Differential Item Functioning and Measurement Invariance of Self- and Proxy-Reports: An Evaluation of Objective Quality of Life Measures for People with Intellectual and Developmental Disabilities

A DISSERTATION
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Renee A. Hepperlen

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Dr. Elizabeth Lightfoot, Adviser

January 2015

Acknowledgements

I am forever grateful for the support and encouragement I received from many individuals throughout this process. My adviser, Dr. Elizabeth Lightfoot provided guidance, reassurance, and the freedom to explore my research interests. I would also like to thank my dissertation committee. Dr. Mark Davison's guidance with the statistical analyses was invaluable. Dr. Derek Nord's help in securing this secondary dataset and knowledge of the National Core Indicators were essential to this dissertation. I would also like to thank Dr. Colleen Fisher and Dr. James Reinardy for their thoughtful and thorough advice.

I appreciate the financial support I received during my tenure at the University of Minnesota. The research assistant positions I held with Dr. Lightfoot, Dr. Nord and Dr. Piescher helped me hone my research interests and skills. In addition, the traineeship through the University of Minnesota's Leadership Education in Neurodevelopmental Disabilities (LEND) allowed me to expand my interests in disability studies.

I am also grateful to have gone through this process with a wonderful cohort of PhD students. Your support and encouragement through the entire program was a tremendous support to me.

I am also very thankful to my family for their love and continuous encouragement. My parents, Phil and Phyllis, were always encouraging and empathetic. Finally, I owe my greatest gratitude to my husband, Michael, and my children for their ongoing support and understanding through many late nights and weekend distractions.

Dedication

To my family – Michael, Thomas, and Anna Marie – for their love, support, and encouragement throughout this educational journey.

Abstract

The field of intellectual disabilities and developmental disabilities (ID/DD) uses objective quality of life indicators for policy and program development (Verdugo, Schalock, Keith, & Stancliffe, 2005). An ongoing concern in this field is the assessment of quality of life for people who are unable to answer for themselves. In these instances, a proxy-respondent, or someone who knows the person with ID/DD well, will respond on his/her behalf. Research examining the efficacy of using proxy-respondents has yielded mixed results. While some studies failed to show statistically significant differences in responses (McVilly, Burton-Smith, & Davidson, 2000; Rapley, Ridgway, & Beyer, 1998; Stancliffe, 1999), other research has found meaningful differences between matched pairs of self- and proxy-respondents (Rapley et al., 1998). A principle limitation of these previous studies is the reliance on simplistic analytic methods, such as a *t*-test or correlation to determine if similarities existed between these matched groups. Methodologically, the previous studies on self- and proxy-respondents used *t*-tests and correlations to examine the relationship between self- and proxy-responses. The present study extends this body of research through the use of differential item functioning and measurement invariance to examine the use of self- and proxy-respondents. Specifically, this study examined the internal structure of the three objective quality of life measures on the National Core Indicators, including the *Community Inclusion*, *Life Decisions*, and *Everyday Choices* scales. Study findings revealed that several items function differently for these two groups when comparing these respondents based on the total score of the scale, which implies that construct-irrelevant differences impacted some item responses (American Educational Research Association [AERA], American Psychological

Association [APA], National Council on Measurement in Education [NCME], 1999). In addition, an examination of measurement invariance established that metric invariance fits these data well, meaning that it is not possible to compare these two groups. These findings have policy-and program-evaluation implications, since construct irrelevance (AERA, APA, & NCME, 1999) indicates that for the items identified as functioning differently for these groups, responses also include another construct that is separate from the construct that the scale intends to measure. With these differences, it becomes more difficult to conclude that changes in outcome can be attributed to program social justice implications, since differential item functioning and measurement invariance assessments relate to fairness in testing (Huggins, 2013). When items function differently for groups, this means that respondents find these items difficult, which makes full participation challenging. When individuals find items confusing or hard, then responses may not accurately reflect their experiences. These findings have implications for policy and practice, since policy makers and practitioners use these scales to make program decisions for people with ID/DD.

Table of Contents

List of Tables

Chapter I: Introduction

Researchers estimate that 14.9 individuals per thousand people in the United States have intellectual and developmental disabilities (ID/DD) (Larson et al., 2001). Over the years, clinicians and researchers have made considerable gains in identifying, supporting, and treating individuals with ID/DD. A primary working principle has been and is the consideration of quality of life in individuals with ID/DD. Measuring the quality of life for persons with ID/DD can impact policy and program decisions for them.

With the deinstitutionalization of people with ID/DD, researchers have developed various methods to evaluate quality of life in individuals with ID/DD. However, a key methodological limitation has been the assessment of quality of life with people who have significant impairment in their language skills, which makes completion of self-report measures difficult, if not impossible. To address this, researchers accept proxy or surrogate responses as a means to assess quality of life.

This dissertation relies on existing definitions of intellectual disabilities, developmental disabilities and proxies. The American Association of Intellectual and Developmental Disabilities (AAIDD, 2013), a primary research organization in this field, defined intellectual disabilities (ID) as "characterized by significant limitations both in intellectual functioning and in adaptive behavior as expressed in conceptual, social, and practical adaptive skills. This disability originates before age 18" (Schalock et al., 2007, p. 118). In contrast, the Developmental Disability Act of 2000 defined DD as "a severe, chronic disability…that is attributable to a mental or physical impairment…[beginning] before age 22, is likely to continue indefinitely, [and] results in functional limitations in 3 or more areas" (section, 102, 8, A).

