686, 9.6%), black/African American (n = 689, 9.7%), Hispanic (n = 472, 6.7%), and white (n = 5,245, 74.0%). In terms of gender, the sample is comprised of 63.6% women (n = 4,507), 35.4% men (n = 2,514), and < 1% of both transgender/gender nonconforming (n = 25) and students who did not report this (n = 46). In terms of the highest level of formal education for either parent, the sample reported this as high school graduate or less (16.4%), some undergraduate education or a baccalaureate degree (44.2%), and some graduate education/ advanced degree (30.8%). The average age of the sample was 21.4 years (SD = 4.1 years), and in terms of academic standing, the sample has relatively advanced academic standing with 14.2%

of the students identifying as first-year students, 11.4% as sophomores, 22.9% as juniors, and 51.5% as seniors. The average self-reported undergraduate grade-point average of the sample was 3.31 (*SD* = .57) on a four-point scale.

Table 2. Descriptive Statistics for Sample (*N* = 7,092)

Student Demographic Variables	***n***	**%**
Asian/Native Hawaiian/Other Pacific Islander	686	9.6%
Black/African American	689	9.7%
Hispanic	472	6.7%
White	5,245	74.0%
	n	**%**
Woman	4,507	63.6%
Man	2,514	35.4%
Transgender/Gender Nonconforming	25	0.4%
No Response	46	0.6%
	n	**%**
Domestic Undergraduate	6,465	91.2%
International Undergraduate	403	5.7%
Other Undergraduate	224	3.2%
	n	**%**
First-year Undergraduate	1,007	14.2%
Sophomore	811	11.4%
Junior	1,622	22.9%
Senior	3,652	51.5%
	Mean	***SD***
Self-reported Age	21.4	4.1
Self-reported Undergraduate GPA[a]	3.31	0.57
Parental Educational Attainment	***n***	**%**
Graduate/professional degree	2,043	28.8%
Bachelor's degree	2,257	31.8%
Some college	876	12.4%
High school graduate	951	13.4%
Less than high school diploma	216	3.0%
Do not know or missing	72	1.0%

Note. [a]Undergraduate self-reported grade point average is reported on a four-point scale.

Table 3 includes information about the 36 institutions where the sample's students were enrolled. The sample nearly evenly attended public (n = 3,473, 49.0%) and private (n = 3,619, 51.0%) postsecondary educational institutions. In terms of academic areas of study, over half of the sample reported majoring in business (31%) and science/technology/engineering/math (29.6%). In terms of Carnegie Classification (i.e., the particular classification of institutional attributes and outputs), Table 3 provides the breakdown of the 36 institutions; most (26 schools) are either doctoral or master's institutions. Three of the 36 participating institutions were established minority-serving institutions (i.e., one Asian American and Native American Pacific Islander-serving Institution (AANAPISI), one Historically Black University, and one Hispanic-serving Institution), while three other participating institutions were AANAPISI-eligible.

The Global Perspective Inventory

The GPI, administered online, measures the development of a global perspective, respondents' engagement with various curricula, co-curricular opportunities, faculty, and current events, as well as their perceived sense of community at their institution (RISE, 2017b). The GPI is primarily used to measure college students' educational experiences as these relate to the development of a global perspective. There are three versions of the GPI (General, New Student, and Study Abroad forms) administered for various assessment, evaluation, and research purposes.

Table 3. Institutional and Enrollment Information for Sample (N = 7,092)

Students' Academic Field of Study	***n***	**%**
Business	2,201	31.0%
Science, Technology, Engineering, or Math	2,096	29.6%
Other Academic Area Not Listed or Missing	768	10.8%
Social or Behavioral Sciences	688	9.7%
Education or Social Work	514	7.2%
Arts and Humanities	492	6.9%
Communication or Journalism	333	4.7%

Institutional Variables	***n***	**%**
Private Institution Enrollments	3,619	51.0%
Public Institution Enrollments	3,473	49.0%

Institutional Carnegie Classification[a]	**# Institutions**
Doctoral Universities: Highest, Higher, and Moderate Research Activity (combined)	13
Master's Colleges and Universities: Larger, Medium, and Small Programs (combined)	13
Baccalaureate Colleges: Arts and Sciences Focus and Diverse Fields (combined)	8
Associate's Colleges: Mixed Transfer/Career and Technical-High Non-traditional	1
Special Focus Four-year: Engineering Schools	1

Total Institutional Undergraduate Enrollment[b]	**# Institutions**
≥ 30,000	3
20,000 – 29,999	3
10,000 – 19,999	4
5,000 – 9,999	7
2,000 – 4,999	13
≤ 1,999	6

Minority-serving Institutions	**# Institutions**
AANAPISI[c] (Asian American and Native American Pacific Islander-serving Institutions)	4
HBCU (Historically Black Colleges and Universities)	1
HSI (Hispanic-serving Institutions)	1

Note. [a]2016-17 Carnegie Classifications; [b]Source: IPEDS 2015-16, 12-month enrollment; [c]AANAPISI totals include one established and three eligible institutions.

The present study uses data from only the GPI General form for two reasons. First, the General form is the most widely used GPI form due to its varied uses. For instance, the GPI General form can be administered for assessment and research that are both cross-sectional (i.e., as a stand-alone assessment of students' global perspective development at any point in their college career, including senior exit assessment) and longitudinal (i.e., as a pre-test for study abroad or other experiential learning participation and as a post-test for students who completed the New Student form upon college entry or for students who completed the General form as a pre-test related to another experiential learning opportunity in which they engaged; RISE, 2017c). Second, the GPI New Student and Study Abroad forms are administered to more specific student samples. The New Student form is administered to incoming first-year college students to measure a variety of pre-college engagement and their global perspective development upon entry to college. When examining different samples of first-year college students who completed either the GPI New Student or General form, the theorized six-factor structure of global perspective development could not be validated for the GPI New Student form sample (RISE, 2017d). The Study Abroad form is administered to study abroad participants upon return from such an experience to measure components of their engagement with the host country's culture and residents and their global perspective development after studying abroad. Earlier work (e.g., Braskamp, Braskamp, & Merrill, 2009; Engberg, Jourian, & Davidson, 2016) suggests that students' GPI scale scores increase after study abroad compared to their pre-study abroad scores. Excluding both the New Student and Study Abroad form samples from this study therefore makes sense in an attempt to attenuate the effects of particular developmental differences observed within these respective student groups. The three GPI forms

include the same 35 core items, while the remaining items on each form measure particular aspects relevant to their purposes as described above. Though there are 35 core items on each of the GPI's forms, three of these items are not conceptually related to the global perspective development constructs and are intended as stand-alone items for institutional use only. As such, the current study examines measurement invariance for only the GPI's 32 global perspective development items.

The GPI's 32 global perspective development items are Likert-type items measuring respondents' level of agreement on a five-point scale (1=Strongly Disagree to 5=Strongly Agree). Previous exploratory factor analyses using these 32 items suggested a six-factor global perspective structure yielding six conceptually distinct yet related scales (Braskamp et al., 2014). More recent confirmatory factor analyses suggested a hierarchical factor structure with a single second-order global perspective development factor accounting for the same six first-order factors as previously observed (Davidson & Engberg, under review). The GPI's six scales and their composite reliabilities (CR) are as follows. I report CR values instead of Cronbach alpha coefficients given the SEM/CFA framework for this study. The SEM approach to reporting scale reliabilities involves reporting CR values, as—unlike Cronbach alpha coefficients— these consider *varying* factor loadings for each item (i.e., each item's standardized regression weights are used to calculate these) and item error variances (Raykov, 1997). The GPI's scales include the Cognitive Knowing (CR = .59) and Cognitive Knowledge (CR = .81) scales, which measure students' epistemological development as well as intercultural knowledge and awareness, respectively. The Intrapersonal Identity (CR = .80) and Affect (CR = .80) scales measure students' personal values as well as comfort and sensitivity toward difference, respectively. The

Interpersonal Social Interaction (CR = .77) and Social Responsibility (CR = .78) scales measure students' preferences for intercultural relationships and their commitment to giving back and making a difference in society, respectively. Table 1 includes the GPI scales, their CR values, and component items.

Development and Validation of the GPI

Instrument development. In 2007, the Global Perspective Institute developed an initial GPI item pool including several hundred items that were subjected to college student as well as study abroad and student development expert review for clarity and credibility (i.e., to determine face validity; Braskamp et al., 2014). From this, 69 items that measured global perspective development were retained for the initial version of the GPI. After initially administering the GPI in 2007, findings from statistical analyses (i.e., exploratory factor analyses, reliability and internal consistency analyses) and qualitative approaches (i.e., soliciting respondents' feedback about item wording and meaning) informed efforts to eventually—through several versions of the instrument—reduce the number of items to the current 32 global perspective development items (Braskamp et al., 2014). The Global Perspective Institute hosted the GPI until 2015, when Iowa State University's RISE assumed this responsibility. Since its creation, nearly 200 colleges, universities, and other educational organizations—including over 120,000 individuals in the U.S. and abroad—have used the GPI to understand aspects of students' global perspective development (RISE, 2017a).

Validation of the GPI. Extending from earlier exploratory factor analytic work on the GPI (Braskamp et al., 2104), confirmatory factor analyses using the GPI's global perspective development items and two separate samples of undergraduates (N = 11,051) more recently

suggested a single second-order factor and six first-order factor structure of global perspective development (Davidson & Engberg, under review). Such an examination of the GPI provided an empirical validation of both the relatedness of the GPI's six first-order dimensions and that these coalesce into a single hierarchical global perspective development construct (Davidson & Engberg, under review). An important nuance to this finding was the validation of a *single* hierarchical global perspective development construct as opposed to a three-factor hierarchical model that included the cognitive, intrapersonal, and interpersonal dimensions that theoretically represent the development of a global perspective. This particular finding was largely explained in terms of the two GPI scales that measure students' cognitive development; Davidson and Engberg (under review) observed a statistically non-significant, weak relationship ($r = .14$) between the two first-order cognitive scales. Yet the relationships between the two first-order scales related to both the intrapersonal and interpersonal domains of development were both statistically significant and strong ($r = .65$ and .72, respectively). When imposing models that forced (1) the two first-order cognitive constructs to interrelate *only* with one another or (2) a second-order cognitive development construct to explain the two first-order cognitive constructs, model fit, parameter estimates, and the amount of variance explained all worsened (Davidson & Engberg, under review).

Related to other validation studies for the GPI, the only other study to date was a longitudinal study by Anderson and Lawton (2011) that examined the convergent validity of the GPI. They compared the intercultural development of study-abroad participants with undergraduates who remained on campus using both the GPI and the IDI to measure aspects of students' intercultural development before and after engaging in either study abroad or on-

campus curricular opportunities. The authors' motive in using two different instruments in their study was two-fold: (1) to investigate any relationship between the two instruments and (2) to conclude whether using two different measures of intercultural development "would provide a broader assessment of the impact of a study abroad program on its participants" (p. 90). Results from this study suggested weak relationships between the two instruments' scales (i.e., in terms of the strength and statistical significance of the relationships). As such, convergent validity between the GPI and IDI scales was not established in their study, suggesting that the GPI and IDI measure different dimensions of intercultural development.

Data Analyses

Overview of Analytic Approach

First, I used the Statistical Package for the Social Sciences, Version 24.0 to generate overall sample descriptive statistics, create subsamples by ethnoracial group, perform internal consistency reliability testing on the GPI scales, and obtain descriptive and distributional statistics for the GPI's 32 items. Second, I used LISREL 8.8 to examine the factor structure and measurement invariance of the GPI using CFAs and MGCFAs within a SEM framework.

Data Characteristics and Estimation Method

The characteristics of the GPI's 32 items relate to considerations around the ordinal nature and multivariate normality of these variables as well as missing data as these relate to the model estimation approach employed. Table 4 includes the descriptive and distributional statistics for the GPI's items.

Table 4. Descriptive and Distributional Statistics for the 32 GPI General Form Global Perspective Development Items Using the Aggregate Sample (N = 7,092)

GPI Item	Missing Cases[a]	Mean[b]	*SD*	Skewness	*SE* Skewness	Kurtosis	*SE* Kurtosis
COGNITIVE KNOWING							
COGEP01	16	3.00	.95	-.07	.03	-.07	.06
COGEP06	21	3.84	1.13	-.83	.03	-.15	.06
COGEP07	37	3.01	1.14	-.05	.03	-.93	.06
COGEP16	26	4.13	.70	-.69	.03	1.12	.06
COGEP19	35	3.94	.74	-.56	.03	.54	.06
COGEP20	28	3.47	1.01	-.38	.03	-.41	.06
COGEP30	23	3.57	1.04	-.58	.03	-.29	.06
COGNITIVE KNOWLEDGE							
COGKNW08	16	3.63	.95	-.57	.03	-.17	.06
COGKNW13	28	3.68	.84	-.60	.03	.28	.06
COGKNW17	28	3.76	.80	-.56	.03	.30	.06
COGKNW21	50	3.72	.78	-.50	.03	.28	.06
COGKNW27	42	3.87	.79	-.58	.03	.41	.06
INTRAPERSONAL IDENTITY							
IDENT02	6	4.13	.89	-1.00	.03	.78	.06
IDENT03	10	4.26	.66	-.75	.03	1.37	.06
IDENT09	23	4.17	.77	-.92	.03	1.16	.06
IDENT12	23	3.97	.74	-.55	.03	.61	.06
IDENT18	26	4.00	.68	-.41	.03	.49	.06
IDENT28	36	4.02	.77	-.54	.03	.31	.06
INTRAPERSONAL AFFECT							
AFFECT22	24	4.05	.79	-.74	.03	.77	.06
AFFECT23	30	4.03	.78	-.86	.03	1.23	.06
AFFECT25	25	4.36	.68	-.95	.03	1.27	.06
AFFECT31	21	4.28	.71	-.84	.03	1.03	.06
AFFECT33	31	4.06	.71	-.56	.03	.63	.06
INTERPERSONAL SOCIAL RESPONSIBILITY							
SOCRES05	18	3.73	.88	-.43	.03	-.08	.06
SOCRES14	34	3.67	.81	-.22	.03	-.05	.06
SOCRES26	26	3.76	.86	-.45	.03	.11	.06
SOCRES32	28	3.82	.78	-.38	.03	.17	.06
SOCRES34	21	3.72	1.04	-.65	.03	-.13	.06
INTERPERSONAL SOCIAL INTERACTION							
SOCINT04	6	2.52	1.10	.55	.03	-.55	.06
SOCINT24	23	3.90	.96	-.70	.03	-.11	.06
SOCINT29	20	3.52	.93	-.12	.03	-.54	.06
SOCINT35	23	3.32	1.09	-.08	.03	-.88	.06

Note. [a]The number of cases for whom responses were missing for these GPI items; missing values were imputed in PRELIS, but 75 cases (1% of the original sample) were deleted since missing values could not be imputed. As such, all CFAs and MGCFAs were conducted with an overall sample size of 7,017 cases. [b]All 32 GPI items on these scales measure respondents' level of agreement on the same five-point scale (1=Strongly Disagree to 5=Strongly Agree).

Scale coarseness. Given that the observed GPI measures are ordinal, LISREL's default maximum likelihood estimation method is not ideally suited for the present study's CFAs and MGCFAs. First, the issue of scale coarseness is relevant. A measurement scale is considered coarse when a construct that is continuous in nature is instead measured using limited response categories that can collapse different true scores into the same category (Aguinis, Pierce, & Culpepper, 2009). Whenever continuous constructs are measured using Likert-type or ordinal items, measurement scale coarseness exists. This coarseness relates to measurement imprecision, as this does not allow respondents to answer items in ways that are adequately discriminating (Aguinis et al., 2009). Koh and Zumbo (2008) explain that maximum likelihood estimation methods are not particularly suited for ordinal measures given their scale coarseness; the Pearson covariance is distorted with ordinal variables. Treating categorical variables as continuous in CFAs (i.e., using normal theory estimators such as maximum likelihood or robust maximum likelihood methods) can result in incorrect parameter estimates, lower estimates of the relationships among indicators, "pseudofactors" that result from item difficulty or extremeness, and incorrect test statistics and standard errors (Brown, 2015, p. 353; Koh & Zumbo, 2008).

Multivariate non-normality. Koh and Zumbo (2008) explain that because of the small number of discrete response categories of ordinal variables, these variables are frequently subject to multivariate non-normality. As Table 4 illustrates, most items are at least moderately skewed. To account for both the ordinal and non-normal nature of the GPI data used for this study, I employed the diagonally weighted least squares (DWLS) estimation method (Muthén, 1993) within LISREL for all CFAs and MGCFAs, where polychoric correlations among the ordinal variables are weighted by the diagonal item variances of the asymptotic covariance matrix. Both

the polychoric correlation and asymptotic covariance matrices were created and inputted via LISREL. Using DWLS estimation yields more accurate results when analyzing ordinal data by correcting the standard errors and chi-square values resultant from non-continuous, multivariate non-normal variables (Finney & DiStefano, 2006). Additionally, factor loadings are frequently underestimated with robust maximum likelihood estimation but were found to be more precise and accurate with DWLS estimation, regardless of the number of response categories (Li, 2016). This practice also reinforces the findings of Bernstein and Teng (1989), who found that factor analyses of Likert-type items (i.e., when assuming continuous scales) failed to accurately uncover the data's structure. Pike (2013) cited this very practice as a key statistical limitation in the extant research using factor analyses to validate instrument structure.

Missing data. Typically applied strategies for handing missing data (i.e., listwise or pairwise deletion) can be problematic within an SEM framework due to the loss of statistical power and potential for biased parameter estimates, standard errors, and test statistics (Brown, 2015). Relative to the present study's dataset, Table 4 indicates the number of missing cases for each of the 32 items under study. While listwise or pairwise deletion typically produces consistent (unbiased) parameter estimates when data are missing completely at random (MCAR), these techniques can result in a substantial loss of sample size and relate to biased standard errors compared to multiple imputation (Brown, 2015).

In addition to retaining adequate sample sizes and avoiding biased estimates, multiple imputation is the ideal strategy to use to estimate missing data given the DWLS estimation method employed in this study (i.e., maximum likelihood methods for estimating missing data are not suited in these cases; Allison, 2003; Brown, 2015). I used PRELIS—an application

within LISREL used to transform data and impute missing data—to handle missing data through multiple imputation. From the total sample of 7,092 cases, PRELIS was able to impute missing data for 7,017 cases; only 1% of cases could not be imputed and were deleted so that no missing data were included in the CFAs or MGCFAs.

Criteria for Evaluating Model Fit

To answer this study's research questions—which require examinations of both the GPI's overall baseline factor structure as well as measurement invariance across ethnoracial groups—particular determinants of model fit are used to understand how well particular models imposed on the GPI fit the data. With reference to testing the initial overall baseline factor structure, the fit of this completely unconstrained (i.e., all parameters are freely estimated) model is determined using particular goodness-of-fit indices. Hu and Bentler (1999) recommend using measures of absolute fit (i.e., root mean square error of approximation [RMSEA], standardized root mean square residual [SRMR]) and relative fit (i.e., comparative fit index [CFI], and nonnormed fit index [NNFI]), using RMSEA $\leq .06$, SRMR $\leq .08$, and CFI and NNFI $\geq .95$ as determinants of good model fit. It is important to consider fit indices from multiple fit categories (i.e., absolute fit such as SRMR, parsimony correction such as RMSEA, and comparative fit indices such as CFI and NNFI) as well as parameter estimates (Brown, 2015). This study bears this in mind and reports all of those in Chapter 4 and in the corresponding tables.

Evaluating measurement invariance involves determining changes in model fit with increasingly more constraints placed upon a model. The central concern is whether placing such constraints worsens model fit. To test hypotheses in SEM by comparing the fit of a less constrained, baseline model (i.e., more estimated parameters) to a more constrained, nested

comparison model (i.e., fewer estimated parameters typically obtained by fixing parameters, constraining parameters as equal, or indicating invariant parameters), the chi-square difference test is widely used to determine whether a comparison model's fit is significantly worsened by imposing particular constraints (Bryant & Satorra, 2012). Dimitrov (2010) explains that invariance of particular parameter estimates is typically determined if the maximum likelihood chi-square ($ML\chi^2$) difference test for nested models is not statistically significant (i.e., usually using an alpha level of .05; this suggests that there is no significant change in model fit with the more constrained model). However, the $ML\chi^2$ statistic is sensitive to both sample size and multivariate normality, both of which are salient to the present study's dataset. The DWLS estimation method in LISREL 8.8 used in these analyses instead provides the Satorra-Bentler scaled chi-square test statistic ($SB\chi^2$; Satorra & Bentler, 2001), which adjusts the $ML\chi^2$ statistic that is biased from multivariate non-normality through a scaling correction factor that considers the multivariate kurtosis that biases this test statistic; the $SB\chi^2$ is calculated by dividing a model's $ML\chi^2$ value by this scaling correction factor. However, the difference in $SB\chi^2$ values to compare nested models does not follow a chi-square distribution (Satorra, 2000), so one cannot simply use the difference between two SB χ^2 values with their associated difference in degrees of freedom to determine whether the difference is statistically significant. As such, the scaled chi-square difference test must be used. In order to generate a scaled WLS chi-square difference test from LISREL results, I will use the macro provided by Bryant and Satorra (2013), as the scaling correction factor used differs based on the SEM software employed for analyses (Bryant & Satorra, 2012).

Large subsample sizes can incorrectly inflate MLχ^2 and SBχ^2 values, and these statistics are also sensitive to model complexity (i.e., the number of factors or observed indicators per factor; Brown 2015; Bryant & Satorra, 2012; Dimitrov, 2010). Specific to investigating measurement invariance, several scholars recommend examining the chi-square statistic in large subsamples, but supplementing this with an examination of changes in alternative fit indices across nested models (e.g., Cheung & Rensvold, 2002; Meade, Johnson, & Braddy, 2008; Vandenberg & Lance, 2000). Cheung and Rensvold (2002) conducted a simulation study that examined the changes in 20 alternative fit indices across nested models; they determined the change in CFI (ΔCFI) was a particularly robust indicator of measurement invariance. Cheung and Rensvold (2002) recommend using a ΔCFI < -.01 (a decrease in CFI larger than .01) as an indicator that the parameters under examination are not invariant. Meade et al. (2008) reiterated Cheung and Rensvold's (2002) recommendations in their rigorous simulation study of the use of various alternative fit indices in determining measurement invariance; they also recommended reporting the ΔCFI. I report the scaled chi-square difference test results and ΔCFI in my results. For the sake of model comparison, I will rely on the ΔCFI in determining measurement invariance given the subsample sizes and their impact on the SBχ^2 statistic.

Sample Size Considerations

In studies employing SEM, the number of cases influences model convergence, precision, statistical power of the models' parameter estimates, and the reliability of fit indices (Brown, 2015). However, the recommendations for determining an appropriate sample size for SEM analyses vary and are often not model-specific (Wolf, Harrington, Clark, & Miller, 2013). Various scholars explain the often-employed approaches used to determine minimum samples

sizes in SEM: employing a standard minimum sample size (e.g., a minimum of 200 cases), including a particular number of cases per observed measure (e.g., $N/p \geq 10$; the ratio of N to the number of observed variables in a model), including a particular number of cases per estimated parameter (e.g., $N/q \geq 5$; the ratio of N to the number of estimated parameters in a model), and conducting power analyses (Myers, Ahn, & Jin, 2011; Wolf et al., 2013). Muthén and Muthén (2002) also explain that the question of sample size depends on a number of things, including model size, multivariate normality of the data, amount of missing data, reliability of the variables, and the number and strength of the indicator-factor relationships.

Specific to sample sizes using the DWLS estimation method employed in the present study, Jöreskog and Sörbom (1996) recommend that estimating asymptotic covariance matrices requires a minimum sample size of $(p+1)(p+2)/2$, where p represents the number of observed variables. While the initial CFA that examines the baseline model will involve the overall sample (i.e., pooling all ethnoracial groups included in the study), the subsequent models will involve MGCFAs, using four separate subsamples representing the individual ethnoracial groups. In the present study, there are 32 observed GPI measures under study, requiring a minimum of 561 cases per subsample using Jöreskog and Sörbom's recommendation. Three of this study's groups (ANHPI, black, and white groups) meet this minimum number of cases, while the Hispanic group size falls below this ($n = 467$).

MGCFA Models and Unbalanced Group Sizes

Yoon and Lai's (2018) findings suggest that when using MGCFAs to examine measurement invariance, large differences in group sizes can lead to incorrectly concluding invariant model parameters. In the event of unbalanced group sizes, Yoon and Lai recommend

examining measurement invariance by subsampling any relatively larger groups—effectively making the groups used for MGCFAs comparable in size—to ensure more accurate conclusions. The present study's white sample consisted of 5,200 cases, while the three other samples consisted of 675 (ANHPI), 673 (Black), and 467 (Hispanic) cases; this created extremely large sample size ratios between 8 and 11. As such, I modified Yoon and Lai's subsampling approach and created a separate subsample of white cases by randomly sampling approximately 15% of the original white sample (n = 766 for the random subsample). While Yoon and Lai recommend using *many* bootstrap samples in their subsampling approach, a technical anomaly within LISREL 8.8 unfortunately prevents the execution of the repeat command (i.e., using a single syntax file across many datasets at once, such as the bootstrap samples produced in PRELIS) with imputed data or while using a large number of resamples. As a workaround, I reran all baseline models and all MGCFAs using the random subsample of white students for all analyses and report these fit statistics along with those obtained for the overall sample.

Analytic Procedures for Examining Measurement Invariance

Forming Subgroups for Invariance Testing

Responses to the GPI's demographic item *Please indicate your racial/ethnic background (mark all that apply)* were used to form the present study's ethnoracial subgroups for invariance testing. For students who responded to this item, they could select *all that apply* from the following six categories: American Indian or Alaska Native, Asian, black or African American, Hispanic (of any race), Native Hawaiian or other Pacific Islander, and white. Four types of details about excluded cases are important to the present study. First, cases were removed from this study's dataset if they completed the survey more than once (i.e., only their initial response

was retained). Second, cases with missing data for this demographic item were excluded from the study's dataset. Third, students who identified as multiracial (i.e., by indicating belonging to two or more categories; $n = 469$) were excluded from the separate ethnoracial groups to which they indicated belonging. In addition, these students were not included as a stand-alone multiracial group in this study in alignment with not treating such an ethnoracially diverse group monolithically (Rockquemore et al., 2009). Fourth, students who identified as American Indian or Alaska Native ($n = 70$) were excluded from the dataset given these extremely small subsample sizes. In addition to these excluded cases, I combined the Asian and Native Hawaiian or other Pacific Islander subsamples into a single group given the small latter subsample.

Scaling Latent Variables

In order to be estimated, the metric of each factor must be defined since latent constructs are unobserved and therefore have no defined unit of measurement (Brown, 2015). There are two typical approaches used to identify a factor's scale: (1) selecting a marker variable within each factor and fixing its value to one, (2) fixing the variance of each factor to one, which standardizes all the factor loadings (Brown, 2015). Chen et al. (2005) explain that the first approach is more appropriate when examining measurement invariance of hierarchical CFA solutions across groups, though they also caution researchers to carefully select the particular marker variable for each first- and second-order factor. Though this concern is widely held among researchers examining measurement invariance, the procedures for selecting a marker variable vary in terms of methodological complexity and their intentionality of use (Bowen & Masa, 2015; Chen et al., 2005; Cheung & Rensvold, 1999).

To define each factor's metric, I examined the factor loadings in each of the four group-specific CFA baseline models and considered three aspects: (1) the loading for each item with the smallest difference in magnitude across groups could suggest the most invariant factor loading (Cheung & Rensvold, 1999), (2) the statistical significance of the factor loadings was critical; invariant loadings needed to be statistically significant across all groups (Cheung & Rensvold, 1999), and (3) whether there were substantive, theoretical explanations to guide the selection of the most invariant item (Dimitrov, 2010). These fixed factor loadings remained the same for all MGCFAs. Note that LISREL does not require a second-order coefficient to be fixed to 1.0 to define the metric of the second-order construct; rather the variance of that single second-order factor is fixed to 1.0 so that each second-order coefficient can be freely estimated.

Testing Invariance Across Groups

The present study employs the forward, or sequential constraint imposition approach to invariance testing, which begins with the least constrained solution (i.e., complete lack of invariance) and gradually imposes more restrictions on particular model parameter estimates to determine whether the model fit worsens by such impositions (Dimitrov, 2010). As described earlier, determinants of decrement in model fit will be determined by ΔCFI values. The approach taken in this study is preferable to the backward, step-down, or sequential constraint release strategy for invariance testing for several reasons outlined by Brown (2015). First, within complex MGCFA models (e.g., several factors, more than two groups), it can be difficult to discern the multiple sources of non-invariance in a poor-fitting model held to full invariance across all parameters given that there are so many. Second, some of the forms of invariance tested in this study require evidence for already-existing invariance across other parameters. A

fully constrained model obscures the results of less restricted models; Brown (2015) argues that it is more logical to proceed gradually forward from a less restricted model to understand any model decrement in smaller increments. In alignment with the approach explained by Chen et al. (2005), each of the steps in the forward approach is described sequentially here in terms of the models tested.