A proxy, an individual who knows the person with ID/DD well, can "provide a satisfactory approximation of the responses the individual [with a disability] would give" (Stancliffe, 1999, p. 185). Nevertheless, researchers recommend relying on a proxy judiciously, to be consulted only when no other reliable source is available (Schalock, 2010; Schalock et al., 2002; Stancliffe, 2000; Verdugo, Schalock, Keith, & Stancliffe, 2005). The use of a proxy presents measurement concerns, since a proxy-response represents another individual's viewpoint, which may or may not be consistent with the perceptions of the individual they represent (Schalock et al., 2002).

The International Association for the Scientific Study of Intellectual Disabilities (IASSID) has developed a structure that outlines quality of life for individuals with ID/DD, consisting of five components. These include: (a) ubiquitous nature of the concept (meaning that everyone regardless of disability status experiences this), (b) fulfillment of needs and wants, (c) incorporation of personal and observable aspects, (d) ability to express personal preferences and control and (e) inclusion of various components (Schalock et al., 2002).

Using several quality of life measures, various researchers have studied the use of self- and proxy-responses to determine if these two groups of individuals respond in equivalent ways. They matched individuals with ID/DD with either family members or paid staff (Perry & Felce, 2002; Rapley, Ridgway, & Beyer, 1998; Stancliffe, 1999). These findings revealed equivocal results. Some studies found no differences between the two respondents (Perry & Felce, 2002; Rapley et al., 1998; Stancliffe, 1999), while others found significant distinctions between these pairings (Rapley et al., 1998).

This paper seeks to further the research of self- and proxy-respondents through the use of differential item functioning and measurement invariance, which represents a methodological departure from previous studies. In a differential item functioning (DIF) analysis, comparisons are made on individual items of a scale or measure between self- and proxy-respondents who have the same overall score on the measure (Dorans & Holland, 1993; Holland & Thayer, 1988; Osterlind & Everson, 2009). This statistical method allows the researcher to examine fairness of items (Huggins, 2013). The second method, measurement invariance, establishes whether scales have the same structure, factor loadings and intercepts for two groups based on the latent construct (Brown, 2006). This analysis assesses the entire scale and determines whether it is possible to make comparisons between groups, such as a *t*-test or ANOVA (Brown, 2006).

The analyses in this dissertation explore the internal structure of the three objective quality of life measures of the National Core Indicators (NCI). The NCI is used to assess the experiences including quality of life of individuals with ID/DD in many states (Agency for Healthcare Research and Quality ([AHRQ], 2010). The overarching research question assesses whether the items and overall three scales of the NCI (*Community Inclusion, Life Decisions* and *Everyday Choices*) function in similar ways when considering self- and proxy-respondents.

Five research inquiries are examined in this dissertation. The first appraises the internal consistency reliabilities for the three objective quality of life scales. Second, it assesses whether there are group differences between self and proxy-respondents for all of the items on the three scales. Third, it considers whether these items function differently depending on who answers the items and uses both the Mantel-Haenszel

method and logistic regression analysis. Fourth, this dissertation seeks to determine if each of the three scales are uni-dimensional for the two respondent types. Finally, it investigates measurement invariance for each of the three scales.

This study found several points of concern with respect to the internal validity of these scales. The first is that for self-reports, the internal consistency reliability was considerably lower than acceptable for the *Community Inclusion* and *Everyday Choices* scales. In addition, self-reports on the *Everyday Choices* scale reported that they had considerable choice on all of the items. Secondly, item impact was detected with meaningful differences for most of the items on all three scales. Third, there were a number of items in which self- and proxy-respondents answered in systematically different ways. Fourth, all of the scales did show a uni-dimensional structure for both self- and proxy-respondents. Finally, all three scales demonstrated metric invariance, which indicates that it is not possible to compare means between these two groups without statistically adjusting for these differences.

Chapter II: Literature Review

Prior to the deinstitutionalization and legal changes that emerged in the 1970s, 1980s and 1990s, individuals with ID/DD had limited opportunities for services. Supports that were available offered either institutional care or informal care at home (Braddock & Parish, 2001; Brockley, 2004, Doll, 1937). This lack of community support as well as family poverty frequently led to institutional placements (Brockley, 2004; Noll, 1995). Between the Great Depression and World War II, institutions became overcrowded and an option for only the most difficult cases (Noll, 1995).

In the early history of this population, mainstream society and policy makers did not consider the quality of life for individuals with ID/DD. A gradual renaissance began with the emergence of advocacy movements in the 1960s (Scotch, 1989; White, Simpson, Gonda, Ravesloot, & Coble, 2010). This advocacy culminated in the deinstitutionalization of individuals with ID/DD (Braddock & Parish, 2001). The concept of quality of life underpinned many of these changes, such as the concept of normalization (Wolfensberger, 1972), which encouraged the full participation of people with disabilities in their communities, families and other social networks.

Legislative and funding changes introduced in the 1970s supported quality of life for individuals with ID/DD. The Intermediate Care Facilities for Individuals with Mental Retardation (ICF/MR) program through Title XIX (Medicaid) modified federal reimbursement and allowed individuals to move from large institutions to community nursing facilities (Lakin, Hayden, & Abery, 1994). Advocates lobbied for the addition of Title V to the Rehabilitation Act Amendments, which banned discrimination against individuals with disabilities where federal funds were the source of payment (Scotch,

1989). The Americans with Disabilities Act (ADA) of 1990 established civil rights protections for people with disabilities (Pfeiffer, 1993). With the 2008 amendment to the ADA, the federal government formally expanded several quality of life aspects for people with disabilities (U.S. Government, 2013).

Review of the Quality of Life Concept

The quality of life construct began in the social indicators research in the 1960s, which sprang from President Johnson's War on Poverty (Rapley, 2003). Previous research using economic measures was successful in forming domestic economic policy, but no comparable system was available for measuring social indicators that could inform social policy development (Land, 1983). From this emergence of social indicators, researchers identified three types of research: (a) normative welfare indicators, (b) satisfaction indicators, and (c) descriptive social indicators (Land, 1983).

From this research, the concept of quality of life in individuals with ID/DD emerged as a means to assess individuals with ID/DD. In the 1980s, the quality of life concept fit well with the emphasis on deinstitutionalization and normalization (Lyons, 2010; Schalock, 2000). As research progressed, additional reasons for including quality of life surfaced. Schalock and colleagues (2002) indicated that the concept of quality of life was relevant for studying individuals with ID/DD for several reasons. First, Schalock and collaborators noted, individuals with ID/DD tend to have more restricted choice options than others who did not have intellectual impairments. This is because individuals with ID/DD rely on assistance and support from others (Schalock et al., 2002). Second, individuals with disabilities may experience barriers to attaining full

involvement in their communities. Researchers surmised that the quality of life concept addresses these aspects of life.

From the quality of life literature, three objective principal concepts in ID/DD research emerged: (a) community inclusion, (b) choices, and (c) rights, which are objective. This dissertation assesses the psychometrics of objective quality of life measures. It defines these terms by considering historical antecedents and ways in which the research community evaluates these concepts.

Community inclusion. A first step toward community inclusion is the involvement of individuals with ID/DD in their communities (Lakin, Hayden, & Abery, 1994). Historically, mainstream society has viewed individuals with disabilities as being separate from conventional communities (Jaeger & Bowman, 2005). With persuasion from self-advocates, parent-advocates and other individuals supportive of change, several legal revisions have led to increased inclusion (Braddock & Parish, 2001; Scotch, 1989). Researchers have identified community inclusion as an essential concept, broad in scope and involving multiple aspects, such as types of environmental supports and adaptations, involvement of professionals and family members, service providers and individual preferences (Fujiura, 1994). From a measurement perspective, this concept is more challenging, since community inclusion in effect measures aspects of life that in other studies would be considered random measurement error (Fujiura, 1994).

Choice. The modern disability movement also considers choice to be another important concept, which grew from self-advocacy and self-determination (Wehmeyer & Abery, in press). Historically, individuals with disabilities have had few choices. People in institutions once received uniform care, which was not individualized to a person's

needs (Braddock & Parish, 2001). A key component of self-determination is the ability to make choices about the types of services and supports that a person with disabilities wants and needs and is consistent with quality of life (Abery, 1994). Abery (1994) defined choice making as one skill embedded in self-determination. This skill involves: (a) conveying one's needs to others, (b) consideration of available options, and (c) making a choice.

Rights. Although legal changes that took effect in the 1970s and were made explicit in the ADA established that individuals with ID/DD should have legal protection, there are multiple examples in which others usurped the rights of individuals with disabilities (Brockley, 2004; Ladd-Taylor, 2011). The passage of the ADA was a major legislative accomplishment that established universal rights for people with disabilities. "…[In] enacting the ADA, Congress recognized that physical and mental disabilities in no way diminish a person's right to fully participate in all aspects of society" (U.S. Government, 2013, p. 5).

Through advocacy and legal changes, the service options initiated in the 1970s and 1980s for people with ID/DD continue to expand. Novak Amado, Stancliffe, McCarron, and McCallion (in press) note that individuals with disabilities experience greater integration today than they ever did in the past. Community inclusion, choices, and rights are important concepts that now represent fundamental notions of quality of life for individuals with ID/DD.

Emergence of the quality of life definition. Today, researchers in various disciplines, including health, policy, and psychology, routinely apply the concept of quality of life as an outcome measure (Nussbaum & Sen, 1993). However, the definition

of the construct often varies, depending on the discipline. In medicine, for example, the World Health Organization (WHO) convened a workgroup to define health-related quality of life (H-R QoL) and to develop an assessment measure. At the onset, the WHO Quality of Life Group (1995) concluded that there is no universally accepted meaning of H-R QoL. The WHO Quality of Life Group was able to develop a definition of health-related quality of life to guide their work. The workgroup defined quality of life as and "individuals' perception of their position in life in the context of the culture and value systems in which they live and in relation to their goals, expectations, standards and concerns" (WHO, 1995, p. 1405). From this definition, the workgroup developed a multi-dimensional measure, which includes: (a) physical aspects, (b) psychological features, (c) level of independence, (d) social relationships, (e) environmental considerations, and (f) spirituality/religion/personal beliefs (WHO Quality of Life Group, 1995). Interestingly, this definition of health-related quality of life shares some similarities with the WHO's definition of health, which is "complete physical, mental well-being and not merely the absence of disease or infirmary" (WHO, 1948, p. 1).

In a similar vein, the developmental disabilities field has for years used the quality of life concept without a consistent definition. Researchers have developed a myriad of quality of life definitions in the ID/DD literature (Lyons, 2010), and some researchers estimate that over 100 definitions exist (Schalock, 2000; Schalock & Verdugo, 2002). These various definitions are specific to the: (a) researcher (Landesman, 1986), (b) topic studied (Rapley, 2003) and/or (c) unit of analysis (Land, 1983; Rapley, 2003). The lack of a uniform definition led Cummins (1991) to conclude that "there exists no generally agreed definition to its meaning, composition or how it should be measured" (p. 259).

Over time, experts in the field have collaborated to develop consistency in the construct. Researchers have agreed that quality of life is a multi-dimensional construct that is complex and hierarchical (Cummins, 1991; Cummins, 2005; Rapley, 2003; Schalock, 2000; Schalock & Felce, 2004).