Separate model estimation in each group (preliminary models). First, the hypothesized hierarchical CFA model will be estimated within each of the four ethnoracial groups separately and evaluated using the previously described determinants of model fit (i.e., RMSEA, SRMR, CFI, and NNFI cutoff values). Each of the disaggregated, group-specific preliminary CFA models (and the subsequent baseline MGCFA model used to test configural invariance explained next) require evidence suggesting adequate model fit in order to proceed with subsequent invariance testing. If adequate model fit is demonstrated across all groups, the examination of configural invariance will commence.

Multiple-group configural invariance (Model 1). In examining configural invariance, a baseline model with the same pattern of free and fixed factor loadings for the first- and second-order factors will be imposed across all four groups. While this pattern of factor loadings will be constrained as equal across all groups, all other parameters will be freely estimated across the separate groups. Model fit will be evaluated in this baseline model using the same determinants of model fit as in the preliminary models (i.e., RMSEA, SRMR, CFI, and NNFI). If configural invariance is established, I will then begin placing constraints on particular model parameters as detailed below.

Invariance of first-order factor loadings (Model 2). In examining this level of invariance, I will first constrain *all* first-order factor loadings as equal across groups to determine whether relationships between the first-order latent constructs and their indicators are equivalent across groups. All other parameter estimates will be freely estimated. This model is nested within Model 1, so ΔCFI will be examined to determine any model decrement from these constraints. If evidence does not support full invariance, I will examine partial invariance, using Steenkamp and Baumgartner's (1998) process to differentiate invariant and non-invariant factor loadings and again examine ΔCFI in determining model fit.

Invariance of second-order factor coefficients (Model 3). In examining this level of invariance, I will constrain the invariant first-order factors (from Model 2) *and* second-order coefficients as equal across groups to determine whether relationships between the first-order factors and the second-order factor are equivalent across groups. This model is nested within Model 2, so ΔCFI will be examined to determine any model decrement from these constraints. If evidence does not support full invariance, I will examine partial invariance, using Steenkamp and Baumgartner's (1998) process to differentiate invariant and non-invariant coefficients and again examine so ΔCFI in determining model fit.

Invariance of item response category thresholds (Model 4). In models that use ordinal data, invariance related to the thresholds—or item response category distribution cut points—must also be examined. To examine this, I will constrain all invariant first-order factor loadings, second-order coefficients, and the item thresholds as equal across groups. This model is nested within Model 3 so ΔCFI will be examined to determine model decrement from these constraints. If evidence does not support full invariance, I will examine partial invariance, using Steenkamp

and Baumgartner's (1998) process to differentiate invariant and non-invariant thresholds and again examine ΔCFI in determining model fit.

Invariance of first-order factor disturbances (Model 5). Next, I will examine whether the first-order factor disturbances are equal across groups. In hierarchical factor structures, the second-order factor accounts for the covariance among the first-order factors (Brown, 2015). The disturbances of these first-order factors represent variance unexplained by the second-order factor. To examine this aspect of invariance, I will constrain all invariant first-order factors, second-order coefficients, the invariant item thresholds, and the disturbances of the first-order factors as equal across groups. This model is nested within Model 4 so ΔCFI will be examined to determine model decrement from these constraints. If evidence does not support full invariance, I will examine partial invariance, using Steenkamp and Baumgartner's (1998) process to differentiate invariant and non-invariant disturbances and again examine ΔCFI in determining model fit.

Invariance of measured variables' error variances (Model 6). I will next examine whether the item error variances (item uniqueness) are equal across groups. To examine this, I will constrain all invariant first-order factor loadings, second-order coefficients, the invariant thresholds, the invariant disturbances of the first-order factors, and the error variances of the 32 GPI items as equal across groups. This model is nested within Model 5 so ΔCFI will be examined to determine model decrement from these constraints.

Cross-group invariance of second-order factor mean (Model 7). If there is evidence of at least partial scalar invariance during measurement invariance testing, Chen et al. (2005) explain that one can also use MGCFAs to estimate any cross-group differences in structural

parameters (i.e., factor means and any applicable factor covariances). This can extend to examining group differences relative to the GPI's second-order factor mean, addressing whether the groups under study have different average levels of global perspective development. To examine this, I will constrain the first-order factor loadings, second-order coefficients, thresholds, disturbances of the first-order factors, error variances of the GPI items, and the single second-order factor mean as invariant across groups. This model is nested within Model 5 so ΔCFI will be examined to determine model decrement from these constraints. If this model does not demonstrate a worsening in fit, the factor mean is invariant across the groups tested. However, if the model fit worsens from this imposed constraint, further testing will commence to compare the factor means across groups. To do this, I will fix one group's second-order factor mean to zero (an arbitrary reference group), while the remaining three groups' second-order factor means are freely estimated (comparison groups). I will change the reference group until all groups are tested. Because in this model the invariant item thresholds are constrained as equal across groups, the factor mean reflects *differences* in the mean level of the latent construct between the reference and comparison groups, not an absolute mean value. As such, these results would indicate that, on average, one group scores a particular number of units higher/lower than another; the units are based on the metric of the marker variable.

Examining Additional Validity Evidence

The GPI's Hierarchical Factor Structure

Once I complete measurement invariance testing to answer the study's first six research questions, I will review the statistical output for evidence that can answer the study's final research question. This last research question asks whether there is evidence to support the

GPI's hierarchical factor structure and the convergent and discriminant validity of the GPI's scales. Determining whether a hierarchical factor structure fits the data well requires an evaluation of the (1) determinants of model fit (i.e., RMSEA, SRMR, CFI, and NNFI) and (2) strength and statistical significance of the model parameters. The first evaluation will have already occurred while establishing a final baseline model to use for the study's measurement invariance testing. The second evaluation, however, is critical in evaluating the GPI's hierarchical factor structure. Particular to hierarchical CFA, one must examine relationships between the second-order and first-order latent constructs. To evaluate these relationships, I will use the final baseline model used to conduct the study's measurement invariance testing, closely examining and reporting the second-order coefficients' values and statistical significance.

In addition to examining model parameter estimates, I will report the reliability of the GPI's first- and second-order factors. Reliability relates to how well the first-order factors (and items on their respective scales) all measure the same thing. In addition to the composite reliabilities already reported for the GPI's first-order factors (see Table 1), another determinant of reliability in SEM is the average variance extracted (AVE), or the average percentage of variation in the first-order factors accounted for by the higher-order factor (and—as will be discussed in the next chapter— the average percentage of variation in the items accounted for by their latent constructs). The AVE is important because it explains the degree to which a scale's items converge around the same latent construct. In SEM, the AVE explains the degree to which (1) the first-order factors and (2) a scale's items converge around the same latent construct (Hair, Hult, Ringle, & Sarstedt, 2014). The AVE for the GPI's higher-order factor can be calculated using the final baseline model's parameter estimates. I will take the sum of the square of the

second-order coefficients and divide that by the number of first-order factors; an AVE ≥ 50% demonstrates a reliable second-order factor (Hair et al., 2014).

Convergent and Discriminant Validity of the GPI's Scales

Convergent validity aids in understanding the dimensionality of scales since this represents the extent to which items on their respective scales converge, or measure the same underlying construct; in other words, evidence of convergent validity suggests that a scale's items effectively measure the underlying construct (Hair et al., 2014). In a CFA framework, convergent validity is evaluated using a scale's factor loadings (should all be statistically significant and > .50), composite reliability (CR should be > .70), and the AVE (should be > .50; Hair et al., 2014). I will use the final baseline model used for measurement invariance testing to evaluate these aspects.

Discriminant validity of an instrument's scale explains the extent to which that latent construct measures something unique that the other latent constructs in the model do not measure (Henseler, Ringle, & Sarstedt, 2015). In terms of the GPI, it is theoretically expected that there is some degree of relation between the six developmental dimensions, but these dimensions should also measure distinct aspects of students' development. Evidence for this type of validity related to first-order factors requires that the factors not correlate excessively with one another. In SEM, discriminant validity is important because if there is no evidence supporting this, the latent constructs exert an influence on the variation of the observed indicators beyond the ones to which they are theoretically related (Henseler et al., 2015). To determine discriminant validity, I will use the final baseline model used for measurement invariance testing. I will compare the square root of a construct's AVE to the inter-factor correlations among the first-order factors; if

the former is greater in value than the latter, this provides evidence of discriminant validity. I will report all of these additional validity findings in Chapter Four and discuss them, including their associated implications, in more detail in Chapter Five.

Study Limitations

There are several factors that may limit the generalizability of the present study's findings. First, one of this study's purposes is to provide evidence for practitioners and researchers that informs their ability to compare students' global perspective development across ethnoracial groups. In proceeding with this particular purpose, I acknowledge the complexities of examining ethnoracial differences related to a variety of educational experiences and outcomes and that most educational surveys use conventional racial and ethnic categories (i.e., Asian, black, Hispanic/Latino/a, and white) that do not account for panethnic diversity within these categories, the racialization of immigrant status, or indigenous students' experiences (Irizarry, 2015). Additionally, as previously discussed, there is much complexity and variability in individuals' survey response processes about their race and ethnicity (i.e., how respondents understand their own racial and ethnic identities, how their racialized selves are understood by others, and the available racial and ethnic categories from which to select; Rockquemore et al., 2009). As with other large-scale educational surveys, understanding cross-group differences related to students' outcomes and their engagement has historically been of interest to those using the GPI to understand nuances related to undergraduates' development. As such, I proceed with this purpose while acknowledging the inherent limitations of collapsing potential intra-group diversity and relying on students' self-reported responses considering this.

Second, like many other large-scale postsecondary educational surveys, the GPI is comprised of students' self-reported data. Self-reported data are subject to inaccuracies for a variety of reasons, including students' expectations that they ought to be learning from their educational experiences (Bowman & Hill, 2011), their pre-college characteristics (Astin & Lee, 2003), and the extent to which they perceive that their educational context values student learning and development (Bowman, 2011; Pike, 1993). While self-reported data pose reliability and validity threats, the direct assessments of students' global perspective development is often not feasible, namely on the scale that larger survey efforts permit. In particular, social desirability bias may be relevant to the GPI for two reasons. First, it is well established that college students generally engage in "positive impression management" when they participate in research; they often consciously bias their responses to portray a favorable, positive self-image (Ferrari, Bristow, & Cowman, 2005, p. 9). Second, relative to how students perceive particular norms within their college contexts, indicating that one has made learning or developmental gains while in college is socially desirable because of the value higher education places on such development (Bowman & Hill, 2011). Previous work (e.g., Ferrari et al., 2005) suggests that as students' understanding of their institutional missions and values increases, so do their tendencies to respond to items about these aspects in socially desirable ways. In particular, higher education's focus around global learning and internationalization (Green, 2013), the public good (Kezar et al., 2004), and diversity more broadly conceived (Smith, 2009) serve as prominent values. The developmental dimensions measured by the GPI, in particular, are either explicitly or implicitly integrated into many institutions' missions and curricular/co-curricular

learning outcomes. Students' socialization around these norms complicates indirect ways of assessing their attitudes, learning, and development related to such norms.

Third, the theoretical background discussed in Chapter Two on the culturally variable nature of the developmental dimensions under study emphasized the role of students' educational contexts in examining these. In particular, the literature discussed the influence of where particular social identities are situated within specific contexts relative to understanding others and one's self in intergroup exchange. Students' perceptions of their campus climate, the amount and quality of intergroup interactions, and the structural ethnoracial diversity at their institutions all influence the extent to which developmental outcomes—such as those measured by the GPI—might be actualized in any given context (Milem et al., 2005). The present study does not examine measurement invariance across groups of students classified by particular psychological, behavioral, or structural aspects of the campus climate for key reasons outlined below.

Currently, the GPI does not include items that explicitly measure students' perceptions of their campus climate relative to ethnoracial diversity (e.g., the extent to which they perceive their campuses as welcoming, accommodating, divisive, or discriminatory toward their ethnoracial identity in particular). Similarly, the GPI does not include items that measure the *quality* of students' interracial/interethnic relationships or interactions on campus. Though the GPI General form includes both Sense of Belonging and Intercultural Engagement scales, the ways in which both scales are operationalized on the instrument limit their use in understanding the current sample's perceptions of their campus climates. For instance, the GPI's Sense of Belonging items do not measure students' perceptions related to the extent to which *their campus* is perceived as

welcoming, divisive, discriminatory, inclusive, or structured in a way that encourages meaningful intergroup exchange. On this scale, the only contextual measure involves students' perception about the extent to which they feel their campus honors diversity and internationalization. The other five measures on this scale are about *individuals'* perceptions (i.e., *they* understand the mission, *they* are both challenged/supported, *they* have been encouraged/supported) rather than about the broader campus or others within it. And while the Intercultural Engagement Scale measures the *number* of curricular and co-curricular engagements around multicultural and global/international topics, these items do not measure aspects related to the perceived quality or other psychological aspects related to these opportunities.

Finally, given the validation purpose of the present study, I had hoped at the outset (i.e., before obtaining the final dataset) that my group sizes would be large enough to randomly divide the total sample so that a subset could be used for initial model specification/re-specification and partial invariance testing (if necessary) while the remaining cases could be used for cross-sample validation (i.e., using the final re-specified model for invariance testing on an entirely different group). However, given the requisite group sizes for MGCFAs using the DWLS estimation method and number of estimated parameters in my models (Jöreskog & Sörbom, 1996), I could not risk potentially losing many cases for one of the groups (i.e., the total number of cases in the Hispanic group was already slightly below the recommended minimum). As such, I used the overall sample to determine the measurement model for this study and then imposed the final re-specified model for group-specific CFAs and MGCFAs on the same sample. I recognize this approach as a limitation since cross-sample validation provides stronger evidence that model re-

specification and fit did not result from sample-specific conditions. However, I carefully proceeded with model re-specification informed by the specific theoretical and empirical considerations outlined in the next chapter.

CHAPTER FOUR

RESULTS

This chapter presents the results from the confirmatory factor analyses (CFAs) and multiple-group confirmatory factor analyses (MGCFAs) executed to answer the study's seven research questions. I first evaluate the baseline measurement model used for invariance testing. Next, I present the results of the preliminary CFAs using both the aggregate sample and separate ethnoracial groups. I then present the invariance testing results (i.e., MGCFAs using all four groups at once). Finally, I present the findings related to the GPI's factor structure as well as convergent and discriminant validity evidence that could inform instrument refinement efforts. I use this chapter to detail findings related to my research questions and Chapter Five to more fully discuss nuances related to these findings and their associated implications.

Measurement Model Used for This Study

Based on earlier psychometric examination of the GPI (Davidson & Engberg, under review), in order to determine a baseline factor structure for invariance testing, I initially imposed Baseline Model 1 on the aggregate data (N = 7,017). This hierarchical model included a single second-order global perspective development factor, the six first-order developmental factors, and all 32 global perspective development items. I initially retained all 32 items in this model in alignment with how the GPI—and its developmental scales—are currently understood and reported by RISE (2017b). This initial CFA model did not converge. Typically, model convergence (i.e., successfully obtaining parameter estimates for one's model) occurs when the

model's estimation method yields results through an iterative process. The parameter estimates change from iteration to iteration, becoming more precise, until these changes are so small that estimates (along with their standard errors and statistical significance) are finally generated. When models do not converge, this often signifies that the imposed model does not fit the data well. Though I manually increased the number of allowable iterations (i.e., the number of times LISREL would permit the iterative process to run), the iterations stopped, and LISREL provided estimates through what it terms an intermediate solution rather than a final solution (Sörbom & Jöreskog, 1989). This intermediate solution presents estimates that are less accurate than would normally be produced and should not be used for reporting (Sörbom & Jöreskog, 1989). As such, this warranted exploring the dataset to determine any causes contributing toward poor model fit.

To understand model fit issues related to Baseline Model 1, I used principal axis factoring (PAF) to explore the underlying dimensionality of the dataset. PAF is an exploratory factor analytic approach that is useful in identifying latent constructs since it seeks to explain a dataset in terms of the number of factors that emerge by grouping correlated items together on the same factors. While the a priori factor structure used for Baseline Model 1 was informed by previous findings and the GPI's current scale construction, I aimed to explore whether any specific items did not appear associated with the constructs to which they should theoretically relate. Results from the PAF suggested multidimensionality for two items on the GPI's Cognitive Knowing scale, *I consider different cultural perspectives when evaluating global problems* and *I take into account different perspectives before drawing conclusions about the world around me*. Instead of loading on the Cognitive Knowing factor, where both are

theoretically situated, I observed in the PAF analysis that these two items loaded on both the Cognitive Knowledge and Intrapersonal Affect factors. Another item, *I am developing a meaningful philosophy of life*, did not load on any factor during the PAF analysis. Davidson and Engberg (under review) obtained nearly identical results in their earlier validation study. In the exploratory phase of their study, the same two items revealed multidimensionality across those same domains (and were therefore eliminated in their subsequent CFAs), and the last item—while retained in their CFAs—had a low factor loading (standardized $\lambda = .39$).

Given the PAF results—and that these aligned with earlier findings—I omitted the three items mentioned above, leaving a total of 29 GPI items for all subsequent CFA and MGCFA analyses. I imposed Baseline Model 2 on the aggregate sample data, which consisted of a single second-order factor, the six first-order factors, and 29 GPI items (see Figure 1). This re-specified model demonstrated marginal fit: SB $\chi^2(371, N = 7{,}017) = 10{,}926.15$, RMSEA = 0.06, SRMR = 0.09, CFI = 0.94, NNFI = 0.94 (see Table 5). Three of the four fit indices fell outside of the recommended thresholds that determine good model fit (i.e., RMSEA $\leq .06$, SRMR $\leq .08$, and CFI and NNFI $\geq .95$; Hu & Bentler, 1999), suggesting that model fit could be improved. Additionally, though statistically significant ($p < .01$), three first-order factor loadings were quite low with standardized loadings ranging from .10 to .29 in this baseline model (see Table 6), suggesting these items were not strong measures of the constructs to which they theoretically relate.

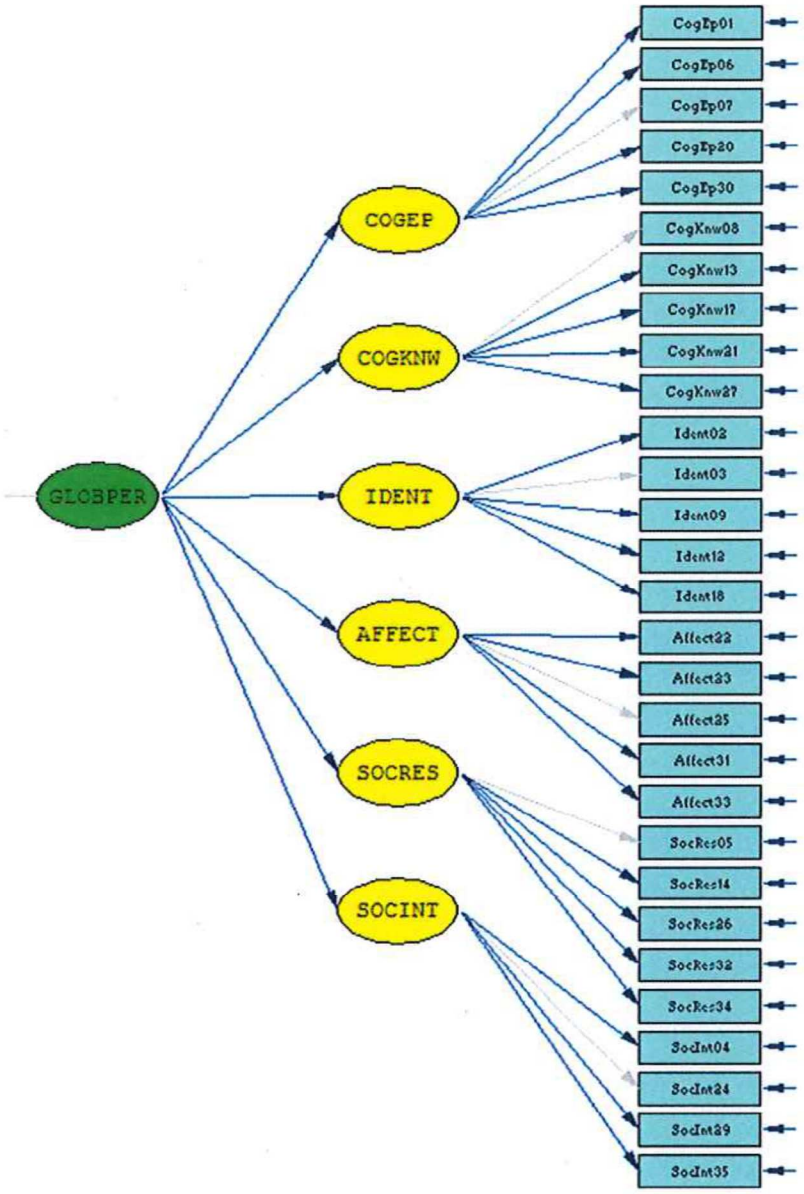

Figure 1. Conceptual diagram of the GPI's hierarchical factor structure using 29 items

Table 5. Overall and Group-specific Goodness-of-fit Statistics for Measurement Model Using Baseline Model 2 Hierarchical Factor Structure (using 29 GPI items)

Fit of Initial Baseline Hierarchical Factor Structure for Overall Samples						
			Measures of Absolute Fit		**Measures of Relative Fit**	
Group Tested for Baseline Models	**SB χ^2 [a]**	***df***	**RMSEA**	**SRMR**	**CFI**	**NNFI**
Aggregate Sample (N = 7,017)	10,926.148	371	0.0637	0.0850	0.942	0.937
Fit of Preliminary, Unconstrained Models for Separate Ethnoracial Groups						
			Measures of Absolute Fit		**Measures of Relative Fit**	
Group Tested for Preliminary Models	**SB χ^2**	***df***	**RMSEA**	**SRMR**	**CFI**	**NNFI**
Asian/Native Hawaiian/Other Pacific Islander (n = 675)	1,394.142	371	0.0640	0.0878	0.943	0.937
Black/African American (n=673)	1,426.649	371	0.0651	0.0921	0.952	0.948
Hispanic/Latinx (n =467)	1,061.450	371	0.0632	0.0853	0.946	0.941
White (n = 5,200)	8,696.753	371	0.0657	0.0876	0.935	0.929

Note. [a]The DWLS estimation method in LISREL 8.8 provides the Satorra-Bentler scaled chi-square test statistic (SB χ^2; Satorra & Bentler, 2001), which adjusts the goodness-of-fit chi-square statistic that is biased from multivariate nonnormality through a scaling correction factor that considers the multivariate kurtosis that biases this.

Table 6. Standardized First-order Factor Loadings, Second-order Coefficients, Variance Explained, Composite Reliability, and Average Variance Extracted for Baseline Model 2 for 29 Items Using the Aggregate Sample (N = 7,017)

GPI Item	Aggregate Sample	
COGNITIVE KNOWING	**λ**	**R^2**
COGEP01 – When I notice cultural differences, my culture tends to have the better approach. (r)	.58	.33
COGEP06 - Some people have culture and others do not. (r)	.24	.06
COGEP07 – In different settings what is right and wrong is simple to determine. (r)	.54	.29
COGEP20 – I rely primarily on authorities to determine what is true in the world. (r)	.52	.27
COGEP30 – I rarely question what I have been taught about the world around me. (r)	.45	.21
C.R.	**.59**	
AVE	**23.2%**	
COGNITIVE KNOWLEDGE	**λ**	**R^2**
COGKNW08 – I am informed of current issues that impact international relations.	.61	.37
COGKNW13 – I understand the reasons and causes of conflict among nations of different cultures.	.61	.37
COGKNW17 – I understand how various cultures of this world interact socially.	.74	.55
COGKNW21 – I know how to analyze the basic characteristics of a culture.	.62	.39
COGKNW27 – I can discuss cultural differences from an informed perspective.	.80	.65
C.R.	**.81**	
AVE	**46.6%**	
IDENTITY	**λ**	**R^2**
IDENT02 – I have a definite purpose in my life.	.58	.34
IDENT03 – I can explain my own personal values to people who are different from me.	.79	.62
IDENT09 – I know who I am as a person.	.69	.48
IDENT12 – I am willing to defend my views when they differ from others.	.51	.26
IDENT18 – I put my beliefs into action by standing up for my principles.	.75	.57
C.R.	**.80**	
AVE	**45.4%**	
AFFECT	**λ**	**R^2**
AFFECT22 – I am sensitive to those who are discriminated against.	.62	.39
AFFECT23 – I do not feel threatened emotionally when presented with multiple perspectives.	.54	.29
AFFECT25 – I am accepting of people with different religious and spiritual traditions.	.71	.50
AFFECT31 – I enjoy when my friends from other cultures teach me about our cultural differences.	.74	.54
AFFECT33 – I am open to people who strive to live lives very different from my own life style.	.74	.55
C.R.	**.81**	
AVE	**45.4%**	

GPI Item	Aggregate Sample	
SOCIAL RESPONSIBILITY	**λ**	**R^2**
SOCRES05 – I think of my life in terms of giving back to society.	.63	.40
SOCRES14 – I work for the rights of others.	.72	.51
SOCRES26 – I put the needs of others above my own personal wants.	.55	.30
SOCRES32 – I consciously behave in terms of making a difference.	.79	.62
SOCRES34 - Volunteering is not an important priority in my life. (r)	.10	.01
C.R.	**.71**	
AVE	**36.8%**	

GPI Item	Aggregate Sample	
SOCIAL INTERACTION	**λ**	**R^2**
SOCINT04 – Most of my friends are from my own ethnic background. (r)	.29	.08
SOCINT24 – I frequently interact with people from a race/ethnic group different from my own.	.72	.52
SOCINT29 – I intentionally involve people from many cultural backgrounds in my life.	.79	.63
SOCINT35 – I frequently interact with people from a country different from my own.	.63	.39
C.R.	**.71**	
AVE	**40.5%**	
FIRST-ORDER FACTOR (η)	**γ**	**R^2**
Cognitive Knowing	.01	.00
Cognitive Knowledge	.78	.61
Intrapersonal Identity	.62	.38
Intrapersonal Affect	.80	.65
Interpersonal Social Responsibility	.84	.71
Interpersonal Social Interaction	.79	.62
AVE	**49.5%**	

Note. (r) Indicates a reverse-worded item; these items were recoded so that a high mean score signifies more positive levels related to the specific dimension of development; λ = Standardized first-order factor loadings; all estimated loadings were statistically significant with $p < .01$; R^2 = Squared multiple correlation coefficient (the amount of variance explained in (1) each GPI item by its latent construct and (2) each first-order factor by the second-order factor); C.R. = composite reliability, the SEM approach for estimating scale reliability using the items' standardized factor loadings, error variance, and item R^2; AVE = Average variance extracted, or the average percentage of variation in the (1) items accounted for by their first-order latent constructs and (2) first-order factors accounted for by the second-order construct

The three items were: *Some people have culture and others do not* (Cognitive Knowing), *Volunteering is not an important priority in my life* (Social Responsibility), and *Most of my friends are from my own ethnic background* (Social Interaction). Theoretically, the latter two items are distinct from the other items on their respective scales. For instance, the other items on

the Social Responsibility scale explicitly ask about altruistic attitudes, whereas the item mentioned above asks about the importance of volunteering in one's life. These items seem to be measuring different dimensions of social responsibility (i.e., perceived civic *commitment* as opposed to civic *action*). Similarly, the other items on the Social Responsibility scale explicitly ask about interaction across *difference*, while the item mentioned above asks about *intra-ethnic* friendships. While the reverse-worded nature of this particular item presents it as the inverse of what the construct measures (i.e., it is technically a polar opposite reverse-worded item), it is noteworthy that this results in measuring something different compared to the other items on this scale, which could explain its low factor loading. All three of these items—along with four other items on the Cognitive Knowing scale—are reverse-worded items. In attempting to mitigate various response biases (i.e., acquiescence, inattention), these items often introduce confusion on the part of respondents and compromise validity (van Sonderen, Sanderman, & Coyne, 2013). Given that CFA is a theory-driven approach, I considered the above in deciding to remove these three items from remaining analyses.