Operationalizing Quality of Life

Construct dimensionality. Research shows that quality of life has multiple domains that constitute the breadth of the construct (Verdugo, Schalock, Keith, & Stancliffe, 2005) and represent the diverse elements of life in which individuals with ID/DD encounter quality of life (Lyons, 2010). These domains show general consistency among studies of individuals with ID/DD, but researchers have often interpreted concepts differently. For example, in Schalock's (2000) review article of the quality of life literature, he indicated that eight core domains are present, while Cummins (1991) identified seven. Schalock noted that when comparing his dimensions to Cummins', Cummins' domain of *productivity* could reflect two of Schalock's dimensions, personal development and self-determination. Felce and Perry (1995) identified five common domain themes, (a) physical well-being, (b) material well-being, (c) social well-being, (d) development and activity, and (e) emotional well-being. Table 1 compares the various domains and information offered by Felce and Perry (1995), Schalock (2000), Verdugo, Schalock, Keith, and Stancliffe (2005) and Cummins (1991). Essentially, the core elements are present in all of these conceptualizations, but the number of domains depends on how the researcher chose to categorize the core elements. Bonham and colleagues (2004) and Schalock (2004) drew similar conclusions when they noted that the

number of domains was not as critical as the point that the quality of life construct has a broad scope that includes various features.

Table 1

Items included in Felce and Perry's (1995) quality of life domains contrasted with Verdugo (2005), Schalock et al. (2004), and *Cummins (1991)*

Domains	Verdugo, 2005	Schalock, 2004	Cummins, 1991	Felce & Perry, 1995
Emotional well-being	X	X	X	X
Physical well-being	X	X	X	X
Material well-being	X	X	X	X
Social well-being	X	X		X
Interpersonal relationships	X	X	X	
Development and activity	X	X	X	X
Self-determination	X	X		
Rights	X	X		

Note: Verdugo indicated additional domains, which included family, recreation and leisure and safety/security. Cummins identified safety and health as additional separate domains.

Researchers have also indicated that quality of life may have a hierarchical component. Schalock (2000) identifies a pyramid structure similar to the level of Maslow's hierarchy of needs in which the base domain is physical well-being with other domain levels building upon this foundation. The structure from the bottom to the top includes: (a) physical well-being, (b) material well-being, (c) rights, (d) social inclusion, (e) interpersonal relationships, (f) self-determination, (g) personal development, and (h) emotional well-being. Similar to Maslow's hierarchy of needs, Schalock noted that an

individual must fulfill the components at the base and each successive level to progress to higher quality of life.

Based on concerns related to understanding, measuring and applying the quality of life construct, the International Association for the Scientific Study of Intellectual Disabilities (IASSID) convened a group of researchers to develop recommendations related to conceptualizing quality of life. From this group, the quality of life framework emerged, consisting of five components, as summarized by Schalock and colleagues (2002). First, quality of life is a universal concept that all individuals share, irrespective of disability status. Second, the satisfaction of an individual's needs and wants is a central tenet of perceived positive quality of life as is the possibility to follow goals and interests. Third, quality of life consists of subjective and objective appraisals of an individual's life. Fourth, quality of life includes "individual needs, choices and control" (Schalock et al., 2002, p. 460). Fifth, the construct is multi-dimensional and "influenced by personal and environmental factors, such as intimate relationships, family life, friendships, work, neighborhood, city or town of residence, housing, education, health, standard of living, and the state of one's nation" (Schalock et al., 2002, p. 460).

Quality of life measurement. Although researchers in the field of intellectual disabilities have indicated that individuals with ID/DD should respond for themselves whenever possible (Perry & Felce, 2002; Verdugo et al., 2005), there are circumstances in which individuals with ID/DD are unable to answer for themselves. For example, an individual with ID/DD who has a significant impairment in expressive language skills will likely be unable to answer. Lyons (2010) indicates that the assessment of individuals with significant ID/DD is one of the most difficult aspects in assessing quality of life.

In an effort to increase self-reports, researchers have developed recommendations to promote an individual with ID/DD to reply independently, such as providing: (a) unambiguously worded questions and responses (Verdugo et al., 2005), (b) picture answers (Perry & Felce, 2002), (c) alternative and augmentative communication (Schalock et al., 2002) and (d) questions that are concrete, without abstract implications (Heal & Sigelman, 1995; Perry & Felce, 2002).

Measurement challenges. Even with the application of these methods to increase self-reports, researchers remain concerned about the following areas: (a) the ability to respond, (b) lack of consistency of responses, (c) relationships to other information and (d) response biases (Perry & Felce, 2002). A non-responsive individual does not reply to questions or provides unrelated answers to questions (Perry & Felce, 2002). A person's response consistency is the stability with which he/she responds to questions, and which a test-retest correlation assesses. The relationship with other information is established through the invariability between an individual's response and other sources, such as document reviews. While these first three challenges are relatively intuitive, response biases are more complex, and separate sections provide a review of the research related to *acquiescence* and *recency effect*.

Acquiescence. Acquiescence is a response bias in which an individual responds in the affirmative, regardless of the question's content. The ID/DD literature has extensively studied acquiescence (Sigelman, Budd, Spanhel, & Schoenrock, 1981). Heal and Sigelman (1995), in their review article related to response bias, indicate that acquiescence is more pronounced in individuals with lower IQ scores than in those with higher IQs. Perry and Felce (2002), in a replication study exploring the impact of bias in

self-responses, found that acquiescence occurred more often in individuals with approximate language scores of less than nine years using the *British Vocabulary Scale*.

Researchers have used acquiescence bias as a pre-screen to identify individuals with ID/DD who respond 'yes' to reverse worded questions. As an example, Stancliffe (1999) excluded 11 participants from a study intending to measure similarities between self- and proxy-responses when self-respondents answered 'yes' to both positive and negative questions covering the same content.

Recency. Researchers define the recency effect as a patterned response. In this, individuals consistently choose the last option rather than answer queries with a variety of all possible response choices (Heal & Sigelman, 1995; Perry & Felce, 2002). Heal and Sigelman (1995) found that the recency effect is not as prevalent as acquiescence in this population. In studying the impact of biased responses, Heal and Sigelman concluded that the recency effect is more pronounced in individuals with more significant language delays than in individuals with less language impairment.

To safeguard findings due to response biases, Cummins (1993) developed a pre-screening method to identify individuals who exhibit response bias or inconsistencies in their responses. Individuals proceed through a pre-screening component of the Comprehensive Quality of Life Scale – Intellectual Disabilities (COMQoL-ID), which involves responding to three successive tasks that begin as concrete concepts then move to more abstract ideas (Cummins, McCabe, Romeo, Reid, & Waters, 1997). Based on the pre-screening results, interviewers then ask individuals with ID/DD to answer questions on the COMQoL-ID using two response options when individuals with ID/DD can only respond to the concrete pre-screen questions (Cummins et al., 1997). Those individuals

with ID/DD who are successfully able to complete all three levels of pre-assessment can respond to subjective questions on the COMQoL-ID using either a 5-point scale or a 7-point scale, depending on the question (Cummins et al., 1997).

Use of proxies. An ongoing problem associated with assessing the quality of life of individuals with ID/DD is assessing people who have significant impairments in their language skills, which make self-reports challenging, if not unattainable. Proxies then provide a response when individuals with ID/DD are unable to do so. According to Rapley, Ridgeway, and Beyer (1998), while proxies can accurately report the experiences of verbal people with a disability, they should also be able to represent the experiences of individuals who re non-verbal. Researchers have indicated that when individuals are unable to respond for themselves or when biases related to acquiescence and recency are likely, proxy-responses are acceptable (Cummins et al., 1997; Perry & Felce, 2002; Stancliffe, 1999). Rapley and colleagues (1998) acknowledge that the conceptual difference between proxies for individuals who are verbal versus those who are non-verbal is notable, since the proxy-respondent can only provide an estimate of the individual's experience.

Use of proxies allows researchers to glean some information about the lives of individuals with ID/DD who are unable to answer for themselves. In addition, proxy use allows for a larger and more representative sample, since it is possible to gather information about all additional individuals with ID/DD and not just self-report (Cusick, Brooks, & Whiteneck, 2001). However, one concern is that proxies may not have the necessary information to provide an accurate response. Another is that a proxy might potentially implicate himself/herself as providing poor service if an individual whom they

support has a low quality of life rating. Researchers have evaluated the validity of proxy use, including to provide answers to subjective and objective questions.

Subjective measures. An individual's perception of his/her experience is the cornerstone of subjective items (Campbell & Converse, 1972). Subjective measures then relate to a person's inner thoughts and feelings (Schalock & Felce, 2004). Diener and Suh (1997) noted that the subjective measures include three categories: (a) satisfaction, (b) pleasurable feelings and (c) unpleasant affect. One subjective item found on a quality of life questionnaire that measures pleasurable feelings in individuals with ID/DD is: "How much fun and enjoyment do you get out of life?" (Schalock & Keith, 1993, p. 1)

Studies comparing subjective measures from self-reports with proxy-reports have yielded mixed results. Some studies have found significant differences between matched pairs of self-reports and proxy-reports while other studies detected no differences between the self/proxy pairs. McVilly, Burton-Smith, and Douglas (2000) used the Comprehensive Quality of Life Scale – Intellectual Disabilities (COMQoL-ID) measure to evaluate the relationship between self- and proxy-reports. Using 24-matched pairs of self- and proxy-respondents, McVilly and collaborators found no statistically significant differences between the self- and proxy-reports in a related-samples *t*-test ($\mu_{self} = 3.82$; $\mu_{proxy} = 3.64$; $t(23) = 1.20$, $p = .24$). Based on the overall satisfaction scores, the self- versus proxy-ratings were also not statistically significant on a *t*-test analysis ($\mu_{self} = 4.52$; $\mu_{proxy} = 4.57$; $t(23) = .26$, $p = .80$). The authors suggested that the proxies knew each other well (parents served as 30% of the proxies), which likely contributed to the study's findings.

Other studies have not been as positive as the findings of McVilly, Burton-Smith, and Davidson (2000). An early study of satisfaction with residential services comparing 75 people with disabilities and matched proxies found no significant relationship between the self- and proxy-reports, $r = 0.11$, $p = 0.17$ (Burnett, 1989). Rapley, Ridgway, and Beyer (1997) studied 13 proxy-and 2-self-report pairs with Schalock and Keith's (1993) QOL-Q measure. Rapley and collaborators compared group means between self-reports and proxy-reports and found significant differences on the satisfaction scale of this measure ($\mu_{self} = 23.7$; $\mu_{proxy} = 22.6$; $t(12) = -1.75$, $p = .05$). Specifically, proxy-respondents underestimated the satisfaction of people with a disability on many items.

Cummins and colleagues (1997) and Perry and Felce (2002) all used the COMQoL-ID inventory to assess the statistical relationships between self- and proxy-responses using the satisfaction component of the measure. Cummins and collaborators found that self-and proxy-reports were significantly correlated on three of the seven domains (health: $p < .05$; productivity: $p < .01$; safety: $p < .01$), while the other four domains (material well-being, intimacy, community inclusion, and emotion) were not significantly associated. Finally, the Perry and Felce study compared reports from 154 people with a disability and their proxies (paid staff), and concluded that there was no relationship between self- and proxy-responses on the subjective COMQoL-ID scales. This finding led Perry and Felce to conclude: "These findings, therefore, provide no rationale for the use of staff as proxy-respondents in relation to subjective issues" (p. 453).