After removing these three items, I imposed Baseline Model 3 on the aggregate sample data. This hierarchical factor structure consisted of a single second-order factor, the six first-order factors, and 26 GPI items (see Figure 2). Baseline Model 3 demonstrated better model fit compared to Baseline Model 2: SB $\chi^2(293, N = 7{,}017) = 6{,}455.78$, RMSEA = 0.05, SRMR = 0.07, CFI = 0.96, NNFI = 0.96 (see Table 7). All four fit indices fell within recommended standards to suggest good model fit. In addition, all standardized first-order factor loadings ranged from .41 to .81 and were statistically significant ($p < .01$), and five of the standardized second-order coefficients ranged from .62 to .82 and were also statistically significant ($p < .01$;

see Table 8). The relationship between the higher-order global perspective development factor and the Cognitive Knowing factor, though statistically significant ($p < .01$), was inverse and quite small (the standardized coefficient (γ) = -.07). Table 8 provides the standardized first-order factor loadings and corresponding squared multiple correlations (SMCs), while Table 9 provides the second-order coefficients and corresponding SMCs. The SMCs (R^2) for the individual GPI items reflect the extent to which the latent construct (or first-order factor) accounts for variance in the GPI items. In CFA analyses, the SMCs are a measure of the reliability of the items relative to their latent constructs; higher SMCs for an item reflect that more of its variability is related to the latent construct and not other unmeasured reasons. The SMCs (R^2) for the first-order factors reflect the extent to which the second-order global perspective development factor accounts for variance in the first-order factors. In other words, since the second-order factor is related to and used to explain the first-order factors, higher SMCs for the first-order factors indicate that the second-order factor is more defined by those first-order factors.

It is particularly useful to examine Table 9 for the SMCs for the first-order factors. Not surprisingly given the findings just mentioned, the second-order factor explains nothing in the Cognitive Knowing factor. These early observations about the Cognitive Knowing scale suggest theoretical and methodological implications that are discussed toward the end of this chapter. Given the theory-driven nature of CFA analytical approaches, I retained the Cognitive Knowing factor and four of its items in alignment with the theoretical underpinnings of the GPI's global perspective development construct. Considered together, the GPI's theoretical foundation along with the goodness-of-fit statistics, parameter estimates, and statistical significance, Baseline

Model 3 provides the best fit. Given this observation, I used this measurement model in all subsequent group-specific CFAs and MGCFAs discussed next.

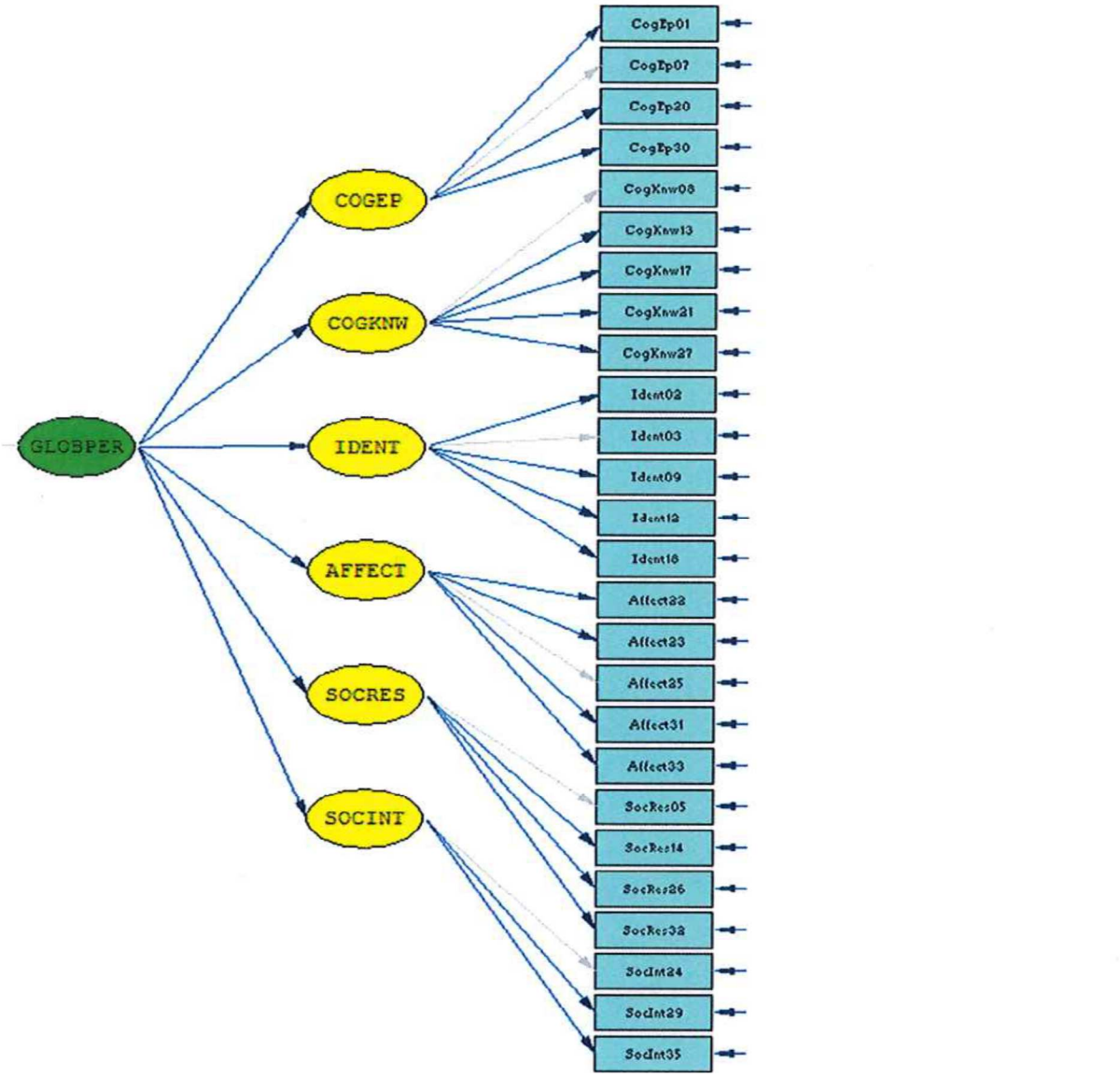

Figure 2. Conceptual diagram of the GPI's hierarchical factor structure using 26 items

Table 7. Overall and Group-specific Goodness-of-fit Statistics for Measurement Model Using Baseline Model 3 Hierarchical Factor Structure (using 26 GPI items)

Fit of Initial Baseline Hierarchical Factor Structure for Overall Samples						
Group Tested for Baseline Models	**SB χ^2 [a]**	***df***	**Measures of Absolute Fit**		**Measures of Relative Fit**	
			RMSEA	**SRMR**	**CFI**	**NNFI**
Aggregate Sample (N = 7,017)	6,455.780	293	0.0548	0.0678	0.963	0.959
Fit of Preliminary, Unconstrained Models for Separate Ethnoracial Groups						
Group Tested for Preliminary Models	**SB χ^2**	***df***	**Measures of Absolute Fit**		**Measures of Relative Fit**	
			RMSEA	**SRMR**	**CFI**	**NNFI**
Asian/Native Hawaiian/Other Pacific Islander (n = 675)	809.126	293	0.0511	0.0691	0.969	0.966
Black/African American (n=673)	870.579	293	0.0542	0.0731	0.972	0.969
Hispanic/Latinx (n =467)	723.411	293	0.0561	0.0728	0.964	0.960
White (n = 5,200)	5,203.037	293	0.0568	0.0723	0.958	0.954

Note. [a]The DWLS estimation method in LISREL 8.8 provides the Satorra-Bentler scaled chi-square test statistic (SB χ^2; Satorra & Bentler, 2001), which adjusts the goodness-of-fit chi-square statistic that is biased from multivariate nonnormality through a scaling correction factor that considers the multivariate kurtosis that biases this.

Table 8. Standardized First-order Factor Loadings, Variance Explained, Composite Reliability, and Average Variance Extracted for Baseline Model 3 for 26 Items for the Aggregate Sample and Separate Ethnoracial Groups (N = 7,017)

GPI Item	Aggregate Sample		Asian/Native Hawaiian/ Other Pacific Islander		Black/ African American		Hispanic		White	
COGNITIVE KNOWING	**λ**	**R^2**	**λ**	**R^2**	**λ**	**R^2**	**λ**	**R^2**	**λ**	**R^2**
COGEP01 – When I notice cultural differences, my culture tends to have the better approach.(r)	.54	.29	.39	.16	.54	.29	.48	.23	.56	.31
COGEP07 – In different settings what is right and wrong is simple to determine.(r)	.51	.26	.35	.12	.38	.15	.41	.17	.57	.32
COGEP20 – I rely primarily on authorities to determine what is true in the world.(r)	.62	.38	1.00	1.00	.84	.71	.60*	.36	.50	.25
COGEP30 – I rarely question what I have been taught about the world around me.(r)	.41	.17	.38	.14	.36	.13	.49	.24	.44	.19
C.R.	**.59**		**.64**		**.62**		**.57**		**.60**	
AVE	**27.6%**		**35.5%**		**32.0%**		**25.0%**		**26.8%**	
COGNITIVE KNOWLEDGE	**λ**	**R^2**	**λ**	**R^2**	**λ**	**R^2**	**λ**	**R^2**	**λ**	**R^2**
COGKNW08 – I am informed of current issues that impact international relations.	.61	.37	.60	.35	.60	.36	.61	.38	.61	.37
COGKNW13 – I understand the reasons and causes of conflict among nations of different cultures.	.61	.37	.64	.41	.56	.31	.63	.39	.61	.37
COGKNW17 – I understand how various cultures of this world interact socially.	.74	.55	.71	.51	.71	.51	.73	.54	.75	.57
COGKNW21 – I know how to analyze the basic characteristics of a culture.	.63	.39	.69	.47	.71	.50	.65	.42	.60	.36
COGKNW27 – I can discuss cultural differences from an informed perspective.	.80	.64	.76	.58	.78	.60	.82	.67	.80	.64
C.R.	**.81**		**.81**		**.81**		**.82**		**.81**	
AVE	**46.6%**		**46.4%**		**45.6%**		**48.0%**		**46.2%**	
IDENTITY	**λ**	**R^2**	**λ**	**R^2**	**λ**	**R^2**	**λ**	**R^2**	**λ**	**R^2**
IDENT02 – I have a definite purpose in my life.	.58	.34	.49	.24	.66	.44	.62	.38	.58	.33
IDENT03 – I can explain my own personal values to people who are different from me.	.78	.61	.75	.56	.72	.52	.77	.59	.79	.63
IDENT09 – I know who I am as a person.	.69	.48	.66	.43	.73	.53	.64	.42	.69	.48

GPI Item	Aggregate Sample		Asian/Native Hawaiian/ Other Pacific Islander		Black/ African American		Hispanic		White	
IDENTITY (cont'd)	**λ**	**R^2**	**λ**	**R^2**	**λ**	**R^2**	**λ**	**R^2**	**λ**	**R^2**
IDENT12 – I am willing to defend my views when they differ from others.	.51	.26	.43	.18	.53	.28	.49	.24	.53	.28
IDENT18 – I put my beliefs into action by standing up for my principles.	.75	.56	.65	.43	.77	.60	.70	.49	.76	.58
C.R.	**.80**		**.74**		**.82**		**.78**		**.81**	
AVE	**44.9%**		**36.8%**		**47.4%**		**42.4%**		**46.0%**	
AFFECT	**λ**	**R^2**	**λ**	**R^2**	**λ**	**R^2**	**λ**	**R^2**	**λ**	**R^2**
AFFECT22 – I am sensitive to those who are discriminated against.	.62	.39	.53	.28	.59	.34	.56	.31	.64	.41
AFFECT23 – I do not feel threatened emotionally when presented with multiple perspectives.	.55	.30	.54	.29	.60	.36	.45	.20	.56	.32
AFFECT25 – I am accepting of people with different religious and spiritual traditions.	.71	.50	.75	.56	.73	.53	.72	.52	.69	.48
AFFECT31 – I enjoy when my friends from other cultures teach me about our cultural differences.	.73	.54	.71	.50	.80	.64	.71	.50	.73	.53
AFFECT33 – I am open to people who strive to live lives very different from my own life style.	.74	.55	.67	.45	.80	.63	.79	.62	.73	.54
C.R.	**.80**		**.78**		**.83**		**.79**		**.80**	
AVE	**45.4%**		**41.6%**		**50.0%**		**43.0%**		**45.6%**	
SOCIAL RESPONSIBILITY	**λ**	**R^2**	**λ**	**R^2**	**λ**	**R^2**	**λ**	**R^2**	**λ**	**R^2**
SOCRES05 – I think of my life in terms of giving back to society.	.63	.40	.64	.41	.62	.39	.62	.38	.62	.38
SOCRES14 – I work for the rights of others.	.72	.52	.73	.53	.68	.46	.68	.47	.74	.54
SOCRES26 – I put the needs of others above my own personal wants.	.55	.30	.51	.26	.56	.32	.42	.18	.58	.34
SOCRES32 – I consciously behave in terms of making a difference.	.79	.63	.67	.44	.83	.68	.81	.65	.80	.64
C.R.	**.77**		**.74**		**.77**		**.73**		**.78**	
AVE	**46.1%**		**41.0%**		**46.3%**		**42.0%**		**47.5%**	

GPI Item	Aggregate Sample		Asian/Native Hawaiian/ Other Pacific Islander		Black/ African American		Hispanic		White	
SOCIAL INTERACTION	**λ**	**R^2**	**λ**	**R^2**	**λ**	**R^2**	**λ**	**R^2**	**λ**	**R^2**
SOCINT24 – I frequently interact with people from a race/ethnic group different from my own.	.74	.55	.67	.45	.69	.47	.74	.54	.73	.53
SOCINT29 – I intentionally involve people from many cultural backgrounds in my life.	.81	.66	.76	.58	.77	.59	.71	.51	.83	.68
SOCINT35 – I frequently interact with people from a country different from my own.	.65	.42	.71	.50	.63	.39	.60	.36	.63	.40
C.R.	**.78**		**.76**		**.74**		**.73**		**.78**	
AVE	**54.2%**		**51.0%**		**48.3%**		**47.0%**		**53.7%**	

Note. [(r)] Indicates a reverse-worded item; these items were recoded so that a high mean score signifies more positive levels related to the specific dimension of development; λ = Standardized first-order factor loadings (* = estimated loadings were statistically non- significant with $p > .05$; all other estimated loadings were statistically significant with $p < .05$); R^2 = Squared multiple correlation coefficient (the amount of variance explained in each GPI item by its latent construct; C.R. = composite reliability, the SEM approach for estimating scale reliability using the items' standardized factor loadings, error variance, and item R^2; AVE = Average variance extracted, or the average percentage of variation in the items accounted for by their first-order latent constructs

Table 9. Standardized Second-order Coefficients, Variance Explained, and Average Variance Extracted for Baseline Model 3 for Aggregate Sample and Separate Ethnoracial Groups (N = 7,017)

First-order Factor (η)	Aggregate Sample		Asian/Native Hawaiian/ Other Pacific Islander		Black/ African American		Hispanic		White	
	γ	**R^2**	**γ**	**R^2**	**γ**	**R^2**	**γ**	**R^2**	**γ**	**R^2**
Cognitive Knowing	-.07	.01	-.16	.03	-.26	.07	-.02*	.01	.02*	.00
Cognitive Knowledge	.80	.63	.89	.78	.87	.76	.82	.67	.76	.58
Intrapersonal Identity	.62	.39	.83	.68	.75	.57	.74	.55	.58	.33
Intrapersonal Affect	.79	.62	.82	.68	.81	.66	.87	.75	.78	.60
Interpersonal Social Responsibility	.82	.68	.78	.61	.88	.77	.86	.74	.82	.66
Interpersonal Social Interaction	.75	.57	.80	.63	.79	.63	.82	.67	.74	.55
AVE	**48.3%**		**56.8%**		**57.7%**		**56.5%**		**45.3%**	

Note. η = First-order factor; γ = Standardized second-order coefficient (* = estimated coefficients were statistically non- significant with $p > .05$; all other estimated coefficients were statistically significant with $p < .01$); R^2 = Squared multiple correlation coefficient (the amount of variance explained in each first-order factor by the second-order factor); AVE = Average variance extracted, or the average percentage of variation in the first-order factors accounted for by the second-order construct

Preliminary Baseline Models

After determining the best model fit for the aggregate sample, I imposed the factor structure of Baseline Model 3 on each of the four separate ethnoracial groups to evaluate the models' goodness of fit. All fit indices fell within ranges that suggest good model fit across all four ethnoracial groups; see Table 7 for the goodness-of-fit statistics for these group-specific models. The first-order factor loading for one item on the Cognitive Knowing scale, *I rely primarily on authorities to determine what is true in the world*, was statistically non-significant for the Hispanic group. All other first-order factor loadings were statistically significant ($p <$ .05) across all four groups (see Table 8). Additionally, the relationship between the higher-order global perspective development factor and the Cognitive Knowing factor was statistically non-significant for both the Hispanic and white groups and while statistically significant across the other two groups, this coefficient was quite low and comparable to the aggregate sample's results above (see Table 9). These early observations suggest no relationship between the Cognitive Knowing factor and higher-order global perspective development factor. In other words, this empirical evidence suggests that global perspective development is not at least partially defined by the dimension of development measured by the Cognitive Knowing scale. This is theoretically unexpected, and I will discuss this particular finding in much more detail later in the chapter. Given the overall findings related to Baseline Model 3, these group-specific baseline models suggested that I could commence with invariance testing.

Measurement Invariance Testing

The findings related to invariance testing are presented below. Given the subsampling approach employed to address group size imbalances, I present results from both the original

models run (i.e., using the entire white sample) and the second set of models run with the random subsample of white students for each research question.

Research Q1: Is Equal Form Observed Across Ethnoracial Groups?

As described in Chapter Three, invariance testing begins with the least constrained model (i.e., complete lack of invariance) and sequentially imposes more equality constraints on particular model parameters to determine whether the model fit worsens by such impositions. The study's first six research questions were structured to test—in order—the different types of invariance potentially present in a higher-order factor structure. Results from each invariance test are presented below to answer each research question.

To answer the first research question, I examined configural invariance across the four ethnoracial groups to determine whether there was equal form (i.e., same number of factors and an identical pattern of indictor-factor loadings) across groups. To examine this, I imposed Model 1, which constrained only the pattern of factor loadings as equal across all four groups; all other parameters were freely estimated. Table 10 and Table 11 provide all goodness-of-fit statistics for each gradually constrained model (Model 1 – Model 7) for the overall sample and random subsample, respectively. Model 1 demonstrated good model fit for both the overall, SB χ^2(1,172, N = 7,017) = 7,701.741, RMSEA = 0.056, SRMR = 0.073, CFI = 0.959, NNFI = 0.954, and random samples, SB χ^2(1,172, N = 2,581) = 3,388.260, RMSEA = 0.054, SRMR = 0.073, CFI = 0.965, NNFI = 0.961. All 26 of the unstandardized first-order factor loadings were statistically significant across all four groups with the exception of the single factor loading for the Cognitive Knowing item mentioned in the group-specific baseline model results above; this factor loading was statistically non-significant for the Hispanic group, and the relationship between the higher-

order global perspective development factor and the Cognitive Knowing factor was statistically non-significant for both the Hispanic and white groups. The goodness-of-fit statistics for both samples indicated good model fit, suggesting configural invariance. I concluded that the same factor structure (i.e., equal form) for the GPI holds across the four groups and proceeded to answer my second research question.

Table 10. Goodness-of-fit Statistics for Evaluating Measurement Invariance (using 26 GPI items) for Overall Sample (N = 7,017)

Model Fit for Gradually Constrained Models									
						Measures of Absolute Fit		Measures of Relative Fit	
Model Tested	**Models Compared**	**ΔCFI[a]**	**Scaled χ^2 Difference Test**	**SB χ^2 [b]**	***df***	**RMSEA[c]**	**SRMR[c]**	**CFI[c]**	**NNFI[c]**
Model 1: Invariant pattern of factor loadings	N/A	N/A	N/A	7,701.741	1,172	0.0564	0.0728	0.959	0.954
Model 2: Invariant first-order factor loadings	Model 2 – Model 1	-.001	$p > .05$	7,904.164	1,232	0.0556	0.0767	0.958	0.955
Model 3: Invariant second-order coefficients	Model 3 – Model 2	-.002	Inadmissible[d]	8,170.402	1,250	0.0562	0.0801	0.956	0.954
Model 4: Invariant item response category thresholds	Model 4 – Model 3	-.001	$p > .05$	8,396.594	1,328	0.0551	0.0801	0.955	0.956
Model 5: Invariant first-order factor disturbances	Model 5 – Model 4	.000	$p < .01$	8,371.088	1,346	0.0546	0.0812	0.955	0.957
Model 6: Invariant error variances of	Model 6 – Model 5	+.001	$p > .05$	8,298.884	1,424	0.0525	0.0812	0.956	0.960

GPI items									
Model 7: Invariant second-order factor mean	Model 7 – Model 5	.000	$p > .05$	8,290.731	1,421	0.0525	0.0812	0.956	0.960

Note. [a]ΔCFI < -.01 indicates lack of invariance (Meade et al., 2008). [b]The DWLS estimation method in LISREL 8.8 provides the Satorra-Bentler scaled chi-square test statistic (SB χ^2; Satorra & Bentler, 2001), which adjusts the goodness-of-fit chi-square statistic that is biased from multivariate nonnormality through a scaling correction factor that considers the multivariate kurtosis that biases this. [c]I follow SEM reporting practices outlined by Hu and Bentler (1999), using measures of absolute fit (i.e., RMSEA, SRMR) and relative fit (i.e., CFI, NNFI), using RMSEA ≤ .06, SRMR ≤ .08, and CFI and NNFI ≥ .95 as stringent determinants of model fit. [d]While using Bryant and Satorra's (2013) macro to generate a scaled chi-square difference test from LISREL results, I obtained negative differences in chi-square values—when comparing Model 2 to Model 1 using the overall sample. I employed the improved scaling correction procedure and new scaled difference test (Satorra & Bentler, 2010; see Bryant & Satorra, 2012 for a detailed discussion on the development and execution of this formula). However, in doing so I obtained inadmissible scaling correction factors for the new scaled difference test in testing the fit of Model 2 compared to Model 1 in the overall sample given a technical anomaly in LISREL 8.8 discussed by Bryant and Satorra (2012).

Table 11. Goodness-of-fit Statistics for Evaluating Measurement Invariance Using Random White Subsample (N = 2,581)

Model Fit for Gradually Constrained Models									
						Measures of Absolute Fit		**Measures of Relative Fit**	
Model Tested	**Models Compared**	**ΔCFI[a]**	**Scaled χ^2 Difference Test**	**SB χ^2 [b]**	***df***	**RMSEA[c]**	**SRMR[c]**	**CFI[c]**	**NNFI[c]**
Model 1: Invariant pattern of factor loadings	N/A	N/A	N/A	3,388.260	1,172	0.0542	0.0728	0.965	0.961
Model 2: Invariant first-order factor loadings	Model 2 – Model 1	-.001	$p < .01$	3,518.587	1,232	0.0537	0.0757	0.964	0.962
Model 3: Invariant second-order coefficients	Model 3 – Model 2	-.001	$p < .01$	3,587.806	1,250	0.0539	0.0775	0.963	0.962
Model 4: Invariant item response category thresholds	Model 4 – Model 3	.000	$p > .05$	3,689.588	1,328	0.0525	0.0775	0.963	0.963
Model 5:	Model 5 –	-.001	$p = .02$	3,719.857	1,346	0.0523	0.0782	0.962	0.964

Invariant first-order factor disturbances	Model 4								
Model 6: Invariant error variances of GPI items	Model 6 – Model 5	+.001	$p > .05$	3,746.769	1,424	0.0503	0.0782	0.963	0.966
Model 7: Invariant second-order factor mean	Model 7 – Model 5	+.001	$p > .05$	3,743.079	1,421	0.0504	0.0782	0.963	0.966

Note. [a]ΔCFI < -.01 indicates lack of invariance (Meade et al., 2008). [b]The DWLS estimation method in LISREL 8.8 provides the Satorra-Bentler scaled chi-square test statistic (SB χ^2; Satorra & Bentler, 2001), which adjusts the goodness-of-fit chi-square statistic that is biased from multivariate nonnormality through a scaling correction factor that considers the multivariate kurtosis that biases this. [c]I follow SEM reporting practices outlined by Hu and Bentler (1999), using measures of absolute fit (i.e., RMSEA, SRMR) and relative fit (i.e., CFI, NNFI), using RMSEA ≤ .06, SRMR ≤ .08, and CFI and NNFI ≥ .95 as stringent determinants of model fit.

Research Q2: Are the First-order Factor Loadings and Second-order Coefficients Invariant Across Ethnoracial Groups?

Invariance of first-order factor loadings. To answer the second research question, I examined metric invariance across the four ethnoracial groups to determine whether the strength of relationships between the observed indicators and the factors to which they relate were the same. Evidence of metric invariance suggests the ethnoracial groups under study understand the measured constructs equivalently. To examine this, I imposed Model 2, which constrained all first-order factor loadings as equal across the four groups; all other parameter estimates were freely estimated.

In order to evaluate model fit after constraining the first-order factor loadings as invariant across groups, I relied on ΔCFI < -.01 (a decrease in CFI larger than .01) as a determinant of model fit worsening (i.e., indicating that the parameters under examination are *not* invariant;

Cheung & Rensvold, 2002; Meade et al., 2008). To obtain the ΔCFI, the CFI value of the nested, more restrictive model (i.e., fewer estimated parameters, higher *df*) is compared to the CFI value obtained with the baseline, less restrictive model (i.e., more estimated parameters, smaller *df*). I report both ΔCFI and the scaled chi-square difference test results in Table 10 and Table 11. For each model, I used Bryant and Satorra's (2013) macro to generate a scaled chi-square difference test from LISREL results to test whether the nested models' fits significantly worsened compared to the baseline models. However, large subsample sizes can incorrectly inflate SBχ^2 values; it is quite common to obtain statistically significant test statistics with any sample > 400 (de Beurs, Fokkema, de Groot, de Keijser, & Kerkhof, 2015), and this statistic is also sensitive to model complexity (Brown 2015). Both of these issues are salient to the present study given the model complexity (i.e., hierarchical structure, number of factors and observed indicators) and that all group sizes are > 400. The ΔCFI between the nested model (Model 2) and baseline model (Model 1) was -.001 for both the overall and random samples, suggesting negligible decrement in model fit. The scaled chi-square difference test was statistically significant ($p < .01$) in comparing Model 2 to Model 1 for the random sample (this statistic was statistically non-significant for the overall sample). However, considering the aforementioned information about large sample sizes and the scaled chi-square test statistic—and that the CFI is relatively unaffected by sample size—I relied on the ΔCFI to support that the first-order factor loadings appear invariant across the four groups.

Invariance of second-order coefficients. Given the hierarchical factor structure of the GPI, answering this second research question also necessitated testing the invariance of the second-order coefficients across groups**.** Doing so tests whether relationships between the six

first-order factors and the single second-order factor are equivalent across the four groups. For Model 3, I constrained all first-order factor loadings and all second-order coefficients as equal across the four groups. The ΔCFI between the nested model (Model 3) and baseline model (Model 2) was -.002 and -.001 for the overall and random samples, respectively (see Table 10 and Table 11). The scaled chi-square difference test was inadmissible for the overall sample. While using Bryant and Satorra's (2013) macro to generate a scaled chi-square difference test, I obtained negative differences in chi-square values—when comparing Model 2 to Model 1 using the overall sample. I employed the improved scaling correction procedure and new scaled difference test (Satorra & Bentler, 2010; see Bryant & Satorra, 2012 for a detailed discussion on the development and execution of this formula). However, in doing so I obtained inadmissible scaling correction factors for the new scaled difference test in testing the fit of Model 2 compared to Model 1 in the overall sample given a technical anomaly in LISREL 8.8 discussed by Bryant and Satorra (2012). The scaled chi-square difference test was admissible and statistically significant ($p < .01$) for the random sample; however, I relied on the ΔCFI to conclude that constraining the first-order factor loadings as well as the second-order coefficients as invariant across all four groups did not worsen model fit. As such, I concluded that the first-order factor loadings and second-order coefficients are invariant across the four groups. The four groups appear to understand the GPI's developmental constructs equivalently, and the relationships between the higher-order global perspective development construct and the six first-order factors also appear to be equivalent across groups.

Research Q3: Are the GPI's Item Thresholds Invariant Across Ethnoracial Groups?

In models with ordinal indicators, invariance related to the thresholds—or item response category distribution cut points—must also be tested to determine whether particular items function differently across the groups under study. More specifically, testing for threshold invariance determines whether individuals' levels of the latent constructs correspond to the same response choices on the items related to those constructs. In other words, groups with the same level of development should have the same probability of providing a certain response; if this is not the case for a given item, one can conclude that the item systematically functions differently as a condition of their group membership.