Objective measures. In contrast to the subjective evaluations, objective measures have three aspects: (a) universal to all individuals (Cummins, 2005), (b) observable and

(c) a general measure of welfare (Diener & Suh, 1997). An example of an objective-measure question is: "In the last month, did you go shopping?" (National Association of State Directors of Developmental Disabilities Services [NASDDDS] and Human Services Research Institute [HSRI], 2003, p. 40).

A number of studies have assessed the validity of self- and proxy-responses to objective measures. As with the subjective measures, researchers obtained mixed results for these scales. McVilly, Burton-Smith, and Davidson (2000) assessed the objective scales of the COMQoL-ID measure and found that in their matched-pair sample, which included 24 people with disabilities and their proxy-respondents, there were no significant differences for these groups based on a *t*-test analysis.

Rapley, Ridgway, and Beyer (1998) evaluated the social belonging/community integration scale and the empowerment scale, measuring choice, of the QOL-Q instrument, which are objective scales within this measure. In their sample of 13 matched pairs of people with disabilities and their proxies, they found that there were statistically significant differences between these groups on the social belonging/community inclusion scale ($t(12) = 3.53$, $p = .00$), but no significance on the empowerment scale ($t(12) = 1.38$, $p = .09$). Further assessment of the social belonging/community inclusion scale revealed that the proxy-reports indicated higher community inclusion than the self-respondents (Rapley et al., 1998). However, findings of the Rapley et al. study should be interpreted with caution due to its small sample size.

Following Rapley, Ridgway, and Beyer's (1998) findings, Stancliffe (1999) attempted to replicate the findings of the empowerment scale of the QOL-Q assessment. In this study, researchers administered the QOL-Q to 63 individuals with cognitive

disabilities and 2 paid staff proxies who were familiar with them. Scores for staff 1 and staff 2, as well as the average scores of the staff correlated with the self-responses. All of the correlations showed moderate association (r = .59, .62 and .64, respectively). In addition, Stancliffe found no statistically significant differences between self and proxies when comparing group mean scores with a t-test ($t(62) = 1.37$, $p = .18$).

Perry and Felce (2002) used Stancliffe's (1995) objective measure to assess the degree of choice in people with disabilities. In this sample, Perry and Felce found that there was a significant correlation between self-respondents and proxies (n = 56, *rho* = .74, $p < .01$).

Methodologically, all of these studies used pairs between a verbal self-respondent and a proxy-report, as well as matched t-tests and correlations. While the results indicate mean differences between groups, the findings can lack specificity. When considering t-tests, the function is to compare group mean differences, which do not allow for item difference detection. As such, the research to date has focused on general differences that occur due to variance between groups (Ackerman, 1992). In addition, researchers have thus far only considered the validity of self- and proxy-reports for verbal individuals with a disability, but not the validity of the use of proxies for individuals with disabilities who are unable to express their perspectives verbally.

An important aspect to quality of life is its measurement. The next section gives an overview of measurement theory with attention to modern measurement theory. Then the second portion considers various instruments that researchers developed to capture quality of life.

Chapter III: Measurement

Psychometric Measurement Considerations

Two important concepts of measurement theory are reliability and validity. This dissertation offers a general definition of each. Reliability is the consistency of scores (Carmines & Zeller, 1979; Thorndike & Thorndike-Christ, 2011). From a Classical Test Theory perspective, researchers define reliability as the ratio of true-score variance to observed-score variance (Haladyna & Downing, 2004). Offering another definition, Nunnally and Bernstein (1994) view reliability as the absence of random error, which does not correlate with any measured value or the true score. The many origins of random error include: (a) guessing, (b) errors in giving the measure, (c) scoring mistakes and (d) examinee fatigue (Crocker & Algina, 1986).

Reliability. Four types of reliability are: (a) stability or test-retest, (b) equivalence or parallel-form, (c) stability-equivalence, and (d) internal consistency or split-half (Thorndike & Thorndike-Christ, 2010). Researchers determine which reliability estimate to use, based on which threat they believe is most salient to generalization in their study (Cortina, 1993). In test-retest reliability, the goal is to establish the stability of a measure over two different testing times. The relationship between the two administrations of the same test yields information about the stability of a measure (Thorndike & Thorndike-Christ, 2010). In the second type of reliability, parallel-form (also referred to as alternate form), researchers give two different tests intended to measure the same construct or skill, in order to examine the relationship between them (Bobko, 2001). While Osburn (2000) noted that this is the preferred type of reliability, the development of a truly parallel test is complex (Nunnally & Bernstein,

1994). The correlation between the two tests provides information about the consistency of scores for these two measures. In stability-equivalence, the third type of reliability assessment, individuals complete two different tests at different times. In this type of reliability, both content and time can impact the results of the scores.

The last type of reliability, internal consistency, is the most commonly reported form of reliability (Cortina, 1993). Internal consistency reliability considers variance associated with subjects and subject-by-item interactions (Cortina, 1993). Here researchers select different statistics based on the level of measurement, including Kuder-Richardson Formula 20 (KR-20), for use with dichotomously scored items (Thorndike & Thorndike-Christ, 2010), and Lambda 2 (Osburn, 2000) and coefficient alpha for ordinal, interval, or ratio data (frequently referred to as Cronbach's alpha), which is the average of all split-halves (Osburn, 2000). There are several key problems with coefficient alpha, especially related to the number of items and multiple dimensions found in a measure (Cortina, 2000). Cortina (2000) developed a *Monte Carlo* study to assess internal consistency under a number of conditions and found that as item numbers increased, so too did internal consistency estimates. The same findings also held for multi-dimensional measures.

Messick (1989b) defined validity as "an evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (p. 13). Messick added that test scores are to include "any observed consistency, not just on tests as ordinarily conceived but also on any means of observing or documenting consistent behaviors or attributes" (Messick, 1989a, p. 5). Validity then

addresses systematic error, which is not random, but specific to individuals or groups, and is predictable (Carmines & Zeller, 1979; Haladyna & Downing, 2004; Thorndike & Thorndike-Christ, 2011). Validity is "the most fundamental consideration in developing and evaluating tests" (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999, p. 9).

In recent years, measurement theory has undergone some important theoretical changes. Depending on the theoretical measurement orientation, researchers consider reliability and validity in different ways. Traditional measurement theory builds on the three-part concept of validity that includes content, construct and criteria validity. In contrast, modern measurement theory uses a unified view of validity based on building evidence to support interpretations.

Traditional measurement theory. More traditional models of validity identify three key categories: (a) content, (b) criterion and (c) construct validity with face validity considered as a fourth aspect (Bobko, 2001). The traditional model of measurement theory defines validity as whether a test measures what it set out to measure when applied in specific ways (Bobko, 2001), and that validity is a property of the test (Zumbo, 1999). Kane (2001) noted that researchers used each validity type in specific situations. For example, researchers interested in assessing topics related to industrial-organizational psychology frequently used criterion validity to support job-placement decisions. Content validity was identified in achievement tests and construct validity was important for theoretically oriented research (Kane, 2001).

In the 1980s, measurement researchers critiqued the three-criteria model of measurement, which eventually grew into modern measurement theory. Messick (1980) noted that the content, criteria, and construct conceptualization would require researchers to simplistically "focus on one or another of the types of validity, as though any one would do, rather than on the specific inferences they intend to make from the scores" (p. 1014). In addition, Messick noted that the three criteria perspective does not consider the social or value consequences of scores. Messick and other colleagues outlined a model of validity in which researchers provide evidence to support the degree to which a measure accomplishes the intended inferences and uses of the measure as well as consequences of a given score (Messick, 1980). Gorin (2007) noted that the importance of considering inferences pertinent to what the test is measuring has added complexity to the validation process.

In contrast to the three-component model, modern measurement theory does not recognize distinctions between these types of validity, because in practice it is difficult to identify divisions between the concepts since the three-criteria frequently overlap, or researchers cannot meet the conditions for each type. As an example, Guion (1977) noted that there are multiple stipulations associated with content validity. These stipulations are that content validity should consider an operational definition that has explicit margins with behaviorally observable measures that capture the full extent of the construct and are reliable. In short, Guion indicates that these conditions are difficult if not impossible to achieve. Messick (1980) proposed that a test score's validity involve several considerations related to (a) content, (b) responses, (c) internal structure, (d) relationships with other variables and (e) consequences of testing. Messick and Kane

(2006) expanded on this idea of validity that the test scores should ultimately reflect evidence of validity, which researchers support through assessment of observable behaviors and deductive reasoning. Messick's second point is that the use of tests followed "justification of the proposed use in terms of social values" (p. 1012) and tied to social justice concepts.

Modern measurement theory. As such, the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999) recognized a unified form of validity with various types of evidence that support the intended inferences and uses of the test. The Standards for Educational and Psychological Testing outlined six principal sources of validity evidence. In developing and validating a test's use, instrument makers identified salient aspects of validity evidence and sought to assess the evidence for or against an intended use or score inference. The six sources of evidence outlined in the Standards for Educational and Psychological Testing include: (a) construct, (b) content, (c) response processes, (d) internal structure, (e) relationships to other variables, and (f) consequences. The following section considers validity evidence with special attention to construct, internal structure and consequences since they are pertinent to the research questions. This dissertation does not consider content, response processes, and relationships to other variables with as much detail since these concepts are not germane to the study purpose.

Construct. In the social sciences, it is usually not possible or is impractical to observe a trait or a construct directly, so researchers use behaviors, checklists, or other methods to gather information about the construct (Thorndike & Thorndike-Christ, 2010). Researchers then develop theories to organize these traits or constructs.

Cronbach and Meehl (1955) indicated that score interpretation associates findings to constructs. It seeks to answer the question: What does a score of 'X' on a measure represent when considering construct 'Y'?. Shadish, Cook, and Campbell (2002) further clarified construct validity as the extent to which inferences associated with specific study characteristics are applicable to a theoretical construct or what Messick (1980) identified as the nomological net.

There are two threats to construct validity, which the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999) identified as: (a) construct underrepresentation and (b) construct-irrelevance. In construct underrepresentation, the measure does not include the full breadth of the trait (Messick, 1995). In construct-irrelevance, an instrument measures other aspects above that of the intended construct. Messick outlined two forms of construct-irrelevancy: construct-irrelevant difficulty and construct-irrelevant easiness. When construct-irrelevant difficulty is present, people score lower on exams. Construct-irrelevant difficulty for various groups is a cause of bias that impacts a group's test score, which then shapes interpretation of scores (AREA, APA, & NCME, 1999). An example of construct-irrelevant difficulty in school assessments is when English Language Learners (ELL) complete a math test that includes story problems. If the construct studied is math computations, this test could present added difficulty for an ELL student due to challenges associated with reading and understanding English. Differential item functioning is one method to assess the possibility of bias and uncover construct-irrelevant difficulty in groups.

In contrast to construct-irrelevant difficulty, construct-irrelevant easiness allows some individuals to correctly answer questions that do not relate to the studied construct.

An example of this is test-wiseness (Rogers & Bateson, 1991), which with individuals can answer questions correctly because they are able to glean clues from the test to arrive at correct answers.

Content. Messick (1998) noted that this type of validity evidence intends to address the question: Is the measure examining the correct aspects of the construct in an even way? Content validity evidence considers the assessment of the full breadth of knowledge, skills, and abilities of a construct with measures high in content validity, including an array of questions that covers the entire domain of the construct (Carmines & Zeller, 1979). Messick (1998) noted that there are three types of evidence to consider: (a) representativeness, (b) relevance, and (c) technical quality. Researchers use several methods to establish content validity evidence. A frequently used method is a test blueprint in which test developers determine the number, type, and degree of difficulty associated with questions to capture the construct (Thorndike & Thorndike-Christ, 2010), which builds evidence associated with content representativeness.