Table 12. Item Thresholds for 26 GPI Items for All Four Ethnoracial Groups

GPI Item	White Sample				ANHOPI[a] Sample				Black Sample				Hispanic Sample			
Cognitive Knowing	**τ1**	**τ2**	**τ3**	**τ4**	**τ1**	**τ2**	**τ3**	**τ4**	**τ1**	**τ2**	**τ3**	**τ4**	**τ1**	**τ2**	**τ3**	**τ4**
CogEp01	-1.63	-0.71	0.57	1.59	-1.32	-0.44	0.66	1.70	-1.16	-0.39	0.80	1.60	-1.49	-0.70	0.56	1.50
CogEp07	-1.36	-0.35	0.28	1.43	-1.55	-0.62	-0.06	0.93	-0.90	-0.02	0.75	1.63	-1.29	-0.25	0.34	1.30
CogEp20	-1.97	-0.98	-0.12	1.08	-1.59	-0.66	0.20	1.31	-1.60	-0.86	-0.19	0.77	-1.67	-0.90	-0.01	1.03
CogEp30	-1.93	-1.01	-0.34	0.94	-1.59	-0.86	-0.12	0.99	-1.38	-0.67	-0.10	0.82	-1.74	-0.80	-0.17	0.94
Cognitive Knowledge	**τ1**	**τ2**	**τ3**	**τ4**	**τ1**	**τ2**	**τ3**	**τ4**	**τ1**	**τ2**	**τ3**	**τ4**	**τ1**	**τ2**	**τ3**	**τ4**
CogKnw08	-2.10	-1.01	-0.32	1.07	-2.22	-1.15	-0.39	0.98	-2.22	-1.29	-0.49	0.75	-2.38	-1.22	-0.42	0.84
CogKnw13	-2.33	-1.27	-0.37	1.22	-2.52	-1.51	-0.54	1.00	-2.44	-1.42	-0.45	0.87	-2.30	-1.47	-0.53	0.96
CogKnw17	-2.65	-1.40	-0.45	1.18	-2.97	-1.77	-0.72	0.77	-2.26	-1.56	-0.59	0.73	-2.63	-1.72	-0.71	0.72
CogKnw21	-2.65	-1.48	-0.43	1.23	-2.52	-1.38	-0.39	1.01	-2.62	-1.72	-0.42	0.89	-2.63	-1.56	-0.45	1.07
CogKnw27	-2.62	-1.55	-0.59	0.95	-2.97	-1.79	-0.76	0.81	-2.75	-1.79	-0.76	0.58	-2.63	-1.88	-0.73	0.63
Identity	**τ1**	**τ2**	**τ3**	**τ4**	**τ1**	**τ2**	**τ3**	**τ4**	**τ1**	**τ2**	**τ3**	**τ4**	**τ1**	**τ2**	**τ3**	**τ4**
Ident02	-2.38	-1.56	-0.90	0.29	-2.22	-1.26	-0.47	0.70	-2.44	-1.84	-1.14	-0.14	-2.30	-1.63	-0.93	0.32
Ident03	-3.00	-2.21	-1.42	0.39	-1.82	-1.10	0.52		-2.97	-2.62	-1.43	0.02	-2.63	-1.91	-1.35	0.30
Ident09	-2.67	-1.81	-1.10	0.40	-2.75	-1.72	-0.80	0.59	-2.52	-1.93	-1.09	0.06	-2.49	-1.74	-1.06	0.30
Ident12	-2.74	-1.99	-0.84	0.80	-2.52	-1.44	-0.41	1.01	-2.97	-1.80	-0.86	0.47	-2.86	-1.72	-0.66	0.75
Ident18	-3.10	-2.09	-0.87	0.84	-2.97	-1.96	-0.67	0.94	-2.75	-2.17	-0.93	0.49	-2.86	-2.12	-0.90	0.71
Affect	**τ1**	**τ2**	**τ3**	**τ4**	**τ1**	**τ2**	**τ3**	**τ4**	**τ1**	**τ2**	**τ3**	**τ4**	**τ1**	**τ2**	**τ3**	**τ4**
Affect22	-2.56	-1.78	-0.85	0.66	-2.75	-1.79	-0.70	0.65	-2.52	-1.73	-0.99	0.01	-2.49	-1.91	-1.04	0.32

Affect23	-2.59	-1.72	-0.98	0.65	-2.37	-1.55	-0.66	0.83	-2.31	-1.55	-0.84	0.49	-2.63	-1.70	-0.87	0.52
Affect25	-2.89	-2.20	-1.35	0.17	-2.97	-2.26	-1.38	0.12	-2.97	-2.37	-1.39	-0.17	-2.86	-2.38	-1.77	-0.29
Affect31	-2.74	-2.17	-1.17	0.31	-2.75	-2.44	-1.44	-0.01	-2.52	-1.30	-0.07		-2.49	-1.46	-0.05	
Affect33	-2.83	-1.94	-0.90	0.75	-2.97	-2.31	-1.04	0.45	-2.22	-1.04	0.32		-2.30	-1.24	0.42	
Social Responsibility	**τ1**	**τ2**	**τ3**	**τ4**	**τ1**	**τ2**	**τ3**	**τ4**	**τ1**	**τ2**	**τ3**	**τ4**	**τ1**	**τ2**	**τ3**	**τ4**
SocRes05	-2.29	-1.32	-0.29	1.02	-2.52	-1.37	-0.31	0.87	-2.75	-1.75	-0.63	0.40	-2.63	-1.65	-0.73	0.49
SocRes14	-2.47	-1.54	-0.21	1.11	-2.52	-1.49	-0.10	1.17	-2.26	-1.57	-0.33	0.81	-2.86	-1.63	-0.36	0.83
SocRes26	-2.39	-1.56	-0.44	0.91	-2.04	-1.19	-0.01	1.05	-2.17	-1.41	-0.36	0.65	-2.49	-1.43	-0.35	0.79
SocRes32	-2.62	-1.72	-0.47	1.02	-2.37	-1.50	-0.41	0.90	-1.98	-0.69	0.52		-2.63	-2.07	-0.70	0.71
Social Interaction	**τ1**	**τ2**	**τ3**	**τ4**	**τ1**	**τ2**	**τ3**	**τ4**	**τ1**	**τ2**	**τ3**	**τ4**	**τ1**	**τ2**	**τ3**	**τ4**
SocInt24	-2.33	-1.15	-0.50	0.71	-2.26	-1.36	-0.65	0.41	-2.75	-1.91	-1.08	-0.03	-2.86	-2.03	-1.21	0.08
SocInt29	-2.27	-1.04	0.13	1.17	-2.62	-1.38	-0.42	0.74	-2.44	-1.46	-0.47	0.49	-2.38	-1.40	-0.46	0.67
SocInt35	-1.76	-0.52	0.28	1.19	-2.14	-1.14	-0.41	0.62	-2.10	-1.01	-0.22	0.62	-1.95	-1.18	-0.45	0.61

Note. τ = thresholds, which define the ranges of true scores between ordinal response options. The GPI has five ordinal response options (1,2,3,4,5), so there are four thresholds for each item (τ1, τ2, τ3, and τ4). [a] ANHOPI = Asian, Native Hawaiian, Other Pacific Islander

For reference, the item thresholds for all 26 GPI items from the four separate ethnoracial groups are included in Table 12. To test whether these thresholds were invariant across groups, I imposed Model 4, which constrained all first-order factor loadings, all second-order coefficients, and the item thresholds for all 26 GPI items as equal across the four groups. The scaled chi-square difference test indicated a statistically non-significant change from the baseline model (Model 3) for both the overall and random samples. The ΔCFI between the nested model (Model 4) and baseline model (Model 3) was -.001 and .000 for the overall and random samples, respectively (see Table 10 and Table 11). These findings suggest that constraining the first-order factor loadings, second-order coefficients, and all item thresholds as invariant across all four groups did not worsen model fit. I can conclude that these four groups' level of development (as measured by the corresponding GPI items) corresponds to equivalent response choices across these items. The items do not appear to be systematically biased or operate differently as a function of ethnoracial group membership.

Research Q4: Are Equal Disturbances of the First-order Factors Observed Across Ethnoracial groups?

The GPI's hierarchical factor structure requires an examination of the equality of disturbances of the first-order factors (i.e., the variance in the GPI's six first-order factors not explained by the single second-order factor). This invariance suggests that the amount of any unexplained variance relative to the first-order factors is equivalent across groups. It is useful to examine Table 9 to observe the second-order parameter estimates from all four groups. Related to this particular research question, examining the amount of *explained* variance in each first-order factor (i.e., included in Table 9 as R^2) provides detail about the amount of *unexplained* variance in these factors since $1 - R^2$ = unexplained variance; the lower the R^2 value, the higher the amount of unexplained variance or factor disturbance. There are interesting differences within the aggregate sample relative to the amount of unexplained variance in each first-order factor. In the aggregate sample, the amount of variance unrelated to the higher-order factor (using results from the standardized solution) varied quite a bit across the first-order factors. For instance, the amount of unrelated variance for the Cognitive Knowing (99.5%) and Intrapersonal Identity (60.8%) factors was quite high compared to the Social Interaction (43.4%), Affect (37.7%), Cognitive Knowledge (36.6%), and Social Responsibility (32.2%) factors. In looking across the four ethnoracial groups, at first glance there appears to be the most variation relative to the Identity factor's R^2 (and, therefore, its factor disturbance). But when constraining all factor disturbances as invariant across groups, was there evidence that these differed across groups?

To test for this type of invariance, I imposed Model 5, which constrained all first-order factor loadings, all second-order coefficients, all item thresholds, and the six factor disturbances as equal across the four groups. Though the scaled chi-square difference test indicated a statistically significant change ($p < .05$) from the baseline model (Model 4) for both the overall and random samples, the ΔCFI between the nested model (Model 5) and baseline model (Model 4) was .000 and -.001 for the overall and random samples, respectively (see Table 10 and Table 11). These findings suggest that constraining the first-order factor loadings, second-order coefficients, item thresholds, and factor disturbances as invariant across all four groups did not worsen model fit. I concluded that the amount of unexplained variance in the first-order factors was invariant across the four groups.

Research Q5: Are Equal Item Error Variances Observed Across Ethnoracial Groups?

Item error variance (or unique variance) is the variance in an item that is not explained by the first-order factor. Unique variance involves both *reliable* variance specific to a particular item (i.e., other factors that systematically influence responses to that item) and *random* error variance (i.e., measurement error or unreliability in the item; Brown, 2015). Invariant item error variances suggest that for each GPI item, the latent constructs are measured equivalently across groups. Similar to the last research question, examining the first-order parameter estimates across groups (see Table 8), one can see the amount of *explained* variance in each GPI item (i.e., included in Table 8 as R^2). This, therefore, provides detail about the amount of *unexplained* variance in these items since $1 - R^2$ = unexplained variance.

To test whether the unexplained variance in the GPI items was invariant, I imposed Model 6, which constrained all first-order factor loadings, all second-order coefficients, all item

thresholds, all six factor disturbances, and all GPI item error variances as equal across the four groups. The scaled chi-square difference test indicated a statistically non-significant change ($p > .05$) from the baseline model (Model 5) for both the overall and random samples. The ΔCFI between the nested model (Model 6) and baseline model (Model 5) was .001 for both samples (see Table 10 and Table 11). These findings suggest that constraining the first-order factor loadings, second-order coefficients, item thresholds, factor disturbances, and error variances as invariant across all four groups did not worsen model fit. I concluded that the unexplained variance in the observed indicators was invariant across the four groups and that the GPI items are measured equivalently across groups.

Research Q6: Do Significant Cross-group Differences Relative to the GPI's Second-order Factor Mean Exist?

Finally, after testing for measurement invariance of a higher-order factor structure, one can examine cross-group equivalency of the higher-order factor's mean. In hierarchical factor structures, the first-order factor means are contingent on their second-order factor means. In terms of the GPI, it is important to note that the hypothesized single second-order factor accounts for the GPI's six first-order factors (i.e., the second-order factor explains the variance and covariance associated with the first-order factors). As such, the GPI's six first-order factor means are contingent on the single second-order factor mean and cannot be directly compared in this hierarchical model. Examining the equivalence of the GPI's single second-order factor mean across groups provides evidence to determine whether students' level of global perspective development differs across groups.

To test for this type of invariance, I imposed Model 7, which constrained all first-order factor loadings, all second-order coefficients, all item thresholds, all six factor disturbances, and all GPI item error variances as equal across the four groups (identical constraints as Model 6). However, I also fixed the second-order factor mean for the white group (the arbitrary reference group) to zero and allowed the remaining three groups' (comparison groups') second-order factor mean to freely estimate. The ΔCFI between the nested model (Model 7) and baseline model (Model 5) was .001 for both overall and random samples and the scaled chi-square difference test indicated a statistically non-significant change ($p > .05$) for both the overall and random samples (see Table 10 and Table 11). Given the negligible change in model fit, I did not need to change the reference group to continue to test across groups since the ΔCFI essentially indicated no change, therefore suggesting no differences in the mean level of the second-order global perspective development factor across the four groups.

Summary of Measurement Invariance Results

Considered together, these results suggest strict measurement invariance (invariant first-order factor loadings, second-order coefficients, item thresholds, item error variances, and factor disturbances) of the GPI's global perspective development items across the four ethnoracial groups. Table 13 includes a tabular summary of the study's measurement invariance results. This evidence suggests that these four ethnoracial groups understand the measured constructs equivalently and demonstrate equivalent correspondence between levels of the underlying constructs and their response choices. This validity evidence allows one to compare GPI mean scores and regression coefficients across ethnoracial groups. Additionally, these findings suggest that the GPI's latent constructs are measured equivalently across these four groups; the

amount of extraneous influence on the response process does not differ as a condition of ethnoracial group membership. Finally, the results suggest structural invariance, or the same higher-order factor structure and higher-order factor mean, across the four groups.

Table 13. Summary of Measurement Invariance Testing Results

Type of Invariance Tested	Key Question Answered	Finding/Interpretation
RQ1: Configural invariance	Can equivalent constructs be measured across groups?	**Finding:** The same hierarchical factor structure applies to all four groups. **Interpretation:** The GPI measures the same developmental constructs across the four groups.
RQ2: Metric invariance	Do different groups understand the measured survey constructs in the same way?	**Finding:** The first-order factor loadings and second-order coefficients are equivalent across groups. **Interpretation:** Relationships between the GPI items and the developmental constructs (and between the developmental constructs and higher-order global perspective development construct) are equivalent. The groups understand the measured constructs equivalently.
RQ3: Threshold invariance	Do particular items function differently for individuals with the same level of development?	**Finding:** The item thresholds are equivalent across groups. **Interpretation:** Students *with the same level of development* have the same probability of providing particular responses to GPI items. The items do not appear to be systematically biased as a function of ethnoracial group membership.
RQ4: Factor disturbance invariance	Is the amount of unexplained variance in the first-order factors equivalent?	**Finding:** The first-order factor disturbances are equivalent across groups. **Interpretation:** The amount of unexplained variance in the first-order factors is equivalent across groups.
RQ5: Error (or unique) invariance	Is the amount of unexplained variance in the GPI items equivalent?	**Finding:** The error (unique variance) in the GPI items is equivalent across groups. **Interpretation:** The amount of unexplained variance in the GPI items is equivalent across groups.
RQ6: Second-order factor mean	Does the global perspective development mean score differ across groups?	**Finding:** Students' average estimated level of global perspective development does not differ across the four groups.

Research Q7: Does Evidence Exist for the Hierarchical Factor Structure of the GPI and Convergent and Discriminant Validity of the GPI's Scales?

In addition to investigating measurement invariance of the GPI's global perspective development items across four ethnoracial groups, part of this study's purpose was framed in terms of informing GPI refinement efforts. Validation efforts afford opportunities to closely examine an instrument's psychometric properties, illuminating very specific areas for instrument refinement. To answer this last research question, I present findings related to the GPI's factor structure and the convergent and discriminant validity of the GPI's six scales. Together, convergent and discriminant validity comprise important aspects of construct validity that allow instrument developers to understand the extent that an instrument has been operationalized in ways that accurately reflect its underlying theories. While I present specific findings related to these areas here, I discuss specific implications and suggestions more fully in the next chapter.

The GPI's Factor Structure

The present study's findings illuminate important information about the GPI's hypothesized factor structure. Historically, students' global perspective development—as measured by the GPI—has been understood to include six dimensions that span cognitive, intrapersonal, and interpersonal domains of development (Braskamp et al., 2014). Given this understanding, the GPI is often used to report students' levels of development across the six global perspective development scales. However, investigating the fit of a higher-order model is warranted when there is theoretical justification for a higher-order construct that predicts or explains the lower-order dimensions. In extending this understanding to the GPI, students' level of global perspective development (higher-order construct) would explain their level of development in each of the six developmental (lower-order) dimensions. This theoretically coheres given the conceptually interrelated dimensions of development encapsulated by the

global perspective development construct. Theoretically, development in one area often spurs concomitant development in the other areas. In addition, there must be evidence of strong associations between the lower-order dimensions themselves. The present study provides additional evidence that global perspective development can be understood as predicting or explaining five of the six first-order developmental factors of the GPI, which themselves predict or explain the observed measures captured by the GPI's items. However, the following findings suggest concrete instrument refinement opportunities.

First, a determinant of reliability in SEM is the average variance extracted (AVE), or the average percentage of variation in the first-order factors accounted for by the higher-order factor (and—as will be discussed shortly— the average percentage of variation in the items accounted for by their latent constructs). In SEM, the AVE explains the degree to which (1) the first-order factors and (2) a scale's items converge around the same latent construct (Hair et al., 2014). The AVE for the GPI's higher-order factor is calculated by taking the sum of the square of the second-order coefficients and dividing that by the number of first-order factors; the AVE should be $\geq$ 50% (Hair et al., 2014). In the aggregate sample, the AVE for the GPI's second-order global perspective development factor is 48.3% (see Table 9). As Table 9 also shows, the AVE for the GPI's second-order factor ranges from 45.3% (white group) to 57.7% (black/African American group). These AVE values can all be attributed to the GPI's Cognitive Knowing first-order factor, for which nearly all the variance is unexplained by the GPI's higher-order factor across all four groups (see Table 9).

Relatedly, hierarchical CFA requires an evaluation of the strength and statistical significance of the relationship between the second-order and first-order latent constructs. My

findings suggest there is essentially no relationship between the higher-order global perspective construct and the first-order Cognitive Knowing factor (γ = -.07; see Table 9). While there is strong support for the GPI's hierarchical factor structure given the strong associations between the higher-order global perspective construct and *five* of the first-order factors (standardized second-order coefficients (γ) range from .62 to .82; see Table 9), this is not the case for this first-order factor. Relatedly, the SMC coefficient for the Cognitive Knowing factor (i.e., proportion of variance explained by the second-order factor) indicates that for the aggregate sample, the higher-order factor explains no variance (R^2 = .01) in this first-order factor, leaving all unexplained variance. Further, in examining the correlation matrix of the six first-order factors for the aggregate sample (see Table 14), the Cognitive Knowing factor is only weakly, yet negatively, associated with all five of the other first-order factors (the correlations range from -.05 to -.07). Such a weak, inverse relationship with both the second-order and other first-order factors is theoretically unexpected. For the sake of comparison, the other five first-order factors' inter-factor correlations range from .49 to .66 (see Table 14), which is more theoretically expected.

Table 14. Inter-factor Correlations of the GPI's First-order Factors Using Aggregate Sample

	Cognitive Knowing	Cognitive Knowledge	Identity	Affect	Social Responsibility	Social Interaction
Cognitive Knowing	1.000					
Cognitive Knowledge	-0.057	1.000				
Identity	-0.045	0.499	1.000			
Affect	-0.056	0.629	0.494	1.000		
Social Responsibility	-0.059	0.656	0.516	0.650	1.000	
Social Interaction	-0.054	0.599	0.471	0.594	0.619	1.000

In addition to these unexpected findings, the initial model specification highlighted other issues with particular items currently on the Cognitive Knowing scale. Only four items from this scale were retained in the final measurement model used for the study's CFAs and MGCFAs. However, seven items are included on this scale, and all are currently used for reporting purposes. Two items—*I consider different cultural perspectives when evaluating global problems* and *I take into account different perspectives before drawing conclusions about the world around me*—were multidimensional (neither actually initially loaded onto this scale; they loaded onto two completely different scales), and another item—*Some people have culture and others do not*—was removed because its factor loading was quite low. Finally, one item—*I rely primarily on authorities to determine what is true in the world*—though statistically significant in the overall and three of the group-specific models, was statistically non-significant ($p > .05$) for the Hispanic group. Given the theoretical import of epistemological development relative to the concept of global perspective development, these findings suggests a need to revisit the Cognitive Knowing scale's conceptualization (as discussed in Chapter Two, this is currently

theoretically quite dense) and operationalize the scale's items differently. I discuss these particular suggestions in much more detail in the next chapter, providing specific recommendations.

Reliability and Convergent and Discriminant Validity of the GPI

In evaluating CFA models, determining the reliability as well as convergent and discriminant validity within a measurement model is critical. Each of these aspects is discussed in relation to the GPI's six scales.

Reliability of the GPI scales. First, in terms of the reliability of the GPI's scales—or how well the items on their respective scales all measure the same thing—the composite reliability (CR) for each scale was presented in Table 1. The conventional threshold for a determinant of acceptable internal consistency of a latent construct is .70 (Hair et al., 2014). For the aggregate sample, the Cognitive Knowing CR is .59, below the threshold. The relatively lower CR value for this scale suggests that the Cognitive Knowing scale's items do not measure the same thing. For the aggregate sample, the CR values for the other five GPI scales were all > .70; these ranged from .77 to .81 (see Table 1), suggesting acceptable internal consistency. These other five scales' items appear to all measure the same thing. Implications of these important differences are discussed in the next chapter.

As mentioned earlier, a second determinant of a scale's reliability is the AVE (i.e., in this case, the average percentage of variation in the *items* accounted for by their first-order latent constructs). The AVE is important because it explains the degree to which a scale's items converge around the same latent construct; as such, it is also used as a determinant of convergent validity discussed next (Hair et al., 2014). In the case of determining the AVE for the first-order

factors, this is calculated by taking the sum of the square of the factor loadings and dividing that by the number of indicators on the scale; the AVE should be ≥ 50% for each construct (Hair et al., 2014). Considering all six GPI scales (see Table 8), the AVE was only ≥ 50% for the Social Interaction scale for the aggregate (AVE = 54.2%), ANHOPI (AVE = 51.0%), and white (AVE = 53.7%) groups; the AVE for this scale was < 50% for the black/African American (AVE = 48.3%) and Hispanic (AVE = 47.0%) groups. In addition the AVE for the Affect scale was exactly 50.0% for the black/African American group but was < 50% for all other groups. The AVE values for the other five scales using the aggregate sample were as follows, in order from highest to lowest: Cognitive Knowledge AVE = 46.6%; Social Responsibility AVE = 46.1%; Affect AVE = 45.4%; Identity AVE = 44.9%; and Cognitive Knowing AVE = 27.6%. When the AVE is < 50% for any given construct, there is more error (unique) variance than explained variance. High unique variance like this suggests that items are not reliable, largely explained by other latent factors, or otherwise very different from other indicators on the scale. This typically suggests items should be investigated for item wording or other contributing issues such as multicolinearity (this latter possibility was examined in seeking evidence for discriminant validity, which is discussed shortly).

Convergent validity of the GPI's scales. Convergent validity aids in understanding the dimensionality of a scale since this represents the extent to which items on a scale converge, or measure the same underlying construct; in other words, evidence of convergent validity suggests that a scale's items effectively measure the underlying construct (Hair et al., 2014). In a CFA framework, convergent validity is evaluated using a scale's factor loadings (should all be statistically significant and > .50), composite reliability (CR should be > .70), and the AVE

(should be > .50; Hair et al., 2014). I only found evidence of convergent validity for the modified Social Interaction scale used for this study; this scale's factor loadings were all statistically significant and > .65, CR = .71, and AVE = 54.2%. For the Cognitive Knowing scale, the CR (.59) and AVE (27.6%) values for this scale were both below the thresholds. In addition, though statistically significant, the relationships between the Cognitive Knowing factor and the observed indicators are relatively low. In the baseline model, this scale's standardized first-order factor loadings (λ) ranged from .41 to .62 (see Table 8). Factor loadings explain the strength of the relationship between an observed indicator and its latent construct, so the expected strength of any given factor loading in a CFA model relates to the theoretical association between the indicator and that construct. Low factor loadings, by themselves, do not necessarily represent validity threats. For instance, if the latent construct is best measured by a collection of different types of items, a lower factor loading might be expected. However, low factor loadings often compromise a scale's convergent validity and often suggest opportunities to revisit the operationalization of the latent construct under study.

While all factor loadings for the remaining four scales were statistically significant and > .50, and all CR values were > .70, the AVE values for these four scales were all below 50% as discussed above (see Table 8). As discussed earlier, relatively low AVE for these scales suggests opportunities to refine items to make them more effective measures of their respective latent constructs. I discuss in detail in the next chapter the implications of these lower AVE values in five of the GPI's scales, closely examining particular items' factor loadings on these scales and specifying what this means in terms of item refinement.

Discriminant validity of the GPI's scales. Finally, discriminant validity of an instrument's scale explains the extent to which that latent construct measures something unique that the other latent constructs in the model do not measure (Henseler et al., 2015). Evidence for this type of validity related to first-order factors requires that they not correlate excessively with one another. In SEM, discriminant validity is important because if there is no evidence supporting this, the latent constructs exert an influence on the variation of the observed indicators beyond the ones to which they are theoretically related (Henseler et al., 2015). To determine discriminant validity, one can compare the square root of a construct's AVE to the inter-factor correlations among the first-order factors; if the former is greater in value than the latter, this provides evidence of discriminant validity. In terms of the discriminant validity of all six GPI scales, Table 15 displays the square roots of the AVEs for these scales and compares these to the inter-factor correlations. For all six scales, the square roots of the AVEs are greater than their respective factor inter-correlations (see Table 15), which provides evidence for the discriminant validity of all six GPI scales. In other words, while there is some degree of relation between the scales—as theoretically expected—the six GPI scales appear to measure distinct aspects of students' development.

Table 15. Comparing Inter-factor Correlations and the Square Root of the First-order Factors' AVE to Test for Discriminant Validity of the GPI's Six Scales

	Cognitive Knowing	Cognitive Knowledge	Identity	Affect	Social Responsibility	Social Interaction
Cognitive Knowing	**0.525**[a]					
Cognitive Knowledge	-0.057	**0.683** [a]				
Identity	-0.045	0.499	**0.670** [a]			
Affect	-0.056	0.629	0.494	**0.674** [a]		
Social Responsibility	-0.059	0.656	0.516	0.650	**0.679** [a]	
Social Interaction	-0.054	0.599	0.471	0.594	0.619	**0.736** [a]

Note. [a]The bolded values along the diagonal represent the square root of the constructs' AVE to allow for comparison between these values and the inter-factor correlations. To determine discriminant validity, one can compare the square root of a construct's AVE to the inter-factor correlations among the first-order factors; if the former is greater in value than the latter, this provides evidence of discriminant validity.

Reverse-worded Items

Finally, the GPI currently contains seven reverse-worded items, five of which appear on the GPI's Cognitive Knowing scale. Only four of these reverse-worded items were retained in the final CFA and MGCFA models. I removed three of these reverse-worded items (one item each from the Cognitive Knowing, Social Responsibility, and Social Interaction scales) because of very low factor loadings when examining Baseline Model 2 (these items' standardized λ ranged from .10 to .29; see Table 6). Relatedly, these reverse-worded items' unique variances were all quite high in examining the estimates in Baseline Model 2 (across these seven items, θ_ε ranged from .67 to .99, with three items' $\theta_\varepsilon > .92$; see Table 6 for R^2 values, knowing that $1 - R^2 = \theta_\varepsilon$).

As mentioned, unique variance is caused by extraneous factors that influence responses as well as random measurement error or unreliability in an item (Brown, 2015). Given that the unique variances for these reverse-worded items, in particular, were so high compared to almost all other GPI items, I examined whether the items' unique variance could be considered non-random (or correlated) given that they are all reverse-worded. It is sometimes warranted to specify non-random error in SEM/CFA models, but this decision should always be theoretically justified. For instance, the items' unexplained variance can be influenced by the type of administration, scoring, or item wording (i.e., reverse- and positively-worded items; Brown, 2015). Additionally, if non-random error is not modeled appropriately, the model can suggest the existence of latent constructs that are not conceptually substantive models, but instead reflect a method effect (Brown, 2015). I wanted to determine whether to model any non-random error due to a method effect given these items' shared reverse-wording. To begin, I examined the modification indices and expected parameter changes for Baseline Model 2 to determine whether correlating these reverse-worded items' error variances would improve model fit. None of the highly significant modification indices appeared theoretically appropriate, and none suggested improved model fit after allowing the reverse-worded items' unique variances to covary. These findings suggested that the unique variance for these items appears to be random (uncorrelated), or at the very least not a method effect related to their reverse-worded nature. However, these items should still be refined given their low factor loadings and high unique variances as well as the specific complications that can arise from the use of reverse-worded items (van Sonderen et al., 2013). I discuss in the next chapter additional implications and suggestions related to these reverse-worded items.

CHAPTER FIVE

DISCUSSION

As internationalization efforts continue to expand within U.S. postsecondary education, the process for determining the effectiveness of such efforts has shifted. Programs increasingly need to report not just the global learning opportunities they offer, but what their students are learning from these curricular and co-curricular efforts (Green, 2012) and how such learning aligns with institutional objectives (IAU, 2014). This increased focus on accountability is resultant from complex, interactive forces at multiple levels, including global competition, divestment in and marketization of higher education, accreditation standards, and diverse stakeholder interest. Nevertheless, this shift has centered the assessment and research of students' intercultural competencies within the global learning and international education arenas of U.S. higher education (Deardorff, 2015).

As the broader U.S. landscape becomes both increasingly diverse and divided, postsecondary educators must consider how they can prepare students, the extent to which they have done this, and the factors that both contribute toward and hinder such development. The GPI is an instrument that allows educators to measure aspects of students' learning and scholars to examine students' intercultural competencies. The present study aimed to inform these efforts by using a national sample of undergraduates to examine the validity of the GPI's items across four ethnoracial groups. The purpose of this was three-fold. First, I sought to provide evidence of whether the GPI's developmental constructs were theorized, understood, and measured

equivalently across an ethnoracially diverse group of undergraduates. In doing so, I added to the large-scale postsecondary educational survey validation and cross-cultural validation literature to address the paucity of rigorous cross-cultural validation work in global learning-related instrumentation.

Second, by examining the cross-cultural validity of the GPI, I sought to provide important evidence for practitioners and researchers that informs their ability to compare students' global perspective development across four ethnoracial groups. In proceeding with this particular purpose, I acknowledged in Chapter One the aforementioned complexities of examining ethnoracial differences related to a variety of educational experiences and outcomes (i.e., Irizarry, 2015; Rockquemore et al., 2009). However, as with other large-scale educational surveys, scholars (e.g., Engberg et al., 2016; Engberg & Davidson, 2016) have used the GPI to investigate cross-group differences related to students' outcomes (i.e., usually in terms of the GPI's six developmental scales) and various forms of engagement to understand nuances related to undergraduates' development. I sought to provide important validity evidence for these uses. Finally, examining and improving an instrument's quality should be a primary aim of survey methodologists and a major consideration for consumers of data from survey-based research and assessment efforts (Cizek et al., 2010). Examining cross-cultural, convergent, and discriminant validity provides important evidence related to an instrument's psychometric properties. I sought to understand whether such evidence suggested specific opportunities for refinement of the GPI. If so, I aimed to synthesize those findings in ways that could inform any subsequent refinement efforts.

Currently, there are growing, multifaceted criticisms of outcomes-based assessment in general (i.e., that this is perceived as reductionist in nature, overly reliant on quantitative data, too resource-intensive, not able to yield meaningful or actionable evidence) and particularly related to survey-based assessment methods (i.e., that survey constructs are immeasurable and methodologies are flawed and non-rigorous; e.g., Eubanks, 2017, Gilbert, 2018, Worthen, 2018). Further, others have scrutinized the equity and inclusivity of outcomes-based assessment, reiterating the need for culturally-responsive assessment of students' learning given the diverse learning contexts within U.S. higher education (e.g., Montenegro & Jankowski, 2017). I took a particular interest in responding to these important concerns with the present study, both in terms of what initially drove the study and the implications of the findings. In framing this study, I acknowledged the import of understanding the assessment of intercultural competencies as inherently value-laden. I recognized the tremendous power embedded in identifying the types of learning and development that are prioritized related to global learning, deciding what is measured, and—particularly germane to the present study—*how* global learning is measured.

This study also responded to an existing gap within postsecondary educational survey-based efforts. The increased use of large-scale surveys in postsecondary institutional assessment and research efforts necessitates more commitment to ensuring the validity of these instruments for both scholars and practitioners (Porter, 2011). However, even when instrument validity is explicitly examined, its understanding and evidence vary immensely across studies (Cizek et al., 2010; Kane, 2001; Sireci, 2007), and the culturally-bound process of validation work is often ignored (Kirkhart, 1995). Given the validation purposes of this study, I sought to center both of these concerns in framing, executing, and discussing this study. I aimed to precisely understand

the process (i.e., using the argument-based approach to validation) and considerations related to such a process (i.e., validity related to different uses, considering fairness in testing). In addition, I framed this study in terms of examining the extent to which educational measurement considers the culturally diverse theoretical, methodological, and consequential aspects encapsulated within Kirkhart's multicultural validity concept. This type of precision in validation work will continue to be necessary if U.S. postsecondary education continues to use surveys in its assessment and research efforts, namely as such inquiry unfolds amidst increasing internationalization efforts within more diverse contexts.

Discussion of Measurement Invariance Findings

To test the cross-cultural, convergent, and discriminant validity of the GPI, I examined measurement invariance of the GPI across four ethnoracial groups and evaluated the instrument's psychometric properties. Examining measurement invariance of the GPI allowed me to understand whether the theorization, understanding, and measurement of the GPI's developmental dimensions differed across four ethnoracial groups. Determining this involved sequentially testing several models to determine whether particular aspects of those models (e.g., number of factors, magnitude of factor loadings) were equivalent across the groups. This study's first six research questions were each structured to test the cross-group equivalency of particular GPI model parameters and each answer provided different evidence. I discuss these findings below.

Configural and Metric Invariance

To begin, my first two research questions tested (RQ1) configural invariance (i.e., whether the same constructs are measured across groups) and (RQ2) metric invariance (i.e.,

whether the constructs have the same meaning) across four ethnoracial groups. First, my findings suggest that the GPI measures the same theoretical constructs across the four ethnoracial groups. This particular finding suggests the theorization of the GPI's developmental constructs is cross-culturally appropriate (He & van de Vijver, 2012). Second, my findings suggest that—in addition to the four groups understanding the same developmental constructs the GPI measures—relationships between the GPI's developmental constructs and its items are also equivalent across groups. This implies that the psychological meanings of the underlying developmental constructs are equivalent (Vandenberg & Lance, 2000).

These first two findings relate to Deardorff's (2004) position that the definitions or concepts of intercultural competencies do not, themselves, often appear to be culturally variable. It is particularly interesting to place these initial two findings in conversation with earlier work that has suggested cultural variability related to the *processes* involved in the types of development measured by the GPI. These initial findings may help underscore a distinction between the GPI's developmental constructs themselves (e.g., how one understands the idea of learning about other cultures or interacting across particular differences) and the psychological or behavioral components involved in that type of development (e.g., how easy or difficult it is to actually interact across particular differences). For instance, earlier qualitative work has suggested cultural variability related to the developmental *processes* that comprise the self-authorship concept (i.e., how context and relationships influence catalysts for development; Pizzolato et al., 2012), which undergirds three of the GPI's developmental dimensions. But the present study's findings suggest that students understand the actual attitudes and behaviors that are indicators of the GPI's dimensions of development equivalently. This distinction between

developmental construct and process seems critical in the charge to inclusively theorize students' learning and development.

Threshold Invariance

My third research question (RQ3) examined threshold invariance to test whether the GPI items function differently across the four groups for individuals with the same level of development. My findings confirm that the GPI items' thresholds are invariant across the groups. This indicates that for students with the same levels of development on the GPI's scales, the probability of providing the same survey response is equivalent across the four groups. A survey item may be systematically biased if this is not the case. Bias is observed when item score differences do not correspond with differences related to the underlying construct being measured; differences may instead relate to the non-applicability of items in various cultures, the triggering of additional attributes not measured by that construct, or ambiguous connotations from item phrasing. As such, inter-group comparisons on these biased items yield invalid cross-group comparisons (He & van de Vijver, 2012). To illustrate this, if students from different ethnoracial groups have a different probability of providing particular survey responses, this does not by itself indicate a survey item functions differently across these groups (i.e., it would just mean the groups score differently on that item). However, if students from different ethnoracial groups with the same level of the attribute being measured (i.e., particular levels of development in the GPI's case) have a different probability of providing particular survey responses, this does indicate the survey item functions differently across those groups.

This particular finding of invariant threshold invariance is important for a two reasons. First, part of this study's purpose involved providing validity evidence related to the appropriate

comparison of GPI mean scores and regression coefficients across ethnoracial groups. In order to conduct *t* tests or analyses of variance (ANOVAs), or to compare various groups' regression weights in models, threshold invariance *must* be demonstrated. Validity evidence now exists for these uses. Second—and relatedly—the *Standards* (AERA et al., 2014) cite measurement bias as an important threat to ensuring fairness in testing. I did not find evidence of measurement bias in the GPI's items. Further, there are consequential aspects to measuring students' learning; if the assessment of such learning is not inclusive of diverse learners, this cannot contribute toward all students' success (Montenegro & Jankowski, 2017). The evidence related to these first three research questions (considered together as *strong* measurement invariance; Dimitrov, 2010) suggests the inclusive theorization of the GPI's constructs, that these four groups understand the psychological meaning of the constructs equivalently, and that their levels of development correspond to equivalent survey responses across groups.

Factor Disturbance and Item Error Invariance

My next two research questions examined *strict* measurement invariance (Dimitrov, 2010) by testing whether factor disturbances (RQ4) and item error (or unique variances) (RQ5) were equivalent. Findings related to RQ4 suggest that the unexplained variance (i.e., factor disturbance) of each first-order factor that is not shared by the higher-order factor is equivalent across the groups (Chen et al., 2005). The hypothesized hierarchical factor structure of the GPI indicates that the higher-order global perspective development construct should influence or explain each of the first-order developmental constructs. The second-order coefficients indicate the strength of those relationships, while the SMCs (in this case, the squared second-order coefficients) indicate the amount of *explained* variance in the first-order constructs by the higher-

order construct. Findings related to RQ4 suggest that any *unexplained* variance in the first-order constructs is equivalent. Findings related to RQ5 suggest that the GPI items are measured with the same level of precision across groups. In other words, the amount of unexplained variance (i.e., variance not accounted for by the first-order constructs) within the GPI items is the same across the groups. The findings related to RQ5 are particularly important because error (or unique variance) affects the strength of the correlations among the GPI's items; this type of invariance is required in order to compare inter-item correlations between groups (Tucker, Ozer, Lyubomirsky, & Boehm, 2006). I will discuss unique variance later in this chapter given its relation to the convergent validity of the GPI's scales. Importantly for both RQ4 and RQ5, these types of invariance indicate that the quantity of unexplained variance is the same across groups; these findings do not reveal the *nature* of this unexplained variance and whether this is equivalent.

Second-order Factor Mean Invariance

Once I concluded measurement invariance across the four groups, my sixth research question (RQ6) compared the second-order global perspective development factor mean score across the four groups. I found no difference in the average level of global perspective development across the four groups. This particular finding is novel, as this study represents the first time the GPI's second-order factor mean has been compared. In a hierarchical factor structure, the first-order factor means are conditional on the second-order factor mean (Chen et al., 2005). The GPI's hierarchical factor structure has interesting implications for reporting students' levels of development in terms of an overall average level of development (i.e., using the second-order factor mean) and/or in terms of the average levels of development on the GPI's

six subscales. However, given some of the implications I discuss later in this chapter related to the GPI's factor structure and convergent validity of its six scales, further consideration of this is warranted.

Additional Measurement Invariance Considerations

Given that the present study's findings suggest measurement invariance across the four ethnoracial groups for the GPI's global perspective development items, it is useful to discuss reasons that help explain this. The following discussion is structured with two major considerations relative to this study's cross-cultural validity findings.

Domestic and international students. The sample of undergraduate students for this study was comprised of 91% domestic (n = 6,465) and 9% international students (n = 403). Since the sample overwhelmingly consisted of domestic students (i.e., students who identified as an American student attending an American college/university, as the GPI asks), this may partially account for the measurement invariance observed. In particular, the cultural variability discussed in Chapter Two related to the GPI's developmental dimensions underscored the role of broader cultural considerations and norms. These included divergent socialization around the involvement of authorities in one's life (e.g., Broido & Schreiber, 2016), individuals' baseline intercultural sensitivity (i.e., how individuals are socialized in an incredibly culturally diverse society) and communication style (i.e., preferences for indirect or non-confrontational interactions; Tamam, 2010) and cultural value orientations that emphasize collectivism as opposed to individualism (e.g., Dugan et al., 2008). Given that the student sample was overwhelmingly domestic, it could be expected—though not certain—that this majority was socialized in the U.S. around the types of cultural norms discussed above. As such, it could be

expected that not as much variability in terms of understanding or replying to the GPI's constructs would be present among this sample. Further, earlier unpublished findings using a different sample of undergraduates suggested configural, metric, and item error invariance of the GPI's items across groups of domestic and international students (Davidson, 2015). While threshold invariance was not tested in that earlier exploratory work, those preliminary findings at least suggested that these two student groups understood the GPI's constructs equivalently.

Role of students' educational contexts. I provided a review of the literature in Chapter Two on the culturally variable nature of the GPI's developmental dimensions. Much of this literature emphasized the role of students' educational contexts relative to these types of development. My findings suggest that for the four ethnoracial groups under study, their understanding of and measurement related to these areas of development are equivalent. However, we cannot know from the items currently included on the GPI about whether these students' *experiences* related to these areas of development vary. Though the GPI General form includes both Sense of Belonging and Intercultural Engagement scales, the ways in which both scales are operationalized on the instrument—as discussed in detail within the study's limitations section in Chapter Three—limit their use in understanding the current sample's perceptions of their campus climates. For instance, without items that ask students about important contextual elements related to their development, the GPI's six developmental constructs are isolated from key considerations including their perceptions of the ease, effort, or desirability relative to interacting across difference (i.e., the psychological dimension of the campus climate; Milem et al., 2005). We cannot know students' interest in learning about or interacting with other cultures or the perceived quality or potential costs and benefits of those interactions (i.e., the behavioral

dimension of the campus climate; Milem et al., 2005). And we cannot know students' perceptions related to how much their institutions emphasized these six developmental areas and offered opportunities to engage in these ways (i.e., the organizational dimension of the campus climate; Milem et al., 2005). Given the literature on the variable nature of students' experiences and engagement that relate to the GPI's dimensions of development, I discuss item development related to this area as a future direction for inquiry later in the chapter.

Discussion on Additional Validity Evidence for the GPI

My final research question (RQ7) examined whether evidence exists for the hierarchical factor structure of the GPI and convergent and discriminant validity of the GPI's six scales. Examining such evidence informs instrument refinement opportunities, as item- and scale-specific issues are illuminated. The following section discusses this specific evidence.

Hierarchical Factor Structure

Regarding the GPI's hierarchical factor structure, this study's findings provide evidence that global perspective development can be understood as a higher-order factor predicting or explaining five of the six first-order developmental factors of the GPI (i.e., all except the Cognitive Knowing factor), which themselves predict or explain the observed measures captured by the GPI's items. Considering five of the first-order factors, the strength and statistical significance of the relationships between the second-order and these first-order latent constructs reflect strong, statistically significant relationships. However, findings illuminated specific issues with the GPI's Cognitive Knowing scale, which compromised the overall reliability of the higher-order global perspective development construct.

First, findings suggested that the Cognitive Knowing scale was not related to the higher-order global perspective development construct or to any of the other five first-order developmental factors. This was theoretically unexpected, as epistemological development theoretically influences *and* is influenced by the other domains of development measured by the GPI (Kegan, 1994; King & Baxter Magolda, 2005). Second, three items from the original Cognitive Knowing scale could not be used in the present study due to multidimensionality (i.e., two items related to the Cognitive Knowledge and Intrapersonal Affect scales instead) or extremely low factor loading. Third, for the four items from this scale that were retained in the present study, findings suggested low internal consistency; in their present state, these items do not appear strongly related to one another. These observations raise questions around the operationalization of the GPI's Cognitive Knowing construct and its relation to the higher-order global perspective development construct.

The Cognitive Knowing scale is theorized to measure the degree of complexity of one's views related to understanding the contextual nature of reality and knowledge (RISE, 2017b). Extending from its theoretical underpinnings, this scale should emphasize increasingly complex habits of mind and perceptual processes during intercultural exchange, including the acceptance of uncertainty, evaluative skills, and the realization that reality and knowledge are constructed (Kegan, 1994; King & Baxter Magolda, 2005). These cognitive developmental dimensions that Kegan (1994) and King and Baxter Magolda (2005) explain in their respective frameworks are critical to the development of intercultural competencies and have been reinforced in other developmental models (e.g., Bennett's (1993) developmental model of intercultural sensitivity; Deardorff's (2004) model of intercultural competence). In particular, all of these theoretical

approaches to understanding effective intercultural exchange underscore the development of increasingly complex ways of *considering* cultural difference. Aspects of this (i.e., respect, openness, curiosity, and discovery) are requisites for intercultural knowledge, which itself involves critical thinking skills (i.e., analysis, evaluation, and interpretation of information; Deardorff, 2004), reinforcing the notion that *how* one knows changes *what* one knows and that these two continue to inform one another.

A close examination of this scale's seven items currently suggests these relate to (1) ethnocentrism (e.g., *When I notice cultural differences, my culture tends to have the better approach*, *Some people have culture and others do not*), (2) sources and justification of knowledge (e.g., *In different setting what is right and wrong is simple to determine*, *I rely primarily on authorities to determine what is true in the world*, *I rarely question what I have been taught about the world around me*), and (3) pluralistic orientation (e.g., *I take into account different perspectives before drawing conclusions about the world around me*, *I consider different cultural perspectives when evaluating global problems*). I present implications of this shortly as it relates to opportunities for refining this scale to more closely reflect its theoretical content.

Convergent and Discriminant Validity

I examined the convergent and discriminant validity of the GPI's scales since such evidence allows instrument developers to understand the extent that an instrument has been operationalized in ways that accurately reflect its underlying theories. Specifically, such evidence explains the extent that items on a given scale measure the same underlying construct

(i.e., convergent validity) and the extent to which the underlying constructs measure only what they are theorized to measure and not other areas of development (i.e., discriminant validity).

While I found evidence for the discriminant validity of the Cognitive Knowing scale, I did not find evidence of convergent validity due to its low internal consistency (CR = .59), low average variance extracted (AVE = 27.6%), and one of the factor loadings' low standardized value. As mentioned in the last chapter, when the AVE is < 50%, there is more error (unique variance) than explained variance, so items should be investigated for item wording or other contributing issues. These issues with the Cognitive Knowing scale can likely be attributed to two causes. First, five of the scale's original seven items are reverse-worded, which warrants attention; since reverse-worded items are currently embedded on three GPI scales, specifics related to those particular items are discussed later in this chapter. The second cause—as discussed earlier—may be due to how this scale has been conceptualized and subsequently operationalized into its present items.

In Chapter Four, I presented evidence of the convergent and discriminant validity of the GPI's other five developmental scales. I found evidence of the discriminant validity of each of the GPI's scales. This suggests that these scales are not also measuring other aspects of students' development. However, findings suggested that only the modified version of the Social Interaction scale used for this study showed convergent validity. I did not find evidence of convergent validity for the remaining GPI scales, which suggests opportunities to refine particular items to make them more effective measures of their respective latent constructs.

In a CFA framework, convergent validity is evaluated using a scale's standardized factor loadings (should all be statistically significant and > .50), composite reliability (CR should be >

.70), and the AVE (should be > .50; Hair et al., 2014). The standardized factor loading was < .50 for only one of the 26 items used in the present study (one item on the Cognitive Knowing scale), and all factor loadings were statistically significant ($p < .01$). For the five remaining GPI scales (the Cognitive Knowing scale was discussed separately above), their composite reliabilities were all > .70 (these ranged from .71 to .81). However, the AVE for all scales except the Social Interaction scale was less than 50%. Given these lower AVE values, it is useful to reexamine the items' standardized factor loadings on these scales.

Though Hair et al. (2014) explain that for evidence of convergent validity, the minimum threshold for factor loadings is > .50, in order for a scale's AVE to be > 50%, most of the factor loadings would actually need to be higher than .50. For instance, when an item's factor loading is .71, its SMC (or the amount of variance explained in that item by its latent construct) is exactly 50%. When one sums up all the SMCs on a given scale and divides that total by the number of items, one obtains that scale's AVE. When the AVE is < 50% for a given scale, it is understood that, on average, the item factor loadings on that scale are < .71 (Hair et al., 2014). Given this, instrument refinement efforts *must* focus on items with factor loadings < .50; however, refinement efforts can also focus on items with factor loadings < .71 in order to increase AVE (i.e., making these particular items more effective measures of their respective constructs). Using the baseline measurement model—and excluding the Cognitive Knowing scale, which I discuss separately below—there were 11 items whose standardized factor loadings were > .50 but < .71 (three on the Identity, three on the Cognitive Knowledge, two on the Affect, two on the Social Responsibility, and one on the Social Interaction scales). Shortly, I discuss the

implications of these findings, using this evidence to present specific item refinement suggestions associated with each.

Implications

Considering the validity evidence presented, there are several practical implications of this study's findings. First, this study's methodologies stand to inform subsequent validation efforts of large-scale educational surveys in several specific ways. Second, the study's evidence suggests particular opportunities to refine some of the GPI's items to make them more effective measures of the constructs to which they theoretically relate. Third, evidence also illuminates particular issues with the GPI's reverse-worded items. I discuss implications related to all of these areas in detail here.

Encouraging More Cross-cultural Validation of Surveys

There is a paucity of rigorous validity studies of intercultural competency measures (Schnabel et al., 2015) that can be subsumed under the larger need for more validation work with postsecondary educational surveys more generally (Porter, 2011). This reality might be understood through three perspectives that this study can help transform. First, the definition of validity—and determinants of this—are often not precisely understood among survey researchers (Kane, 2013). Newton and Shaw (2013) remind researchers that given the various types of validity—and that such evidence provides different types of answers—one must approach this work with clear purposes (i.e., examining *particular* interpretations or uses of an instrument) and rigorous methods by which evidence is collected. The present study provides a roadmap for instrument developers and those using large-scale surveys for how to understand various types of validity (i.e., cross-cultural, convergent, discriminant), what drives the examination of those

types of validity with particular attention to theoretical considerations, and how to use such validity evidence in the specific ways outlined.

Second, even if validity and different forms of validity evidence are adequately understood, the process of validation work can still be daunting to those without extensive methodological training. It is my hope that the present study contributes methodologically to both postsecondary educational and intercultural competency survey research so that more of this work is done. Just as I aimed to provide a roadmap for how to *understand* validity—as discussed directly above—I also hope that this study's methods and findings can teach others how to do this important work by outlining important methodological considerations and analytical strategies.

Finally, in addition to the methodological considerations discussed above, I hope that this study encourages more scholars to examine epistemological and theoretical underpinnings of how we in postsecondary education understand and quantitatively measure students' learning and development. Instrument development *and* validation are theoretically driven. The cross-cultural validity purpose of this study was informed by a rich body of literature on the culturally variable nature of college students' learning and development. To support the important charge to approach the assessment of students' learning and development in culturally-relevant ways (Montenegro & Jankowski, 2017), postsecondary education must continue to examine the extent to which the constructs and phenomena studied through surveys are theorized and measured in inclusive ways. While the present study adds directly to this work, I hope the present study also drives others to deeply examine what they measure and how they do so.

Recommendations for Instrument Refinement

Instrument refinement has both methodological and practical implications. Methodologically, instruments should be reliable and valid measures of the phenomena they seek to measure. If an instrument aims to be used in investigating particular dimensions of development or in assessing students' learning and development, determining whether it validly does so is essential. From a practical standpoint, as institutions use the GPI as a tool to understand students' development—and hopefully use those findings to make informed decisions about opportunities that promote such development—precision around the GPI's developmental constructs is paramount. In this section, I begin by discussing refinement opportunities related to the GPI's Cognitive Knowing scale, as my findings suggest issues with the reliability and validity of this scale as well as issues related to the association with this scale and the higher-order global perspective construct to which it theoretically relates. I then discuss additional validity evidence for the five other GPI scales, using that evidence to provide concrete refinement recommendations.

Cognitive Knowing scale. Refining the GPI's Cognitive Knowing items to reflect this scale's original conceptualization will be beneficial. Doing so should involve developing items that operationalize the (1) acceptance of uncertainty and (2) evaluative skills dimensions of this construct, as these appear to be key drivers of epistemological development, namely as this relates to intercultural competencies. As discussed, both of these dimensions theoretically underlie the Cognitive Knowing construct (Kegan, 1994; King & Baxter Magolda, 2005). However, the construct is currently discussed as a unidimensional developmental domain of the

GPI. The unidimensionality of this construct may not be realistic, given the conceptual complexity of epistemological development.

In underscoring the focus on acceptance of uncertainty and evaluative skills, it is useful to understand the developmental tasks required for the increasingly complex habits of mind and perceptual processes related to intercultural exchange explained by Kegan (1994) and King and Baxter Magolda (2005). Kuhn, Cheney, and Weinstock (2000) explain that epistemological development over the lifespan entails coordinating objective and subjective aspects of knowing, moving through four levels of understanding (i.e., pre-absolutist, absolutist, multiplist, and evaluativist). Considering the developmental timeframe for many undergraduates, they may move from absolutist to multiplist orientations, which requires them to shift their view of the source of knowledge from the objective (i.e., that which is knowable) to the subjective (i.e., the knower). "This awareness comes to assume such proportions, however, that it overpowers and obliterates any objective standard that could serve as a basis for comparison or evaluation of conflicting claims," so all claims are viewed as equal (p. 310). Kuhn et al. explain that the shift from multiplist to evaluativist—an epistemological level that entails the type of complex understanding required for effective intercultural exchange—involves the eventual reintegration of the objective into one's knowing by integrating *uncertainty* with an *evaluative frame* to judge the merit of claims. For instance, if one operates with an evaluativist epistemological orientation, reality and knowledge are viewed as uncertain, assertions are judgments that must be evaluated and compared based on particular criteria of argument and evidence, and critical thinking is required for sound claims and to understand comprehensively (Kuhn et al., 2000).

In considering specific opportunities for item refinement for this scale, operationalizing acceptance of uncertainty and evaluative skills (i.e., more cognitively complex and critical thinking) into measureable items is feasible. Regarding acceptance of uncertainty—a critical aspect of how this scale is theorized—three existing items (*In different setting what is right and wrong is simple to determine*, *I rely primarily on authorities to determine what is true in the world*, *I rarely question what I have been taught about the world around me*) have been operationalized to reflect a dualistic level of epistemological understanding. Dualistic understanding entails viewing knowledge as certain, assertions as facts, and reality as directly knowable (Kuhn et al., 2000). These items' current reverse-worded nature measures the inverse of what the construct should measure (i.e., increasingly complex epistemological development). I will discuss shortly specific issues with reverse-worded items and that such issues often warrant not using them. In addition to not asking about the inverse of this construct, items could be developed to measure uncertainty avoidance or tolerance of ambiguity. These would more precisely capture this acceptance of uncertainty aspect of the development measured by this scale.

Second, developing items related to higher-order cognitive skills such as evaluative thinking also seems feasible. However, this needs to be carefully operationalized. Too many times concepts such as *critical thinking* are not effectively operationalized; students are instead simply asked to report the extent that they feel prepared for or actually engage in critical or analytical thinking (see Porter, 2011 for examples). Item comprehension is significantly compromised in these instances since respondents may not understand these concepts at all, or given the concepts' vagueness or ambiguity, respondents may interpret them unreliably, focusing

on any given aspect or application of such complex tasks. In operationalizing these items related to evaluative thinking, the key consideration involves what students should be able to do if they are evaluativists. In alignment with the recommendation of Heine, Lehman, Peng, and Greenholtz (2002), it could be more useful to create behaviorally-oriented items that ask students the extent to which they engage in behaviors that would reflect this dimension of the construct (e.g., how much do they compare different forms of evidence, center assumptions, determine the merit of particular claims). Interestingly, two current items on this scale that seem to involve pluralistic orientation, *I take into account different perspectives before drawing conclusions about the world around me*, *I consider different cultural perspectives when evaluating global problems*, seem to conceptually relate to this area. However—as noted earlier—these items instead loaded onto the Cognitive Knowledge and Intrapersonal Affect factors in this study and in earlier validation work by Davidson and Engberg (under review). Perhaps after refining the other items in the scale as recommended, removing one of these two items altogether (they appear redundant) and making slight refinements to the other, this scale would cohere differently.

While I have offered thoughts on conceptualizing and operationalizing this particular scale differently in an attempt to more precisely align with its theoretical grounding, I also acknowledge the difficulty of measuring students' epistemological development through self-report surveys (see DeBacker et al., 2008 for a review). Scholars have examined the conceptual challenges and dimensionality (Buehl & Alexander, 2001; Duell & Schommer-Aikins, 2001), psychometric properties that fail to validate theorized notions of students' epistemological development (DeBacker et al., 2008), and socially desirable responding and reliability issues

related to self-report measures of students' cognitive development (Bowman & Seifert, 2011).

Specifically related to the epistemological development dimension of the self-authorship concept that undergirds part of the GPI, scholars have argued that this type of development is quite difficult—if not impossible—to quantify. For instance, studying this developmental dimension does not involve understanding the specific content of respondents' thinking; rather, the focus should involve understanding the structure of respondents' meaning making across a variety of domains (Creamer, Baxter Magolda, & Yue, 2010). Only a few scholars have attempted to quantify students' epistemological development, and each of these efforts has produced inconclusive evidence about the reliability, validity, and utility of doing so (Creamer et al., 2010; Goodman & Siefert, 2009; Pizzolato & Chaudhari, 2009). It may be that direct evidence of students' learning and development (i.e., using artifacts that students actually produce)—as opposed to indirect evidence such as surveys, which ask students about their perception of their learning and development—is more suited to provide evidence of students' increasingly complex abilities to thoroughly explain issues, incorporate different perspectives, and understand the influence of context and assumptions in claims. Or, as other scholars have argued, it may be that qualitative indirect approaches to measuring this aspect of development are better suited given its complexity. Nevertheless, the above recommendations aim to improve the conceptual precision and indirect quantitative measurement of this area of learning and development. Such work stands to benefit those in higher education who wish to conveniently and broadly assess educational interventions designed to promote this area of development (Creamer et al., 2010).

Identity scale. Theoretically, this GPI dimension relates to individuals' awareness and acceptance of their identities as they interact across cultural differences. Kegan (1994) explains that intercultural competence is only possible when an internally defined sense of self exists to mitigate the emotional threat of interacting across difference. The intrapersonal dimension of the intercultural maturity model also addresses the integration of students' values and beliefs into how they live their lives, the ways in which students understand their social identities, and the extent to which they rely on others for self-definition (King & Baxter Magolda, 2005).

On the Identity scale, the three items with factor loadings < .71 are *I have a definite purpose in my life*, *I know who I am as a person*, and *I am willing to defend my views when they differ from others*. These first two items seem theoretically related to the internal sense of self dimension of this scale's theoretical grounding. However, these items could be refined in ways that operationalize these more specifically, making them less ambiguous. For instance, King and Baxter Magolda (2005) explain that at a mature stage of development within this intrapersonal dimension, students should be able to explain their beliefs and values, describe how these have been shaped by personal characteristics (e.g., race, social class, other identities), incorporate such understandings into their choices and behaviors, seek challenge to their views, and continuously reexamine their values and beliefs. The GPI currently includes an item with a relatively high factor loading (.78), *I can explain my own personal values to people who are different from me*. However, refining the two items above with the lower factor loadings to more specifically ask students the extent to which they engage in these other types of behaviors could strengthen this scale.

Additionally, the lower factor loading for the third item above, *I am willing to defend my views when they differ from others*, is especially interesting, namely because an item very similar on the scale, *I put my beliefs into action by standing up for my principles*, has a high factor loading (.75). Perhaps the "defend my views" aspect of the item is understood differently than the "standing up for my principles" aspect of this latter item, though they are functionally quite similar. The earlier item also explicitly involves others ("…when they differ *from others*"), while the latter item leaves room to interpret "standing up for my principles" in a variety of ways (i.e., not necessarily involving others, perhaps involving only one's own actions). As discussed above, theoretically this area involves—among other aspects—seeking challenges to one's views and reexamining one's values and beliefs. Instead of asking about perhaps something that could be understood as more *interpersonal*—defending one's views when they differ from others—this item could instead focus on the *intrapersonal* process that would allow someone to do this.

Cognitive Knowledge scale. Theoretically, this scale measures intercultural awareness, or the extent that individuals know intercultural information, similarities, and differences. Intercultural awareness involves two components, self-awareness and cultural awareness, as one must first develop an understanding of aspects of one's own culture in order to understand others' cultures (Bennett, 2009; Chen & Starosta, 1998; Fritz et al., 2001). The first observation in examining this scale's items is that they all theoretically measure awareness of others' cultures or cultural relations, but no items currently measure one's self-awareness of one's own culture. Like the dimensionality of the Cognitive Knowing scale discussed earlier, this construct is currently operationalized as unidimensional on the GPI. However, that self-awareness of one's culture underlies the awareness of others' cultures presents an opportunity to rethink how this

scale is operationalized. Items could be developed that ask respondents about the extent that they are aware of their own cultural norms. It would be interesting to observe in evaluating any new items the dimensionality of this construct if self-awareness and cultural awareness items were both included.

Second, on the Cognitive Knowledge scale, the three items with factor loadings < .71 are *I am informed of current issues that impact international relations*, *I understand the reasons and causes of conflict among nations of different cultures*, and *I know how to analyze the basic characteristics of a culture*. The first two items may be measuring very broad aspects of this developmental domain. Knowledge of global issues and international conflict are undoubtedly important global learning outcomes. However, instead of asking about cultures and international relations more broadly, item development could include items that ask students about the extent of their culture-specific knowledge (e.g., how much does one know about different religious and nonreligious groups, races and ethnicities, political groups?). From the standpoint of using research or assessment findings, this level of specificity would stand to describe students' intercultural knowledge much more specifically and inform the types of educational opportunities that could promote *particular* types of intercultural awareness. As it is currently written, the last item seems to theoretically relate more to the Cognitive Knowing scale, with its emphasis on critical thinking and analytical skills. This item could be operationalized more specifically (i.e., framing this item in terms of what students would actually be able to do if they could analyze cultural characteristics, and clarifying within the item the basic characteristics of a culture) and included in the refined Cognitive Knowing scale discussed above.

Affect scale. Theoretically, this GPI dimension relates to individuals' respect for and appreciation of difference and their emotional awareness in interacting across difference. This domain of development is often explained in terms of intercultural sensitivity, which involves self-monitoring (i.e., detecting situational aspects and aligning behaviors accordingly), concern for and accurately perceiving others' feelings and experiences, and non-judgment (i.e., an ability to listen to culturally different information without prematurely concluding meaning; Chen & Starosta, 2000). Currently, two items on the Affect scale have factor loadings < .71, *I do not feel threatened emotionally when presented with multiple perspectives* and *I am sensitive to those who are discriminated against*.

For the first item, item comprehension may explain this low factor loading. The *not* in this item is a negative particle, which respondents can easily miss, thereby altering the meaning of the item, resulting in inaccurate responses (van Sonderen et al., 2013). This is likely given that three other items on this scale—all with factor loadings > .70—seem particularly related to this item. The second item may need to be operationalized more specifically. Concern for and ability to perceive others' feelings and experiences is a key element to intercultural sensitivity. It may be more useful to consider if students were sensitive to those who are discriminated against, what they could do that reflects this and instead measure *that* (e.g., emotional awareness during interactions, perceiving others' feelings and experiences, withholding judgment).

Social Responsibility scale. First, one item on the Social Responsibility scale, *Volunteering is not an important priority in my life,* could not be included in the present study given its extremely low factor loading. It is one of the reverse-worded items on the instrument I will discuss next that should be refined. But in addition, this item is distinct from the other items

on this scale. The other items on the Social Responsibility scale explicitly ask about civic attitudes, whereas this item asks about the importance of volunteering in one's life. As mentioned in the last chapter, these items seem to be measuring different dimensions of social responsibility (i.e., perceived civic commitment or civic identity as opposed to civic action). In addition to this item, two other items on this scale, *I think of my life in terms of giving back to society* and *I put the needs of others above my own personal wants*, have factor loadings < .71. While both items seem to be measuring aspects of civic commitment or civic identity, both could benefit from slight refinement. For instance, the first item could easily be interpreted as the extent to which altruism is the most important priority in one's life. In the case of the second item, this could be interpreted as a measure of collectivism orientation. If, instead, the aim is to understand the extent to which students consider the needs of society or actually engage in behaviors that address society's needs, it will be useful to instead ask those.

Social Interaction scale. While the Social Interaction scale demonstrated convergent validity, one item on this scale, *Most of my friends are from my own ethnic background,* could not be included in the present study given its low factor loading. It is also a reverse-worded item, so it asks the inverse of what this scale measures (i.e., it was operationalized in way that asks about ethnically homogenous friendships rather than interactions across difference). Another of this scale's items, *I frequently interact with people from a country different from my own*, has a factor loading < .71. Similar to the suggestion above about the Cognitive Knowledge scale, slightly refining this scale's items to ask—in the same way—the extent to which students interact across several *specific* groups (e.g., retaining the interactions across different races/ethnicities item but adding interactions across individuals with different religious/

nonreligious, political, country of origin, or language backgrounds) adds specificity to this scale and allows for more nuanced findings related to students' interactions across difference in ways that reflect intercultural exchange more comprehensively than the current scale allows.

Eliminating Reverse-worded Items

Reverse-worded items consist of two types: (1) negation reverse-worded items that contain negative particles (e.g., not, no) or affixal morphemes (e.g., un-, non-, dis-, -less) and (2) polar opposite reverse-worded items (e.g., using words with an opposite meaning; Zhang & Savalei, 2015). The GPI currently contains seven reverse-worded items; one item is a negation item, while the remaining six are polar opposite items, making these items the inverse of their respective constructs. Given the observed low factor loadings and high unique variance of all seven reverse-worded items on the GPI, refining these items is warranted.

Tourangeau et al. (2000) explain that response bias is best understood as discrepancies that arise relative to the information an instrument developer seeks from a survey respondent and the information provided by the respondent. Importantly, Weijters (2006) differentiates between two types of response bias: response set (i.e., bias related to an item's *content*, such as social desirability bias) and response style (i.e., a general tendency to respond to items in a particular manner, *regardless of the item's content*). In differentiating these, Weijters explains that the utility of reverse-worded items is typically in addressing particular response styles; different response sets (e.g., social desirability) tend to be less impacted by reverse-worded items. van Sonderen et al. (2013) synthesized the various types of response styles that threaten the validity of self-reported instruments and discussed the use of reverse-worded items to combat particular bias (i.e., acquiescence, inattention, and confusion). First, they define acquiescence as a

respondent's inclination to generally agree with survey items, regardless of content. Reverse-worded items are often used to eliminate acquiescent response styles and identify acquiescence bias (i.e., through discrepant responses). Second, van Sonderen et al. explain that respondents can also fail to carefully read survey items and/or response categories, miss the intended meaning of those items, and therefore provide inaccurate responses. Respondents also engage in satisficing, where they exert less cognitive effort in understanding the content and nuances of the items and response categories (Barge & Gehlbach, 2012). Reverse-worded items are also used to address inattentive response styles since these break up items that otherwise resemble each other and introduce subtle differences, which are thought to help address inattentive and satisficing response styles (van Sonderen et al., 2013). Finally, confusion on the part of respondents can arise if items and/or response categories are unclear (Tourangeau et al., 2000). Negation items are particularly confusing for respondents since negative particles (e.g., not, no) or affixal morphemes (e.g., un-, non-, dis-, -less) are used in the survey item (Swain, Weathers, & Niedrich, 2008), and it is easy for both attentive and inattentive respondents to miss these subtle item differences (Zhang & Savalei, 2015).

Though widely used to combat a variety of validity threats, reverse-worded items present important complications. For instance, reverse-worded items can often suggest multidimensionality and spurious factors (i.e., instead of a developmental construct, for instance, a factor can emerge due to a method effect), reduce scale reliability (Yue, Creamer, & Wolfe, 2009), and bias parameter estimates in SEM and CFA analyses (Zhang & Savalei, 2015). Often, respondents do not understand reverse-worded items as the actual opposite of directly worded items (Barnette, 2000), an issue that is particularly salient to the GPI given that six of its reverse-

worded items are polar opposite items. van Sonderen et al. (2013) tested the effectiveness of reverse-worded items in reducing particular problematic response styles and found that not only were reverse-worded items ineffective at eliminating these biases, but measurement error increased due to issues related to item comprehension and inattention. Given the implications of reverse-worded items for both respondents (i.e., item comprehension difficulties) and survey developers (i.e., increased measurement error, biased estimates, spurious factors), instrument refinement efforts should remove these items from the GPI. Above, I have discussed strategies to rethink these seven items on their respective scales. In attending to the larger issues with the Cognitive Knowing scale and the two other reverse-worded items on the Social Responsibility and Social Interaction scales, the suggestions above make it quite possible to refine these items and present them in ways that directly (and not inversely) measure their respective constructs.

In removing reverse-worded items, there are ways to empirically test for the types of response styles that pose validity threats (see van Vaerenbergh & Thomas, 2012 for a comprehensive review). While there are a variety of methods available to detect particular response styles, adding representative indicators of response styles to an instrument allows survey developers to add items that measure specific response styles (i.e., acquiescence response style, disacquiescence response style, extreme response style, midpoint response style) and calculate the amount of those response styles in the dataset (van Vaerenbergh & Thomas, 2012). Weijters, Schillewaert, and Geuens (2008) recommend adding five items per response style examined; these items could be included in a pilot administration of the refined GPI to determine any influence of response styles. If the original reverse-worded items were mainly added to mitigate acquiescence response styles, for instance, adding five items that measure this response

style would allow for the detection of the presence of acquiescence response styles without the reverse-worded items.

Future Inquiry

Given this study's findings and implications, opportunities for several types of future investigations emerge. First, developing new climate items for the GPI would allow for a deeper, more contextual understanding of the GPI's developmental outcomes. Second, there are measurement issues in cross-cultural research—and within measurement invariance testing in particular—that deserve more intentional investigation than the current study permitted. Third, if items are refined and developed, subsequent validation efforts will be necessary to gauge the effectiveness of those efforts. Each is discussed below.

Developing New Campus Climate Indicators

Given the import of students' educational contexts relative to the development the GPI measures, future inquiry could involve developing new climate items for the GPI. This particular area of inquiry relates to a larger issue that involves how the development measured by the GPI is theoretically grounded. In addition to the campus climate frameworks advanced by Milem et al. (2005) and Hurtado, Milem, Clayton-Pedersen, and Allen (1998) that explain the various dimensions of campus climate relative to the types of learning and development measured by the GPI, two other frameworks underscore the import of understanding students' perceptions of their climate as these relate to their development. First, Tanaka's (2002) intercultural effort concept explains that to measure student engagement across difference without also measuring the cost and effort required to engage across such difference assumes this type of engagement is a neutral process (i.e., that intercultural engagement is equally easy,

welcoming, and educative for all students) when it is not. Tanaka (2002) charges those developing and refining large-scale postsecondary educational surveys to consider this in more accurately measuring students' engagement and the learning and development resultant from it.

Second, Astin's (1991) Input-Environment-Outcomes (I-E-O) framework is widely used in assessing students' learning and development since the model accounts for students' inputs (i.e., the knowledge, skills, and attributes they bring with them to college—or that they have prior to engaging in particular opportunities—to be used as baseline measures) as well as environmental aspects related to their college experiences (i.e., their level/types of engagement, perceptions of their climate, and other environmental attributes) as these relate to learning and developmental outcomes (i.e., what students take away from college). Astin underscores the interrelatedness among the framework's three components and explains that the components should not be isolated in framing the assessment of students' learning and development. In I-E-O terms, the GPI contains input (i.e., through the GPI's New Student form or through using the General form as pre-test measures) and outcome measures (i.e., the GPI's developmental dimensions). However, the GPI General form environmental measures are currently almost entirely behavioral (e.g., curricular and co-curricular intercultural engagement, faculty engagement, current event engagement), capturing the quantity or frequency of these types of engagement but not the level of ease/difficulty/comfort in doing so. The two existing clusters of climate indicators involve two items on the Faculty Engagement scale that measure how much students perceive that faculty foster multiple perspectives in class and items on the GPI's Community scale. However, these community items largely focus on individual-level perceptions of very broad aspects (i.e., with the exception of a single item, the other community

items do not measure perceptions of the campus itself, but rather individuals' perceptions of aspects such as affiliation, institutional mission and whether they have been able to develop their strengths and feel part of a close community).

Particular scales on the current GPI warrant attention around developing campus climate indicators. For example, item development could provide more nuance related to the GPI's Social Interaction dimension of interpersonal development. The three items on this scale used in the present study were: *I frequently interact with people from a race/ethnic group different from my own*, *I intentionally involve people from many cultural backgrounds in my life*, and *I frequently interact with people from a country different from my own*. In addition to understanding the frequency of these interactions—as measured by that scale—new items could also ask how students perceive the opportunities for interaction (i.e., is there diversity present on campus, and are there opportunities to engage across it; the perceived structural diversity that Milem et al., 2005 include) or about the desirability, ease, and costs of such interactions (the perceived psychological climate for diversity; Milem et al., 2005). Similarly, the GPI's Identity dimension of intrapersonal development currently asks three items that would also benefit from contextual considerations: *I can explain my personal values to people who are different from me*, *I am willing to defend my own views when they differ from others*, and *I put my beliefs into action by standing up for my principles*. In understanding whether students perceive a sense of agency in enacting deeply personal parts of their identities, the concept of silenced identities (McLean, 2008; Shapiro et al., 2010) underscores the role of context in silencing marginalized groups. New climate items could ask how students perceive the ease/difficulty of doing this and how agentic they actually feel in doing this.

Such a consideration of campus climate would illuminate a more complete understanding of students' development and provide numerous opportunities for different inquiry into the areas of development measured by the GPI. Survey-based efforts require balancing thoroughness and parsimony. In addition to its global perspective development items, the GPI General Form includes several other measures (i.e., curricular and co-curricular activities, engagement with faculty, experiences with current events, and sense of community; RISE, 2017b). Adding climate measures certainly adds to the length of the survey. However, developers of the GPI might consider adding topical modules as the National Survey of Student Engagement (NSSE) has done; institutions can elect to add additional items to the NSSE core survey to learn more about in-depth areas the core survey does not cover (e.g., civic engagement, experiences with diverse perspectives; NSSE, 2018). Or those developing the GPI might elect to develop a separate form of the GPI that measures the same developmental dimensions but that also includes a comprehensive institutional climate assessment, similar to the Higher Education Research Institute's (HERI's) Diverse Learning Environments survey (HERI, 2017).

Measurement Issues in Cross-cultural Research

Response scales. On the part of instrument developers, the use of Likert scales presents unique challenges for cross-cultural research. While the self-reported measurement of students' attitudes, perceptions, and behaviors is imperfect in a broader sense, the relative and subjective nature of Likert scales adds to this complexity. Response scales can obscure cultural variability related to particular psychological constructs when such variation actually exists; this presents an important opportunity to pause and reflect on what can be done. Important to cross-cultural research, the process where respondents map attitudes or perceptions to a response category is

culturally variable (Hui & Triandis, 1989). Heine et al. (2002) explain this in terms of a shifting standards effect; individuals evaluate themselves relative to close comparison groups, not the population as a whole, so survey items "do not provide a context-free assessment of one's absolute standing" (p. 905).

Likert scales are particularly subject to shifting standards effects since the scales are subjective and what is asked about in a survey item is considered relative, usually culturally- or contextually-bound (Heine et al., 2002). If there are actual differences related to the item, these are likely to be concealed by subjective Likert scales. This makes examining measurement invariance across groups particularly challenging in that *lack* of invariance can be easily obscured (Heine et al., 2002). An item on the GPI's Social Interaction scale illustrates the possibility of a shifting standards effect. Like other survey items, *I frequently interact with people from a race/ethnic group different from my own*, is relative. Take, for instance, a student whose context is racially/ethnically homogenous. Given that context, she does not encounter much intergroup interaction; she interacts across racial/ethnic difference only a few times each week. A different student's context is very racially/ethnically diverse. Given this student's surroundings, she interacts across racial/ethnic difference multiple times each day. When both are presented this survey item, the former student could respond "Strongly agree" (based on that student's context and her peers, her interaction *seems* frequent), and the latter student could also provide the same response. But these students' *actual* interaction across racial/ethnic difference is markedly different. The Likert scale obscures the variability in these students' actual experiences.

Future inquiry could develop and test different response scales for the GPI. Though Heine et al. (2002) explain the value of using objective response scales when possible (e.g., reporting actual frequencies of interactions), frequency-based response scales are still subject to inaccurate responses related to inaccurate recall of factual survey items (Tourangeau et al., 2000). An emerging area of inquiry is related to the use of situational judgment items, which require respondents to read scenarios and indicate the likelihood of doing something as a result of the scenario; the constructs under study are operationalized differently (i.e., behaviorally) through these response options (Whetzel et al., 2008). In particular, combining situational judgment items with self-reported measures has been shown to explain more variance in intercultural competency outcomes compared to self-report measures by themselves (Leung, Ahn, & Tan, 2014; Schnabel, Kelava, Seifert, & Kulbrodt, 2015). Leung et al. (2014) explain that the measurement of intercultural competencies ought to align with the type of competencies theorized. For instance, are the intercultural competencies under study theorized as abilities or skills, knowledge, attitudes, behaviors, or come combination of these? In refining and developing items for the GPI, attention to the constructs best measured by self-report methods (i.e., those measuring aspects of one's self-concept) and those better measured by more performance-based measures (i.e., behavioral preferences) could address this measurement challenge in cross-cultural research (Schnabel et al., 2015). Schnabel and Kelava's (2013) work illustrates this point. They developed and validated the Test to Measure Intercultural Competence to explicitly use self-report scales to measure individuals' self-concept and situational judgment items to measure behavioral tendencies. This is particularly interesting given the developmental dimensions measured by the GPI. Five of the six developmental scales

(i.e., all except the Cognitive Knowledge scale) could be operationalized to measure behavioral tendencies reflective of the types of development on the scales. For instance, what should students be able to do to reflect evaluativist epistemological understandings (Cognitive Knowing)? In what do students engage to have their views challenged (Identity)? How often do students engage in particular activities that address society's needs (Social Responsibility)? Future inquiry could seek to refine the GPI's items as discussed above with particular attention to operationalizing these behaviorally where possible. Contributing toward this type of inquiry aligns with what is becoming an emerging need to approach the indirect measurement of students' intercultural competencies using multiple methods (i.e., self-report *and* performance-based measures such as situational judgment items) that adequately capture the complexities of such learning and development across diverse individuals (Schnabel et al., 2015).

Determinants of measurement invariance in cross-cultural research. Finally, this study's findings illuminated an aspect of measurement invariance testing within an SEM framework that is currently unclear in both social science and research methodology literature. In MGCFA, using particular fit indices to compare the fit of increasingly more constrained models to determine measurement invariance is currently a process involving several unclear contingencies and little agreement among researchers. Further complicating this is that much of the research on particular model fit criteria has been conducted using normal-theory maximum likelihood estimation methods (i.e., assuming continuous data) and has excluded ordered, categorical data that require robust estimators (Sass, Schmitt, & Marsh, 2014). Still further, it remains unclear whether model fit criteria apply to hierarchical CFA models, such as the one used in the present study (Rudnev, Lytkina, Davidov, Schmidt, & Zick, 2018).

In this study, when testing invariance across particular parameters, at times I obtained a statistically significant scaled chi-square difference test yet no appreciable difference in CFI. This phenomenon has been widely observed and largely discussed in terms of the high power afforded by large sample sizes of the scaled chi-square difference test to detect otherwise negligible changes in model fit. As such, the use of alternative fit indices (i.e., CFI) have been advanced as the standard in evaluating measurement invariance (Chen, 2007; Cheung & Rensvold, 2002; Meade et al., 2008). However, emerging research has uncovered serious issues in using both chi-square difference tests *and* commonly used alternative fit indices (e.g., inconsistent Type I error rates) and has instead suggested the use of permutation tests (i.e., initially randomly permuting group assignment to yield no initial cross-group differences; Jorgensen, Kite, Chen, & Short, 2017) to detect true measurement invariance. Subsequent validation efforts using GPI data can contribute toward the ongoing understanding of measurement invariance testing by integrating some of these emerging methodological approaches.

Developing, Refining, and Validating the GPI

Considering the aforementioned implications related to instrument refinement and the possibility of developing new climate items, future inquiry should involve refining and developing the items as described above. Instrument development or refinement involves three steps: either deductively or inductively determining the precise content areas, item generation, and instrument construction (Zamanzadeh, Ghahramanian, Rassouli, Abbaszadeh, Alavi-Maid, & Nikanfar, 2015). After refining and developing items, the next logical validation step would involve examining the GPI's content validity, which investigates the extent that the items

comprehensively represent the domains they are meant to measure. During this process, the instrument is subjected to a panel of both content-area experts (i.e., professionals with substantive theoretical or applied knowledge of the constructs under study) and lay audiences (i.e., those who would actually participate in the survey) to determine the representativeness, importance, and clarity of how the items were operationalized. Content validity can be quantified using a content validity index, where the panel rates the relevancy and clarity of all items using an established rating scale; Zamanzadeh et al. provide a comprehensive method for using content validity indices as evidence. Once content validity is established, reliability can be initially assessed by examining the internal consistency of the GPI's refined and developed scales (i.e., whether a scale's items appear to be related) and by assessing test-retest reliability (i.e., whether the same participants' scores are consistent over a period of time). From there, proceeding with exploratory and confirmatory factor analyses to understand and validate (respectively) the refined instrument's dimensionality will be important.

Subsequent validation efforts should continue to examine the cross-cultural validity of the GPI in terms of students' ethnoracial identities and domestic/international status. However, the development of new climate items (as discussed) for future validation efforts would be particularly beneficial. This would permit cross-group validation studies based on important contextual differences. For instance, would the GPI's constructs and refined items function differently as a condition of students' perceptions of particular psychological (e.g., how divisive, discriminatory, welcoming) or behavioral (e.g., extent of opportunities to engage across difference, effortful/comfortable interactions) dimensions of their educational contexts? As I outlined in Chapter Two, ample theory suggests this may be the case; it is a worthy investigation.

In alignment with Heine et al. (2002), who argued that examining cross-cultural variability ought to involve comparing responses across more than one condition within individual cultures, empirically determining the extent that this is so has important implications for both practice and research.

For instance, one could hypothetically examine measurement invariance across two groups (e.g., one group perceives divisive, tense campus contexts, the other group does not). One could determine whether these two groups understand the GPI's developmental constructs differently and whether any systematic measurement bias exists as a condition of their group membership. If findings suggested a lack of invariance, this would indicate that these concepts functioned differently across these. The groups' different understanding of these dimensions of development could be conditional on whether they perceive the campus as divisive. This ought to then raise questions about the role of campus contexts in students' learning and development. For students who view their campuses as divisive, this hypothetical evidence would suggest that they literally understand and/or respond to measures of particular developmental outcomes differently. If educational opportunities are framed in terms of these widely used outcomes, such efforts might not be framed—much less evaluated—inclusively.

Conclusion

This study sought to address gaps related to validation efforts for both large-scale postsecondary educational surveys and, more specifically, measures of intercultural competencies. In this chapter, the study's measurement invariance findings support particular uses of the GPI (i.e., comparing cross-group regression coefficients and means across the four groups under study, understanding the instrument's conceptualization as culturally inclusive).

Other validation evidence from this study stands to inform particular ways the instrument could be refined. The study's findings—and the process of situating these within larger bodies of campus climate and cross-cultural measurement literature—also illuminated several avenues of future inquiry. But the impact of this study extends well beyond the contributions and practical implications discussed earlier.

Executing this study allowed me—at times *required* me—to reflect on my transformation as a doctoral student. I entered this experience interested in learning how to measure the benefits of diverse learning contexts for both college students and the faculty who teach them. On some level, I am leaving this experience equipped to do precisely that. A superficial understanding of my start and end points in this program might suggest I did not deviate much from what I initially endeavored to do. However, the path to make this assertion at this point has been anything but straightforward. At the outset of this experience, I approached research opportunities asking, "Who am *I* to investigate this?" Looking back, this arose from a space many doctoral students occupy at some point: the ongoing concern whether I would *ever* feel efficacious as a scholar.

As I conclude this experience, I pose the same question to myself, only now asking, "*Who am I* to investigate this?" This self-examination emerges after nearly five years of deeply questioning my own identities, contexts, and understandings and grappling with the immense implications of contributing to our field's scholarship. I had to first unlearn particular ways of knowing in order to learn in ways that will continue to inform my research. The critical epistemologies and theoretical perspectives I studied forced me to wrestle with who and what I centered in the concepts and processes that I had not yet interrogated. At times, the tensions I

felt between epistemologies and methodologies felt so pronounced that I considered refocusing my own methodological orientation. But my training also instilled in me a strong desire for theoretical and analytical precision. And, importantly, the concept of precision took on a new personal meaning. It transcended methodological training and located my habits of mind just as centrally to the research process. I believe the domain of survey research to be an ideal space in which to continue thinking in this way.

The development and use of large-scale educational surveys will undoubtedly continue, especially as U.S. higher education finds itself increasingly situated within accountability and evidence-based frameworks. Those of us developing and using educational surveys have tremendous power related to the stories we tell about students' experiences, learning, and development. This is an enormous responsibility that requires us to continuously consider how we *understand* students' learning and development in the first place. What (and who) shapes such an understanding? Based on our own epistemological orientations, what theories of students' learning and development do we aim to operationalize into survey measures? How do we measure all of this? And, finally, how can we ensure that the theorization and measurement are inclusive across diverse student populations? As I leave this experience, I continue to reflect on the immense responsibility involved in personally answering these questions *and* in raising these considerations among survey researchers more broadly.

REFERENCE LIST

Aguinis, H., Pierce, C.A., & Culpepper, S.A. (2009). Scale coarseness as a methodological artifact: Correcting correlation coefficients attenuated from using coarse scales. *Organizational Research Methods, 12*(4), 623-652.

Alcoff, L.M. (2009). Latinos beyond the binary. *The Southern Journal of Philosophy, 47*, 112-128.

Allison, P.D. (2003). Missing data techniques for structural equation modeling. *Journal of Abnormal Psychology, 112*(4), 545-557.

Allport, G.W. (1954). *The nature of prejudice*. Cambridge, MA: Addison-Wesley.

Altbach, P.G., & Knight, J. (2007). The internationalization of higher education: Motivations and realities. *Journal of Studies in International Education, 11*(3/4), 290-305.

American Council on Education. (2012). *Mapping internationalization on U.S. campuses: 2012 edition.* Retrieved from http://www.acenet.edu/news-room/Documents/Mapping-Internationalizationon-US-Campuses-2012-full.pdf

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: Authors.

Ancis, J.R., Sedlacek, W.E., & Mohr, J.J. (2000). Student perceptions of campus cultural climate by race. *Journal of Counseling and Development, 78*(2), 180-185.

Anderson, P.H., & Lawton, L. (2011). Intercultural development: Study abroad vs. on-campus study. *Frontiers: The Interdisciplinary Journal of Study Abroad*, *21*, 86-108.

Andreotti, V. (2010). Postcolonial and post-critical 'global citizenship education.' In G. Elliott, C. Fourali, & S. Issler (Eds.), *Education and social change: Connecting local and global perspectives* (pp. 233-245). London, UK: Continuum International Publishing Group.

Antonio, A.L., Chang, M.J., Hakuta, K., Kenny, D.A., Levin, S., & Milem, J.F. (2004). Effects of racial diversity on complex thinking in college students. *Psychological Science, 15*(8), 507-510.

Applebaum, B. (2003). Social justice, democratic education, and the silencing of words that wound. *Journal of Moral Education, 32*(2), 151-162.

Arminio, J.L., Carter, S., Jones, S.E., Kruger, K., Lucas, N., Washington, J., & Scott, A. (2000). Leadership experiences of students of color. *NASPA Journal, 37*, 496-510.

Association of American Colleges and Universities. (2007). *College learning for the new global century: A report from the National Leadership Council for Liberal Education and America's Promise.* Washington, DC: Author.

Association of American Colleges and Universities. (2017). *Intercultural knowledge and competence VALUE rubric*. Retrieved from https://www.aacu.org/value/rubrics/intercultural-knowledge

Astin, A.W. (1991). *Assessment for excellence: The philosophy and practice of assessment and evaluation in higher education*. Washington, DC: American Council on Education.

Astin, A.W. (1993). *What matters in college? Four critical years revisited*. San Francisco, CA: Jossey-Bass.

Astin, A.W., & Lee, J.J. (2003). How risky are one-shot cross-sectional assessments of undergraduate students? *Research in Higher Education, 44*(6), 657-672.

Atkinson, D.R., Morten, G., & Sue, D.W. (1989). A minority identity development model. In D.R. Atkinson, G. Morten, & D.W. Sue (Eds.), *Counseling American min*orities (pp. 35-52). Dubuque, IA: W.C. Brown.

Aud, S., Fox, M.A., & KewalRamani, A. (2010). *Status and trends in the education of racial and ethnic groups (NCES 2010-15).* Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Barge, S., & Gehlbach, H. (2012). Using the theory of satisficing to evaluate the quality of survey data. *Research in Higher Education, 53*, 182-200.

Barnette, J.J. (1996). Responses that may indicate nonattending behaviors in three self-administered educational attitude surveys. *Research in the Schools, 3*(2), 49-59.

Barnette, J.J. (2000). Effects of stem and Likert response option reversals on survey internal consistency: If you feel the need, there is a better alternative to using those negatively worded stems. *Educational and Psychological Measurement, 60*(3), 361-370.

Bennett, M.J. (1986). A developmental approach to training for intercultural sensitivity. *International Journal of Intercultural Relations, 10*(2), 179-196.

Bennett, M.J. (1993). Towards a developmental model of intercultural sensitivity. In R. M. Paige (Ed.), *Education for the intercultural experience*. Yarmouth, ME: Intercultural Press.

Bennett, M.J. (2009). Defining, measuring, and facilitating intercultural learning: A conceptual introduction to the Intercultural Education double supplement. *Intercultural Education, 20*(S1-2), S1-13.

Bernstein, I.H., & Teng, G. (1989). Factoring items and factoring scales are different: Spurious evidence for multidimensionality due to item categorization. *Psychological Bulletin, 105*, 467-477.

Bettencourt, B.A., Dorr, N., Charlton, K., & Hume, D.L. (2001). Status differences and in-group bias: A meta-analytic examination of the effects of status stability, status legitimacy, and group permeability. *Psychological Bulletin, 127*(4), 520-542.

Bhawuk, D.P.S., & Brislin, R. (1992). The measurement of intercultural sensitivity using the concepts of individualism and collectivism. *International Journal of Intercultural Relations, 16*, 413-436.

Binder, J.R., & Desai, R.H. (2011). The neurobiology of semantic memory. *Trends in Cognitive Science, 15*(11), 527-536.

Bing, V.M., & Reid, P.T. (1996). Unknown women and unknowing research: Consequences of color and class in feminist psychology. In N.R. Goldberger, J.M. Tarule, B.M. Clinchy, & M.F. Belenky (Eds.), *Knowledge, difference, and power: Essays inspired by women's ways of knowing* (pp. 175-202). New York, NY: Basic Books.

Bonilla-Silva, E., & Forman, T.A. (2000). "I am not a racist but . . .": Mapping white college students' racial ideology in the USA. *Discourse and Society, 11*(1), 50-85.

Bowen, N.K., & Masa, R.D. (2015). Conducting measurement invariance tests with ordinal data: A guide for social work researchers. *Journal of the Society for Social Work and Research, 6*(2), 229-249.

Bowman, N.A. (2011). Validity of college self-reported gains at diverse institutions. *Educational Researcher, 40*(1), 22-24.

Bowman, N.A., & Hill, P.L. (2011). Measuring how college affects students: Social desirability and other potential biases in college student self-reported gains. *New Directions for Institutional Research, 150*, 73-85.

Bowman, N.A., & Park, J.J. (2014). Interracial contact on college campuses: Comparing and contrasting predictors of cross-racial interaction and interracial friendship. *The Journal of Higher Education, 85*(5), 660-690.

Bowman, N.A., & Seifert, T.A. (2011). Can college students accurately assess what affects their learning and development? *Journal of College Student Development, 52*(3), 270-290.

Braskamp, L.A., Braskamp, D.C., & Engberg, M.E. (2014). *Global Perspective Inventory (GPI): Its purpose, construction, potential uses, and psychometric characteristics*. Chicago, IL: Global Perspective Institute.

Braskamp, L.A., Braskamp, D.C., & Merrill, K. (2009). Assessing progress in global learning and development of students with education abroad experiences. *Frontiers: The Interdisciplinary Journal of Study Abroad, 18*, 101-118.

Brinkman, U., & van Weerdenburg, O. (2014). *Intercultural readiness: Four competencies for working across cultures*. New York, NY: Palgrave MacMillan.

Broido, E.M. (2000). The development of social justice allies during college: A phenomenological investigation. *Journal of College Student Development, 41*(1), 3-18.

Broido, E.M., & Schreiber, B. (2016). Promoting student learning and development. *New Directions for Higher Education, 2016*, 65-74.

Brown, L.I. (2004). Diversity: The challenge for higher education. *Race Ethnicity and Education, 7*(1), 21-34.

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

Bryant, F.B., & Satorra, A. (2012). Principles and practice of scaled difference chi-square testing. *Structural Equation Modeling, 19*, 372-398.

Bryant, F.B., & Satorra, A. (2013). *EXCEL macro file for conducting scaled difference chi-square tests via LISREL 8, LISREL 9, EQS, and Mplus*. Available from the authors.

Buehl, M.M., & Alexander, P.A. (2001). Beliefs about academic knowledge. *Educational Psychology Review, 13*, 385-418.

Byrne, B.M., Shavelson, R.J., & Muthén, B.O. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*(3), 456-466.

Cabrera, A.F., Nora, A., Terenzini, P.T., Pascarella, E., & Hagedorn, L.S. (1999). Campus racial climate and the adjustment of students to college. *The Journal of Higher Education, 70*(2), 134-160.

Campbell, C.M., & Cabrera, A.F. (2011). How sound is NSSE? Investigating the psychometric properties of NSSE at a public, research-extensive institution. *The Review of Higher Education, 35*(1), 77-103.

Carnegie Council for Ethics in International Affairs. (2017). Challenges of globalization. Retrieved from https://www.carnegiecouncil.org/publications/archive/dialogue/1_11/relevance_social/588

Carter, D.J. (2007). Why the black kids sit together at the stairs: The role of identify-affirming counter-spaces in a predominantly white school. *The Journal of Negro Education, 76*(4), 542-554.

Carter, P.L. (2010). Race and cultural flexibility among students in different multiracial schools. *Teachers College Record, 112*(6), 1529-1574.

Chai, C.S. Hong, H.Y., & Teo, T. (2009). Singaporean and Taiwanese pre-service teachers' beliefs and their attitude towards ICT use: A comparative study. *The Asia-Pacific Education Researcher, 18*(1), 117-128.

Chan, K., & Elliott, R.G. (2004). Relational analysis of personal epistemology and conceptions about teaching and learning. *Teaching and Teacher Education, 20*(8), 817-831.

Chang, M.J. (1999). Does racial diversity matter? The educational impact of a racially diverse undergraduate population. *Journal of College Student Development, 40*, 377-395.

Chang, M.J. (2001). The positive educational effects of racial diversity on campus. In G. Orfield & M. Kurlaender (Eds.), *Diversity challenged: Evidence on the impact of affirmative action* (pp. 175-186). Cambridge, MA: Harvard Education Publishing Group.

Chang, M.J., Astin, A.W., & Kim, D. (2004). Cross-racial interaction among undergraduates: Some consequences, causes, and patterns. *Research in Higher Education, 45*, 529-553.

Chen, F.F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 464-504.

Chen, F.F., Sousa, K.H., & West, S.G. (2005). Testing measurement invariance of second-order factor models. *Structural Equation Modeling, 12*(3), 471-492.

Chen, G.M. (1997, January). *A review of the concept of intercultural sensitivity*. Paper presented at the biennial convention of Pacific and Asian Communication Association. Honolulu, Hawai'i.

Chen, G.M. (2014). Intercultural communication competence: Summary of 30-year research and directions for future study. In X. Dai & G.M. Chen (Eds.), *Intercultural communication competence: Conceptualization and its development in cultural contexts and interactions* (pp. 14-40). Newcastle, UK: Cambridge Scholars Publishing.

Chen, G.M., & Starosta, W.J. (1996). Intercultural communication competence: A synthesis. *Annals of the International Communication Association*, *19*(1), 353-384.

Chen, G.M., & Starosta, W.J. (1997). A review of the concept of intercultural sensitivity. *Human Communication, 1*, 1-16.

Chen, G.M., & Starosta, W.J. (1998). A review of the concept of intercultural awareness. *Human Communication, 2*, 27-54.

Chen, G.M., & Starosta, W.J. (2000). The development and validation of the Intercultural Sensitivity Scale. *Human Communication, 3*(1), 3-14.

Cheung, G.W., & Rensvold, R.B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233-255.

Chickering, A.W., & Reisser, L. (1993). *Education and identity* (2nd ed.). San Francisco, CA: Jossey-Bass.

Cizek, G.J., Bowen, D., & Church, K. (2010). *Sources of validity evidence for educational and psychological tests: A follow-up study*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Denver, CO.

Clark, R., Anderson, N.B., Clark, V.R., & Williams, D.R. (1999). Racism as a stressor for African Americans: A biopsychosocial model. *American Psychologist, 54*, 805-816.

Cone, J.D., & Foster, S.L. (1991). Training in measurement: Always the bridesmaid. *American Psychologist, 46*(6), 653-654.

Cooperative Institutional Research Program. (2017). *Using CIRP surveys in accreditation*. Retrieved from https://heri.ucla.edu/using-cirp-surveys-in-accreditation/

Council for the Advancement of Standards in Higher Education. (2015). *CAS learning and development outcomes*. Retrieved from http://www.cas.edu/learningoutcomes

Creamer, E.G., Baxter Magolda, M., & Yue, J. (2010). Preliminary evidence of the reliability and validity of a quantitative measure of self-authorship. *Journal of College Student Development, 51*(5), 550-562.

Cross, W.E. (1991). *Shades of black: Diversity in African-American identity*. Philadelphia, PA: Temple University Press.

D'Augelli, A.R., & Hershberger, S.L. (1993). African American undergraduates on a predominantly white campus: Academic factors, social networks, and campus climate. *The Journal of Negro Education, 62*(1), 67-81.

Davidson, L.M. (2015). *Examining measurement invariance for the Global Perspective Inventory*. Unpublished manuscript.

Davidson, L.M., & Engberg, M.E. (under review). Examining the psychometric properties of the Global Perspective Inventory. *International Journal of Intercultural Relations*.

Deardorff, D.K. (2004). *The identification and assessment of intercultural competence as a student outcome of internationalization at institutions of higher education in the United States* (Doctoral dissertation). Retrieved from ProQuest Dissertations Publishing. (3128751)

Deardorff, D.K. (2015). *Demystifying outcomes assessment for international educators: A practical approach.* Sterling, VA: Stylus.

DeBacker, T.K., Crowson, H.M., Beesley, A.D., Thoma, S.J., & Hestevold, N.L. (2008). The challenge of measuring epistemic beliefs: An analysis of three self-report instruments. *The Journal of Experimental Education, 76*(3), 281-312.

de Beurs, D.P, Fokkema, M., de Groot, M.H., de Keijser, J., & Kerkhof, A.J. (2015). Longitudinal measurement invariance of the Beck Scale for Suicide Ideation. *Psychiatry Research, 225*(3), 368-373.

Delgado Bernal, D. (2002). Critical race theory, Latino critical theory, and critical raced-gendered epistemologies: Recognizing students of color as holders and creators of knowledge. *Qualitative Inquiry, 8*(1), 105-126.

Denson, N., & Ing, M. (2014). Latent class analysis in higher education: An illustrative example of pluralistic orientation. *Research in Higher Education, 55*, 508-526.

Der-Kerabetian, A., & Metzer, J. (1993). The Cross-cultural World Mindedness Scale and political party affiliation. *Psychological Reports, 72*(3), 1069-1070.

de Wit, H. (2013). Reconsidering the concept of internationalization. *International Higher Education, 70*, 6-7.

Dimitrov, D.M. (2010). Testing for factorial invariance in the context of construct validation. *Measurement and Evaluation in Counseling and Development, 43*(2), 121-149.

Dovidio, J.F., Gaertner, S.L., & Saguy, T. (2009). Commonality and the complexity of "we": Social attitudes and social change. *Personality and Social Psychology Review, 13*(3), 3-20.

Dowd, A.C., Sawatzky, M., & Korn R. (2011). Theoretical foundations and a research agenda to validate measures of intercultural effort. *The Review of Higher Education, 35*(1), 17-44.

Duell, O.K., & Schommer-Aikins, M. (2001). Measures of people's beliefs about knowledge and learning. *Educational Psychology Review, 13*, 419-449.

Dugan, J.P., Komives, S.R., & Segar, T.C. (2008). College student capacity for socially responsible leadership: Understanding norms and influences of race, gender, and sexual orientation. *NASPA Journal, 45*, 475-500.

Dunton, B.C., & Fazio R.H. (1997). An individual difference measure of motivation to control prejudiced reactions. *Personality and Social Psychology Bulletin, 23*, 316-326.

Duster, T. (1991). *The diversity project*. Berkeley, CA: Institute for the Study of Social Change.

Engberg, M.E. (2004). Improving intergroup relations in higher education: A critical examination of the influence of educational interventions on racial bias. *Review of Educational Research, 74*(4), 473-524.

Engberg, M.E. (2007). Educating the workforce for the 21st century: A cross-disciplinary analysis of the impact of the undergraduate experience on students' development of a pluralistic orientation. *Research in Higher Education, 48*(3), 283-317.

Engberg, M.E., Davidson, L.M., Manderino, M., & Jourian, T.J. (2016). Examining the relationship between intercultural engagement and undergraduate students' global perspective. *Multicultural Education Review*, *8*(4), 253-274.

Engberg, M.E., & Fox, K. (2011). Service participation and the development of a global perspective. *Journal of Student Affairs Research and Practice, 48*(1), 85-105.

Engberg, M.E., & Hurtado, S. (2011). Developing pluralistic skills and dispositions in college: Examining racial/ethnic group differences. *The Journal of Higher Education, 82*(4), 416-443.

Engberg, M.E., Jourian, T.J., & Davidson, L.M. (2016). The mediating role of intercultural wonderment: Connecting programmatic components to global outcomes in study abroad. *Higher Education, 71*(1), 21-37.

Engberg, M.E., Meader, E.W., & Hurtado, S. (2003, April). *Developing a pluralistic orientation: A comparison of ethnic minority and White College students*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Eubanks, D. (2017, Fall). A guide for the perplexed. *Intersection of Assessment and Learning*. Retrieved from http://www.aalhe.org/page/Intersection.

Ferrari, J.R., Bristow, M., & Cowman, S.E. (2005). Looking good or being good? The role of social desirability tendencies in student perceptions of institutional mission and values. *College Student, 39*(1), 7-13.

Fine, M., Weis, L., Powell, M., & Wong, F. (1997). *Off-white: Readings on race, power, and society*. New York, NY: Routledge.

Finney, S.J., & DiStefano, C. (2006). Nonnormal and categorical data in structural equation models. In G.R. Hancock, & R.O. Mueller (Eds.), *A second course in structural equation modeling* (pp. 269-314). Greenwich, CT: Information Age.

Fordham, S., & Ogbu, J.U. (1986). Black students' school success: Coping with the "burden of 'acting White.'" *The Urban Review, 18*(3), 176-206.

Frankenberg, R. (1993). *White women, race matters*. Minneapolis, MN: University of Minnesota Press.

Fritz, W., Möllenberg, A., & Chen, G.M. (2001). *Measuring intercultural sensitivity in different cultural context*. Paper presented at the biannual meeting of the International Association for Intercultural Communication Studies. Hong Kong.

Fritz, W., Möllenberg, A., & Chen, G.M. (2002). Measuring intercultural sensitivity in different cultural contexts. *Intercultural Communication Studies, 11*(2), 165-176.

Fritz, W., Möllenberg, A., & Chen, G.M. (2003). Die interkulturelle Sensibilität als Anforderung an Entsandte – Bedeutung und Elemente für ein Meßmodell. In K. P. Wiedmann (Ed.), *Fundierung des Marketing* (pp. 231-258). Wiesbaden, Germany: Gabler.

Fritz, W., Graf, A., Hentze, J., Möllenberg, A., & Chen, G.M. (2005). An examination of Chen and Starosta's Model of Intercultural Sensitivity in Germany and United States. *Intercultural Communication Studies, 14*(1), 53-65.

Gilbert, E. (2018, January 12). An insider's take on assessment: It may be worse than you thought. *The Chronicle of Higher Education*. Retrieved from https://www.chronicle.com/article/An-Insider-s-Take-on/242235

Gillborn, D. (2005). Education policy as an act of white supremacy: whiteness, critical race theory and education reform. *Journal of Education Policy, 20*(4), 485-505.

Glass, C.R., & Braskamp, L.A. (2012). Foreign students and tolerance. *Inside Higher Ed*. Retrieved from http://www. insidehighered.com/views/2012/10/26/essay-how-colleges-should-respond-racism-against-international-students

Glass, C.R., Buus, S., & Braskamp, L.A. (2013). *Uneven experiences: What's missing and what matters for today's international students*. Chicago, IL: Global Perspective Institute.

Goldberger, N.R. (1996). Women's constructions of truth, self, authority, and power. In H. Rosen, & K. Kuehlwein (Eds.), *Constructing realities: Meaning making perspectives for psychotherapists* (pp. 167-194). San Francisco, CA: Jossey-Bass.

Goodman, K.M., & Seifert, T.A. (2009, April). The process of developing a brief survey of self-authorship. Paper presented at the American Educational Research Association National Conference. San Diego, CA.

Goodwin, L.D., & Leech, N.L. (2003). The meaning of validity in the new Standards for Educational and Psychological Testing: Implications for measurement courses. *Measurement and Evaluation in Counseling and Development, 36*, 181-191.

Gonyea, R.M. (2005). Self-reported data in institutional research: Review and recommendations. *New Directions for Institutional Research, 127*, 73-89.

Green, M.F. (2012). *Measuring and assessing internationalization*. Washington, DC: NAFSA Association of International Educators.

Green, M.F. (2013). *Improving and assessing global learning*. Washington, DC: NAFSA Association of International Educators.

Gudykunst, W.B. (1993). Toward a theory of effective interpersonal and intergroup communication: An anxiety/uncertainty management perspective. In R.L. Wiseman, & J. Koester (Eds.), *Intercultural communication competence* (pp. 33-71). Thousand Oaks, CA: Sage Publications.

Gumport, P.J. (2000). Academic restructuring: Organizational change and institutional imperative. *Higher Education, 39*(1), 67-91.

Gurin, P. (1999). Selections from *The Compelling Need for Diversity in Higher Education,* Expert Reports in Defense of the University of Michigan. *Equity and Excellence in Education, 32*(2), 36-62.

Gurin, P., Dey, E.L., Hurtado, S., Gurin, G. (2002). Diversity and higher education: Theory and impact on educational outcomes. *Harvard Educational Review, 72*(3), 330-366.

Hair, J.F., Hult, G.T.M., Ringle, C.M., & Sarstedt, M. (2014). *A primer on partial least squares structural equation modeling (PLS-SEM)*. Thousand Oaks, CA: Sage.

Hamm, J.V., & Coleman, H.L.K. (2001). African American and White adolescents' strategies for managing cultural diversity in predominantly White high schools. *Journal of Youth and Adolescence, 30*(3), 281-303.

Hammer, M.R. (2005). The Intercultural Conflict Style Inventory: A conceptual framework and measure of intercultural conflict resolution approaches. *International Journal of International Relations, 29*(6), 675-695.

Hammer, M.R. (2011). Additional cross-cultural validity testing of the Intercultural Development Inventory. *International Journal of Intercultural Relations, 35*, 474-487.

Hammer, M.R., Bennett, M.J., & Wiseman, R. (2003). Measuring intercultural sensitivity: The Intercultural Development Inventory. *International Journal of Intercultural Relations, 27*, 421-443.

Hanvey, R.G. (1987). Cross-culture awareness. In L.F. Luce, & E.C. Smith (Eds.), *Toward internationalism* (pp. 13-23). Cambridge, MA: Newbury.

Harper, S.R. (2012). Race without racism: How higher education researchers minimize racist institutional norms. *The Review of Higher Education, 36*(1), 9-29.

Harper, S.R., & Hurtado, S. (2007). Nine themes in campus racial climates and implications for institutional transformation. *New Directions for Student Services, 2007*(120), 7-24.

Harper, S.R., & Quaye, S.J. (2007). Student organizations as venues for Black identity expression and development among African American male student leaders. *Journal of College Student Development, 48*(2), 127-144.

Harrison, K.P. (2012). *Measuring youth experiences of youth development programs: Measurement invariance across gender, racial-ethnic group, and age*. (Unpublished dissertation). University of Connecticut, Mansfield, CT.

He, J., & van de Vijver, F. (2012). Bias and equivalence in cross-cultural research. *Online Readings in Psychology and Culture, 2*(2). Retrieved from https://doi.org/10.9707/2307-0919.1111

Heine, S.J., Lehman, D.R., Peng, K., & Greenholtz, J. (2002). What's wrong with cross-cultural comparisons of subjective Likert scales? The reference-group effect. *Journal of Personality and Social Psychology, 82(*6), 903-918.

Helms, J.E. (1995). An update of Helms' white and people of color racial identity models. In J.G. Ponterotto, J.M. Casas, L.A. Suzuki, & C.M. Alexander (Eds.), *Handbook of multicultural counseling* (pp. 181-198). Thousand Oaks, CA: Sage Publications.

Henseler, J., Ringle, C.M., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science, 43(*1), 115-135.

Higher Education Research Institute. (2017). *Diverse Learning Environments survey*. Retrieved from https://heri.ucla.edu/diverse-learning-environments-survey/

Higher Learning Commission. (2017). *Criteria for accreditation*. Retrieved from https://www.hlcommission.org/Policies/criteria-and-core-components.html

Hofer, B.K. (2008). Personal epistemology and culture. In M.S. Khine (Ed.), *Knowing, knowledge and beliefs: Epistemic studies across diverse cultures* (pp. 3-24). Amsterdam, Netherlands: Springer.

Hofer, B.K., & Pintrich, P.R. (1997). The development of epistemological theories: Beliefs about knowledge and knowing and their relation to learning. *Review of Educational Research, 67*(1), 88-140.

Hogg, M.A., Terry, D.J., & White, K.M. (1995). A tale of two theories: A critical comparison of identity theory with social identity theory. *Social Psychology Quarterly, 58*(4), 255-269.

Hollinger, D.A. (2003). Amalgamation and hypodescent: The question of ethnoracial mixture in the history of the United States. *American Historical Review, 108*(5), 1363-1390.

Howard-Hamilton, M. (2000). Programming for multicultural competencies. *New Directions for Student Services, 2000*(90), 67-78.

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

Hubley, A.M., & Zumbo, B.D. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research, 103*, 219-230.

Hui, C.H., & Triandis, H.C. (1989). Effects of culture and response format on extreme response style. *Journal of Cross-cultural Psychology, 20*(3), 296-309.

Hunter, B., White, G.P., & Godbey, G.C. (2006). What does it mean to be globally competent? *Journal of Studies in International Education, 10*(3), 267-285.

Hurtado, A. (1996). Strategic suspensions: Feminists of color theorize the production of knowledge. In N. Goldberger, J. Tarule, B. Clinchy, & M. Belenky (Eds.), *Knowledge, difference, and power: Essays inspired by women's ways of knowing* (pp. 372-392). New York, NY: Basic Books.

Hurtado, S. (1994). The institutional climate for talented Latino students. *Research in Higher Education, 35*, 539-569.

Hurtado, S. (2003). *Preparing college students for a diverse democracy*. Ann Arbor, MI: University of Michigan, Center for the Study of Higher and Postsecondary Education.

Hurtado, S., & Carter, D. (1997). Effects of college transition and perceptions of the campus racial climate on Latino college students' sense of belonging. *Sociology of Education, 70*, 324-345.

Hurtado, S., Carter, D. F., & Spuler, A. (1996). Latino student transition to college: Understanding racial and ethnic differences. *The Journal of Higher Education, 72*, 265-286.

Hurtado, S., Milem, J.F., Clayton-Pedersen, A., & Allen, W.R. (1998). Enhancing campus climates for racial/ethnic diversity: Educational policy and practice. *The Review of Higher Education, 21*, 279-302.

Ibarra, H. (1993). Personal networks of women and minorities in management: A conceptual framework. *Academy of Management Review, 18*, 56-87.

Institute of International Education. (2012). *Open doors 2012 report on international educational exchange*. Washington, DC: Author.

Institute of International Education. (2016). *Open doors 2016 executive summary.* Retrieved from https://www.iie.org/Why-IIE/Announcements/2016-11-14-Open-Doors-Executive-Summary

International Association of Universities. (2014). *Internationalization of higher education: Growing expectations, fundamental values*. Retrieved from http://www.iau-aiu.net/sites/all/files/IAU-4th-GLOBAL-SURVEY-EXECUTIVE-SUMMARY.pdf

Iowa State University. (2015). *Global Perspective Inventory General, New Student, and Study Abroad Forms*. Ames, IA: Research Institute for Studies in Education, Iowa State University.

Irizarry, Y. (2015). Utilizing multidimensional measures of race in educational research: The case of teacher perceptions. *Sociology of Race and Ethnicity, 1*(4), 564-583.

Jay, G.M., & D'augelli, A.R. (1991). Social support and adjustment to university life: A comparison of African American and White freshmen. *Journal of Community Psychology, 19*(2), 95-108.

Jiménez, T.R., & Horowitz, A.L. (2013). When white is just alright: How immigrants redefine achievement and reconfigure the ethnoracial hierarchy. *American Sociological Review, 78*(5), 849-871.

Johnson, T.P., Cho, Y.I., Holbrook, A.L., O'Rourke, D., Warnecke, R.B., & Chavez, N. (2006). Cultural variability in the effects of question design features on respondent comprehension of health surveys. *Annals of Epidemiology, 16*(9), 661-668.

Johnston, M.P., Ozaki, C.C., Pizzolato, J.E., & Chaudhari, P. (2014). Which box(es) do I check? Investigating college students' meanings behind racial identification. *Journal of Student Affairs Research and Practice, 51*(1), 56-68.

Jöreskog, K.G., & Sörbom, D. (1996). *Prelis 2: User's reference guide: A program for multivariate data screening and data summarization.* Chicago, IL: Scientific Software.

Jorgensen, T.D., Kite, B.A., Chen, P., & Short, S.D. (2017). Permutation randomization methods for testing measurement equivalence and detecting differential item functioning in multiple-group confirmatory factor analysis. *Psychological Methods*, Advanced online publication.

Judd, C.M., Drake, R.A., Downing, J.W., & Krosnick, J.A. (1991). Some dynamic properties of attitude structures: Context-induced response facilitation and polarization. *Journal of Personality and Social Psychology, 60*(2), 193-202.

Kane, M.T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*, 319-342.

Kane, M.T. (2013). The argument-based approach to validation. *School Psychology Review, 42*(4), 448-457.

Kegan, R. (1994). *In over our heads: The mental demands of modern life*. Cambridge, MA: Harvard University Press.

Kelley, C., & Myers, J. (1995). *CCAI Cross-cultural Adaptability Inventory*. Minneapolis, MN: National Computer Systems.

Kezar, A.J., Chambers, A.C., & Burkhardt, J.C. (Eds.). (2004). *Higher education for the public good: Emerging voices from a national movement.* San Francisco, CA: Jossey-Bass.

Kim, B.S.K., Atkinson, D.R., & Yang, P.H. (1999). The Asian Values Scale: Development, factor analysis, validation, and reliability. *Journal of Counseling Psychology, 46*, 342-352.

King, G., Murray, C.J.L., Salomon, J.A., & Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review, 98*(1), 191-207.

King, P.M., & Baxter Magolda, M.B. (2005). A developmental model of intercultural maturity. *Journal of College Student Development, 46*(6), 571-592.

King, P.M., Perez, R.J., & Shim, W. (2013). How college students experience intercultural learning: Key features and approaches. *Journal of Diversity in Higher Education, 6*(2), 69-83.

Kirkhart, K.E. (1995). Seeking multicultural validity: A postcard from the road. *Evaluation Practice, 16*(1), 1-12.

Kleinman, A. (1988). *The illness narratives: Suffering, healing, and the human condition*. New York, NY: Basic Books.

Knight, R.G., Chisholm, B.J., Marsh, N.V., & Godfrey, H.P. (1988). Some normative, reliability, and factor analytic data for the revised UCLA Loneliness Scale. *Journal of Clinical Psychology, 44*, 203-206.

Koh, K.H., & Zumbo, B.D. (2008). Multi-group confirmatory factor analysis for testing measurement invariance in mixed item format data. *Journal of Modern Applied Statistical Methods, 7*(2), 471-477.

Kozai Group. (2010). *Global Competencies Inventory*. Chesterfield, MO: Author.

Kuh, G.D., & Ikenberry, S. (2009). *More than you think, less than we need: Learning outcomes assessment in American higher education*. Champaign, IL: National Institute for Learning Outcomes Assessment.

Kuhn, D., Cheney, R., &Weinstock, M. (2000). The development of epistemological understanding. *Cognitive Development, 15*, 309-328.

LaNasa, S.M., Cabrera, A.F., & Trangsrud, H. (2009). The construct validity of student engagement: A confirmatory factor analysis approach. *Research in Higher Education, 50*, 315-332.

Lee, J.J. (2010). International students' experiences and attitudes at a U.S. host institution: Self-reports and future recommendations. *Journal of Research in International Education, 9*(1), 66-84.

Leung, K., Ang, S., & Tan, M.L. (2014). Intercultural competence. *Annual Review of Organizational Psychology and Organizational Behavior, 1,* 489-519.

Li, C.H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods, 48*(3), 936-949.

Lum, D. (1999). *Culturally competent practice: A framework for growth and action*. Pacific Grove, CA: Brooks/Cole.

Madden, M. (2005). Gender and leadership in higher education. *Psychology of Women Quarterly, 29*(1), 3-14.

Major B., & O'Brien L.T. (2005). The social psychology of stigma. *Annual Review of Psychology, 56*, 393-421.

Marsh, H.W. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In R.P. Perry & J.C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 319-383). Dordrecht, The Netherlands: Springer.

McIntosh, P. (2003). White privilege: Unpacking the invisible knapsack. In S. Plous (Ed.), *Understanding prejudice and discrimination* (pp. 191-196). New York, NY: McGraw-Hill.

McLean, K.C. (2008). The emergence of narrative identity. *Social and Personality Psychology Compass, 2*(4), 1685-1702.

McPherson, M., Smith-Lovin, L., & Cook, J.M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology, 27*, 415-444.

Meade, A.W., Johnson, E.C., & Braddy, P.W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology, 93*(3), 568-592.

Melnick, S.A., & Gable, R.K. (1990). The use of negative item stems: A cautionary note. *Educational Research Quarterly, 14*(3), 31-36.

Mendez-Russell, A., Wilderson, Jr., F., & Tolbert, A.S. (2003). *Discovering diversity profile*. Buffalo, NY: RV Rhodes.

Mendoza-Denton R., Downey, G., Purdie, V., Davis, A., & Pietrzak, J. (2002). Sensitivity to status-based rejection: Implications for African American students' college experience. *Journal of Personality and Social Psychology, 83*, 896-918.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Milem, J.F., Chang, M.J., Antonio, A.L. (2005). *Making diversity work on campus: A research-based perspective.* Washington, DC: Association of American Colleges and Universities.

Molix, L., & Bettencourt, B.A. (2010). Predicting well-being among ethnic minorities: Psychological empowerment and group identity. *Journal of Applied Social Psychology, 40*(3), 513-533.

Mollica, K.A., Gray, B., & Treviño, L.K. (2003). Racial homophily and its persistence in newcomers' social networks. *Organization Science, 14*(2), 123-136.

Montenegro, E., & Jankowski, N.A. (2017). *Equity and assessment: Moving towards culturally responsive assessment* (Occasional Paper No. 29). Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment (NILOA).

Mummendey, A., Klink, A., & Brown, R. (2001). Nationalism and patriotism: National identification and out-group rejection. *British Journal of Social Psychology, 40*, 159-172.

Muthén, B.O. (1993). Goodness of fit with categorical and other nonnormal variables. In K.A. Bollen, & J.S. Long (Eds.), *Testing structural equation models* (pp. 205-234). Newbury Park, CA: Sage Publishing.

Muthén, L.K. & Muthén, B.O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 4*, 599-620.

Myers, N.D., Ahn, S., & Jin, Y. (2011). Sample size and power estimates for a confirmatory factor analytic model in exercise and sport: A Monte Carlo approach. *Research Quarterly for Exercise and Sport, 82(*3), 412-423.

National Center for Public Policy and Higher Education. (2008). *Partnerships for public purposes: Engaging higher education in societal changes of the 21st century*. Retrieved from http://www.highereducation.org/reports/wegner/wegner.pdf

National Survey of Student Engagement. (2010). *Factor analysis: 2009 internal structure for deep learning*. Retrieved from http://nsse.indiana.edu/pdf/psychometric_portfolio/Validity_DeepLearning.pdf

National Survey of Student Engagement. (2017). *NSSE accreditation toolkits*. Retrieved from http://nsse.indiana.edu/html/accred_toolkits.cfm

National Survey of Student Engagement. (2018). *NSSE survey instrument*. Retrieved from http://nsse.indiana.edu/html/survey_instruments.cfm

Nelson Laird, T.F., Shoup, R., & Kuh, G.D. (2005). *Measuring deep approaches to learning using the National Survey of Student Engagement.* Paper presented at the Annual Meeting of the Association for Institutional Research. Chicago, IL.

Neuliep, J.W., Chaudoir, M., & McCroskey, J.C. (2001). A cross-cultural comparison of ethnocentrism among Japanese and United States college students. *Communication Research Reports, 18*(2), 137-146.

Newton, P.E., & Shaw, S.D. (2013). *Validity in educational and psychological assessment*. Thousand Oaks, CA: Sage Publications.

Nezlek J.B. (2007). *Naturally occurring interethnic contact: Blacks and Whites in the US*. Paper presented at the Annual Meeting of the Society of Experimental Social Psychology. Chicago, IL.

Nora, A., & Cabrera, A.F. (1996). The role of perceptions of prejudice and discrimination on the adjustment of minority students to college. *Journal of Higher Education, 67*(2), 119-148.

Ogbu, J.U., & Simons, H.D. (1998). Voluntary and involuntary minorities: A cultural-ecological theory of school performance with some implications for education. *Anthropology and Education Quarterly, 29*(2), 155-188.

Olcott, D. (2009). Global connections to global partnerships: Navigating the changing landscape of internationalism and cross-border higher education. *The Journal of Continuing Higher Education, 57*, 1-9.

Paige, R.M., Jacobs-Cassuto, M., Yershova, Y.A., & DeJaeghere, J. (2003). Assessing intercultural sensitivity: An empirical analysis of the Hammer and Bennett Intercultural Development Inventory. *International Journal of Intercultural Relations, 27*(4), 467-486.

Paulhus, D.L. (2002). Socially desirable responding: The evolution of a construct. In H.I. Braun, D.N. Jackson, & D.E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 49-69). Mahwah, NJ: Erlbaum.

Peng, S. (2006). A comparative perspective of intercultural sensitivity between college students and multinational employees in China. *Multicultural Perspectives, 8*(3), 38-45.

Peng, S, Rangsipaht, S., & Thaipakdee, S. (2005). Measuring intercultural sensitivity: A comparative study of ethnic Chinese and Thai nationals. *Journal of Intercultural Communication Research, 34*(2), 119-137.

Phelen, P., Yu, H.C., & Davidson, A.L. (1994). Navigating the psychosocial pressures of adolescence: The voices and experiences of high school youth. *American Educational Research Journal, 31*, 415-447.

Phinney, J.S. (1989). Stages of ethnic identity development in minority group adolescents. *Journal of Early Adolescence, 9*, 34-49.

Phinney, J.S. (1990). Ethnic identity in adolescents and adults: Review of research. *Psychological Bulletin, 108*(3), 499-514.

Phinney, J.S., & Alipuria, L.L. (1990). Ethnic identity in college students from four ethnic groups. *Journal of Adolescence, 13*(2), 171-183.

Pike, G.R. (1993). The relationship between perceived learning and satisfaction with college: An alternative view. *Research in Higher Education, 34*, 23-40.

Pike, G.R. (2013). NSSE benchmarks and institutional outcomes: A note on the importance of considering the intended uses of a measure in validity studies. *Research in Higher Education, 54*, 149-170.

Pilotte, W.J., & Gable, R.K. (1990). The impact of positive and negative item stems on the validity of a computer anxiety scale. *Educational and Psychological Measurement, 50*, 603-610.

Pizzolato, J.E., & Chaudhari, P. (2009, April). *Complicating assessment: Considerations for quantitative measurement of self-authorship.* Paper presented at the American Educational Research Association National Conference. San Diego, CA.

Pizzolato, J.E., Nguyen, T.K., Johnston, M.P., & Wang, S. (2012). Understanding context: Cultural, relational, and psychological interactions in self-authorship development. *Journal of College Student Development, 53*(5), 656-679.

Plant E.A., & Devine P.G. (1998). Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology, 75*, 811-832.

Portalla, T., & Chen, G. (2010). The development and validation of the intercultural effectiveness scale. *Intercultural Communication Studies, 19*(3), 21-37.

Porter, S.R. (2011). Do college student surveys have any validity? *The Review of Higher Education, 35*(1), 45-76.

Rankin, S.R., & Reason, R.D. (2005). Differing perceptions: How students of color and white students perceive campus climate for underrepresented groups. *Journal of College Student Development, 46*(1), 43-61.

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement, 21*(2), 173-184.

Research Institute for Studies in Education. (2017a). *Global Perspective Inventory: Theoretical foundations and scale descriptions*. Ames: Iowa State University.

Research Institute for Studies in Education. (2017b). *Global Perspective Inventory: Scales and component items*. Ames: Iowa State University.

Research Institute for Studies in Education. (2017c). *GPI information*. Retrieved from http://www.gpi.hs.iastate.edu/information.php

Research Institute for Studies in Education. (2017d). *Factor structure of the GPI New Student Form*. Unpublished manuscript, Research Institute for Studies in Education, Iowa State University.

Richeson, J.A., & Shelton, J.N. (2007). Negotiating interracial interactions: Costs, consequences, and possibilities. *Current Directions in Psychological Science, 16*, 316-320.

Riveras, J., & Harrison, M.J. (2016). Restructuring a globalization model to reflect changing dynamics. *The Business and Management Review, 7*(5), 372-378.

Roberts, D.C., & Komives, S.R. (2016). Internationalizing student learning and development. *New Directions for Higher Education, 2016*, 9-21.

Rockquemore, K.A., Brunsma, D.L., & Delgado, D.J. (2009). Racing to theory or retheorizing race? Understanding the struggle to build a multiracial identity theory. *Journal of Social Issues, 65*(1), 13-34.

Roth, L.M. (2004). The social psychology of tokenism: Status and homophily processes on Wall Street. *Sociological Perspectives, 47*(2), 189-214.

Ruben, B.D. (1976). Assessing communication competency for intercultural adaptation. *Group and Organization Studies, 1*, 334-354.

Rudnev, M., Lytkina, E, Davidov, E., Schmidt, P., & Zick, A. (2018). Testing measurement invariance for a second-order factor: A cross national test of the Alienation Scale. *Methods, Data, Analyses, 12*(1), 47-76.

Saenz, V.B., Ngai, H.N., & Hurtado, S. (2007). Factors influencing positive interactions across race for African American, Asian American, Latino, and White college students. *Research in Higher Education, 48*(1), 1-38.

Saguy,T., Dovidio, J.F., & Pratto, F. (2008). Beyond contact: Intergroup contact in the context of power relations. *Personality and Social Psychology Bulletin, 34*, 432-445.

Sanchez-Hucles, J., & Sanchez, P. J. (2007). From margin to center: The voices of diverse feminist leaders. In J.L. Chin, B. Lott, J.K. Rice, & J. Sanchez-Hucles (Eds.), *Women and leadership: Transforming visions and diverse voices* (pp. 211-227). Malden, MA: Blackwell.

Sass, D.A. (2011). Testing measurement invariance and comparing latent factor means within a confirmatory factor analysis framework. *Journal of Psychoeducational Assessment, 29*(4), 347-363.

Sass, D.A., Schmitt, T.A., & Marsh, H.W. (2014). Evaluating model fit with ordered categorical data within a measurement invariance framework: A comparison of estimator. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(2), 167-180.

Satorra, A. (2000). Scaled and adjusted restricted tests in multisample analysis of moment structures. In D.D.H. Heijmails, D.S.G. Pollock, & A. Satorra (Eds.), *Innovations in multivariate statistical analysis: A festschrift for Heinz Neudecker* (pp. 233-247). Dordrecht, Netherlands: Kluwer Academic.

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

Satorra, A., & Bentler, P.M. (2010). Ensuring positiveness of scaled difference chi-square test statistic. *Psychometrika, 75*(2), 243-248.

Scheepers, D., Spears, R., Doosje, B., & Manstead, A.S.R. (2006). Diversity in in-group bias: Structural factors, situational features, and social functions. *Journal of Personality and Social Psychology, 90*(6), 944-960.

Schnabel, D.B.L., & Kelava, A. (July, 2013). *Development and validation of the German Test to Measure Intercultural Competence (TMIC): A combined-method approach.* Paper presented at the 12th European Conference on Psychological Assessment, San Sebastian, Spain.

Schnabel, D.B.L., Kelava, A., Seifert, L., & Kulbrodt, B. (2015) Development and validation of a job-related multimethod Test to Measure Intercultural Competence. *Diagnostica, 61*, 3-21.

Schnabel, D.B.L., Kelava, A., van de Vijver, F.J.R., & Seifert, L. (2015). Examining psychometric properties, measurement invariance, and construct validity of a short version of the Test to Measure Intercultural Competence (TMIC-S) in Germany and Brazil. *International Journal of Intercultural Relations, 49*, 137-155.

Schommer-Aikens, M. (2004). Explaining the epistemological belief system: Introducing the embedded systemic model and coordinated research approach. *Educational Psychologist, 39*(1), 19-29.

Schriesheim, C.A., & Hill, K.D. (1981). Controlling acquiescence response bias by item reversals: The effect on questionnaire validity. *Educational and Psychological Measurement, 41*, 1101-1114.

Schwartz, S.H. (2006). A theory of cultural value orientations: Explication and applications. *Comparative Sociology, 5*, 136-182.

Schwartz, S.J., Zamboanga, B.L., Meca, A., & Ritchie, R.A. (2012). Identity around the world: An overview. *New Directions for Child and Adolescent Development, 2012*(138), 1-18.

Sedlacek, W.E. (1999). Black students on White campuses: 20 years of research. *Journal of College Student Development, 40*(5), 538-550.

Sedlacek, W.E. (2003). Alternative admissions and scholarship selection measures in higher education. *Measurement and Evaluation in Counseling and Development, 35*, 263-272.

Shapiro, D.N., Rios, D., & Stewart, A.J. (2010). Conceptualizing lesbian sexual identity development: Narrative accounts of socializing structures and individual decisions and actions. *Feminism and Psychology, 20*(4), 491-510.

Sharkness, J., DeAngelo, L., & Pryor, J. (2010). *CIRP construct technical report*. Los Angeles, CA: University of California, Los Angeles Higher Education Research Institute.

Shweder, R.A., Goodnow, J.J., Hatano, G., LeVine, R.A., Markus, H.R. and Miller, P.J. (2007). The cultural psychology of development: One mind, many mentalities. In W. Damon (Ed.), *Handbook of child psychology: Volume 1* (pp. 865-937). New York, NY: John Wiley and Sons.

Simon, B., & Brown, R. (1987). Perceived intragroup homogeneity in minority-majority contexts. *Journal of Personality and Social Psychology, 53*(4), 703-711.

Simon, B., Kulla, C., & Zobel, M. (1995). On being more than just a part of the whole: Regional identity and social distinctiveness. *European Journal of Social Psychology, 25*(3), 325-340.

Sireci, S.G. (2007). On validity theory and test validation. *Educational Researcher, 36*(8), 477-481.

Smith, D.G. (2009). *Diversity's promise for higher education: Making it work*. Baltimore, MD: Johns Hopkins University Press.

Sobania, N.W. (Ed.). (2015). *Putting the local in global education: Models for transformative learning through domestic off-campus programs*. Sterling, VA: Stylus.

Sobania, N.W., & Braskamp, L.A. (2009). Study abroad or study away: It's not merely semantics. *Peer Review, 11*(4), 23-26.

Solórzano, D., Villalpando, O. (1998). Critical race theory, marginality, and the experience of minority students in higher education. In C. Torres & T. Mitchell (Eds.), *Emerging issues in the sociology of education: Comparative perspectives* (pp. 211-224). Albany: State University of New York Press.

Solórzano, D., Ceja, M., & Yosso, T. (2000). Critical race theory, racial microaggressions, and campus racial climate: The experiences of African American college students. *Journal of Negro Education, 69*(1), 60-73.

Sörbom, D., & Jöreskog, K. (1989). *LISREL 8: User's reference guide*. Lincolnwood, IL: Scientific Software International, Inc.

Sorensen, N., Nagda, B.A., Gurin, P., & Maxwell, K.E. (2009). Taking a "hands-on" approach to diversity in higher education: A critical-dialogic model for effective intergroup interaction. *Analyses of Social Issues and Public Policy, 9*(1), 3-35.

Spanierman, L.B., Neville, H.A., Liao, H.Y., Hammer, J.H., Wang, Y.F. (2008). Participation in formal and informal campus diversity experiences: Effects on students' racial democratic beliefs. *Journal of Diversity in Higher Education, 1*(2), 108-125.

Spencer, M.B., & Dornbush, S.M. (1990). Challenges in studying ethnic minority youth. In Feldman, S.S. & Elliot, G.R. (Eds.), *At the threshold: The developing adolescent* (pp. 123-146). Cambridge, MA: Harvard University Press.

Stage, F.K. (2007). Answering critical questions using quantitative data. In F.K. Stage (Ed.), *New directions for institutional research: Using quantitative data to answer critical questions* (pp. 5-16). San Francisco, CA: Jossey-Bass.

Steenkamp, J., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research, 25*(1), 78-90.

Steinmetz, H., Schmidt, P., Tina-Booh, A., Wieczorek, S., & Schwartz, S.H. (2009). Testing measurement invariance using multigroup CFA: Differences between educational groups in human values measurement. *Quality and Quantity: International Journal of Methodology, 43*, 599-616.

Stephan W.G., & Stephan, C.W. (2001). *Improving intergroup relations*. Thousand Oaks, CA: Sage Publications.

Stevens, M., Bird, A., Mendenhall, M.E., & Oddou, G. (2014). Measuring global leader intercultural competency: Development and validation of the Global Competencies Inventory (GCI). In J.S. Osland, M. Li, & Y. Wang (Eds.), *Advances in global leadership* (pp.115-154). Bingley, UK: Emerald Group Publishing Limited.

Stier, J. (2006). Internationalisation, intercultural communication and intercultural competence. *Journal of International Communication, 11*, 1-11.

Stinson, K.M. (2007). *Diversity awareness profile: Facilitator's guide*. Hoboken, NJ: John Wiley & Sons, Inc.

Sue, D.W. (2004). Whiteness and monocultural ethnocentrism: Making the "invisible" visible. *American Psychologist, 59*(8), 761-769.

Sue, D.W., & Sue, D. (1990). *Barriers to effective cross-cultural counseling*. New York, NY: Wiley.

Swain,S.D., Weathers, D., & Niedrich, R.W. (2008). Assessing three sources of misresponse to reversed Likert items. *Journal of Marketing Research, 45*(1), 116-131.

Tajfel, H. (Ed.) (1978). *Differentiation between social groups: Studies in the social psychology of intergroup relations*. London, UK: Academic Press.

Tamam, E. (2010). Examining Chen and Starosta's model of intercultural sensitivity in a multiracial collectivist country. *Journal of Intercultural Communication Research, 39*(3), 173-183.

Tanaka, G. (2002). Higher education's self-reflexive turn: Toward an intercultural theory of student development. *The Journal of Higher Education*, *73*(2), 263-296.

Tasaki, K. (2001). Culture and epistemology: An investigation of different patterns in epistemological beliefs across cultures. *Dissertation Abstracts International Section A: Humanities and Social Sciences, 62*(2-A), 463.

Tatum, B.D. (2007). *Can we talk about race? And other conversations in an era of school resegregation*. Boston, MA: Beacon Press.

Thomas, D.A., & Gabarro, J.J. (1999). *Breaking through: The making of minority executives in corporate America.* Boston, MA: Harvard Business School Press.

Thomson, A.M., Smith-Tolken, A.R., Naidoo, A.V., & Bringle, R.G. (2011). Service-learning and community engagement: A comparison of three national contexts. *Voluntas, 22*(2), 214-237.

Tourangeau, R., Rips, L., & Rasinski, K. (2000). *The psychology of survey response*. New York, NY: Cambridge University Press.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133*(5), 859-883.

Trawalter, S., & Richeson, J.A. (2008). Let's talk about race, baby! When whites' and blacks' interracial contact experiences diverge. *Journal of Experimental and Social Psychology, 44*(4), 1214-1217.

Tropp, L.R. (2003). The psychological impact of prejudice: Implications for intergroup contact. *Group Processes and Intergroup Relations, 6*, 131-149.

Tucker, K.L., Ozer, D.J., Lyubomirsky, S., & Boehm, J.K. (2006). Testing for measurement invariance in the satisfaction with life scale: A comparison of Russians and North Americans. *Social Indicators Research, 78*(2), 341-360.

United States Department of Education. (2015a). *Total fall enrollment in degree-granting postsecondary institutions, by level of enrollment, sex, attendance status, and race/ ethnicity of student: Selected years, 1976 through 2014* [Data file]. Retrieved from http://nces.ed.gov/programs/digest/d15/tables/dt15_306.10.asp?current=yes

United States Department of Education. (2015b). *Percentage of 18- to 24-year-olds enrolled in degree-granting postsecondary institutions, by level of institution and sex and race/ ethnicity of student: 1970 through 2014* [Data file]. Retrieved from http://nces.ed.gov/programs/digest/d15/tables/dt15_302.60.asp?current=yes

van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology, 9*(4), 486-492.

van der Zee, K.I., & Brinkmann, U. (2004). Construct validity evidence for the Intercultural Readiness Check against the Multicultural Personality Questionnaire. *International Journal of Selection and Assessment, 12*(3), 285-290.

Van Sonderen, E., Sanderman, R., & Coyne, J.C. (2013). Ineffectiveness of reverse wording of questionnaire items: Let's learn from cows in the rain. *PloS One*, *8*(7), e68967.

van Vaerenbergh, Y., & Thomas, T.D. (2013). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research, 25*(2), 195-217.

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.

Warnecke, R.B., Johnson, T.I., Chávez, N., Sudman, S., O'Rourke, D.P., Lacey, L., & Horm, J. (1997). Improving question wording in surveys of culturally diverse populations. *Annals of Epidemiology, 7*, 334-342.

Weijters, B. (2006). *Response styles in consumer research*. (Doctoral dissertation). Ghent University, Ghent, Belgium.

Weijters, B., Schillewaert, N., & Geuens, M. (2008). Assessing response styles across modes of data collection. *Journal of the Academy of Marketing Science, 36*, 409-422.

Wells, A.S., Holme, J.J., Revilla, A.T., Atanda, A.K. (2009). *Both sides now: The story of school desegregation's graduates*. Berkeley, CA: University of California Press.

Whetzel, D.L., McDaniel, M.A., & Nguyen, N.T. (2008). Subgroup differences in situational judgment test performance: A meta-analysis. *Human Performance, 21*, 291-309.

Whitehead, D.M. (2015). Global learning: Key to making excellence inclusive. *Liberal Education, 101*(3), 6-13.

Whitehead, D.M. (2016). *Essential global learning*. Washington, DC: Association of American Colleges and Universities.

Wolf, E.J., Harrington, K.M., Clark, S.L., & Miller, M.W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement, 76*(6), 913-934.

Wong, B., & Chai, C.S. (2010). Asian personal epistemologies and beyond: Overview and some reflections. *The Asia-Pacific Education Researcher, 19*(1), 1-6.

Worthen, M. (2018, February 23). The misguided drive to measure 'learning outcomes'. *The New York Times*. Retrieved from https://www.nytimes.com/2018/02/23/opinion/sunday/colleges-measure-learning-outcomes.html

Yan, M.C., & Wong, Y.R. (2005). Rethinking self-awareness in cultural competence: Toward a dialogic self in cross-cultural social work. *Families in Society, 86*(2), 181-188.

Yoon, M., & Lai, M.H.C. (2018). Testing factorial invariance with unbalanced samples. *Structural Equation Modeling: A Multidisciplinary Journal, 25*(2), 201-213.

Youn, I. (2000). The culture specificity of epistemological beliefs about learning. *Asian Journal of Social Psychology, 3*(1), 87-105.

Yue, J., Creamer, E.G., & Wolfe, E. (2009, April). *Measurement of self-authorship: A validity study using multidimensional random coefficients multinomial logit model*. Paper presented at the American Educational Research Association National Conference, San Diego, CA.

Zamanzadeh, V., Ghahramanian, A., Rassouli, M., Abbaszadeh, A., Alavi-Majd, H., & Nikanfar, A. (2015). Design and implementation content validity study: Development of an instrument for measuring patient-centered communication. *Journal of Caring Sciences*, *4*(2), 165-178.

Zane, N.W.S., Sue, S., Hu, L., & Kwon, J.H. (1991). Asian American assertion: A social learning analysis of cultural differences. *Journal of Counseling Psychology, 38*, 63-70.

Zhang, X., & Savalei, V. (2015). Improving the factor structure of psychological scales: The expanded format as an alternative to the Likert scale format. *Educational and Psychological Measurement, 76*(3), 357-386.

VITA

Lisa Davidson was born in Cleveland, Ohio and lived in Kent, Ohio until moving to Columbus, Indiana at the age of 12. She traveled back to Ohio to attend Cleveland State University, earning a Bachelor of Arts in psychology with a minor in biology in 2000. She moved to Chicago, Illinois in 2002 and attended DePaul University, earning a Master of Education in counseling and college student development in 2006.

Davidson began her higher education career in 2001 as the Assistant Director of Undergraduate Admission at Cleveland State University. In 2002, she began an 11-year period at DePaul University, where she first worked as the Assistant Director for First-year Students in the College of Liberal Arts and Sciences, then as an Assistant Director and Career Counselor in the Career Center, and—for the last six years there—as the founding Director of the Office for Academic Advising Support. Davidson served as an adjunct faculty member in DePaul University's Liberal Studies Program and its Department of Counseling and Special Education from 2010 – 2013. She has also served as an instructor in Loyola University Chicago's Higher Education program since 2016. Davidson worked as a research assistant alongside Dr. Mark Engberg at Loyola University from 2013 – 2016 and spent a year working as a survey data analyst in Waubonsee Community College's Office of Institutional Effectiveness from 2016 – 2017. Since 2017, Davidson has worked for Interfaith Youth Core's assessment and research team and will begin a new role as the Assessment and Research Manager in May 2018. She currently resides in Chicago, Illinois.