Response processes. This type of evidence is related to the approach that test takers use to answer questions. Downing (2003) expanded on this definition of response process and developed a more comprehensive model. In it, he defined response process as accounting for potential sources of error that occur in test administration. Downing summarized this as quality control of the testing experience from accurately representing the responses to reporting the outcomes.

Internal structure. AERA, APA, and NCME (1999) identified the internal structure as an assessment of relationships between the test items and the construct measured in the test. It seeks to evaluate how well the items contribute to the measure.

This type of validity evidence considers both psychometric assessments of the structure as well as scoring. Researchers evaluate the internal structure of the test through statistical assessments to determine several aspects, such as reliability, uni-dimensionality, relationships with other items, and differential item functioning (Downing, 2003). An exploratory factor analysis completed on test items provides an assessment of the relationships between items as well as a score reliability estimate (AERA, APA, & NCME, 1999). Downing stressed that reliability is an important concept to the internal structure of a measure because it provides evidence associated with score consistency. If a test has low reliability, then the scores change over time.

Another type of analysis used to examine the internal structure of the test is confirmatory factor analysis (CFA). In CFA, researchers identify the structure of a measure prior to the analysis (Brown, 2006; Long, 1983). The construct theory then informs how the researcher completes the analysis with pre-established ideas related to the structure and components of the measure. The purpose of CFA is to assess the model fit between the *a priori* structure and the experimental sample (Brown, 2006). Measurement invariance, or the consistency of responses across groups, is another useful assessment (Millsap, 2012). In this way, comparisons are made between groups to determine the possibility of bias when groups respond in disparate ways.

Differential item functioning (DIF) of an assessment also leads to inferences about score interpretations. The purpose of DIF is to assess whether there are systematic dissimilarities in the ways in which different groups respond when conditioned on the total score (Osterlind & Everson, 2009). As an example, Fleishman, Spector, and Altman (2002) assessed DIF using the 1994/1995 National Health Interview Survey – Disability

Supplement to assess functional disability. The findings revealed that several items on this survey showed DIF when considering gender and age categories. In a comparison of middle-aged women to middle-aged men on levels of functional disability, the women were less disabled than the men ($t(10) = -2.56$, $p = -.10$). However, several items on this scale exhibited DIF when researchers conditioned the item scores on overall functional disability. When items were statistically adjusted to account for DIF, the authors concluded that these gender differences in functional disability were non-significant (Fleishman et al., 2002). This analysis shows that in considering omnibus hypotheses, such as *t*-tests, DIF within test items can contribute to statistically significant findings between groups that are an artifact of item differences rather than true differences in scores.

Relations to other variables. In previous conceptualizations of validity, this type of evidence is closely related to the idea of criterion validity, which includes concurrent, divergent and predictive evidence (AERA, APA, & NCME, 1999). In concurrent validity evidence, the scores of the measure being validated are correlated with scores on another test that assesses a similar construct (Thorndike & Thorndike-Christ, 2010). In completing this type of assessment, the purpose is to show a positive relationship between scores of a new measure and scores of the assessment used to evaluate the same construct. This type of evidence seeks to answer the question: Does current performance on a known test corroborate with score results on a new test? In contrast, a researcher could also establish that there is no relationship between scores on a studied assessment and another unrelated construct to establish divergent evidence. To assess predictive validity, researchers match a test score with a particular outcome (Thorndike &

Thorndike-Christ, 2010). For example, a standardized college admissions scores, which is the score interpretation being considered, is correlated with first-year college grades, or the predictive outcome (Geiser & Studley, 2002). Psychologists use predictive validity in employee selection, for which prospective employees complete an assessment; the scores on this assessment are used to predict future performance in a role or position.

Consequences. This is a controversial aspect of modern validity evidence with divergent views expressed in the literature. For example, Popham (1997) argued that it should not be included, while Shepard (1997) advocated its incorporation. This example of validity evidence was not as well developed as the other components, probably because psychometricians have only explicitly considered it in recent years.

Messick (1989b; 1994; 1998) provided a broad definition of consequential evidence. In this, Messick noted that both the planned and unplanned consequences should be considered for the short and long term. Additionally, consequences for individuals and groups, particularly underrepresented groups, should be considered. Messick (1998) noted "Ideally, there should be no adverse consequences associated with bias in scoring and interpretation, with unfairness in test use, or with negative effects on teaching and learning" (p. 15). Kane (2012) noted that invalidity associated with adverse consequences of the test is more serious than other consequences. Invalidity is tied closely with construct underrepresentation and construct-irrelevant variance. An example of an unintended consequence of test use includes low scores on a test of math ability, because the questions are not high-quality multiple-choice questions. Had the multiple-choice questions been high-quality, students could have performed better. Messick

(1998) also noted that valid assessments can yield negative consequences, which is an issue of social policy, not problems of measurement.

The literature identifies a number of different definitions of consequential validity, with emphasis on a different aspect of this type of evidence. Downing (2003) described the consequences as considering the intended and unintended effects of an assessment on the individuals who took the test and more broadly on how the test's findings could impact others, such as through policy changes. Cizek, Rosenberg, and Koons (2008) detailed the consequences associated with the test's use as having either a positive, negative or benign impact, whereas Shepard (2005) described test-use consequences as a consideration of social justice.

Argument-based approach to validity. Kane (1992; 2012) developed a process of considering the intended inferences and uses of a test and providing validity evidence to support the interpretations of the scores. Kane (2012) espoused the notion that researchers should develop specific arguments for validity consistent with the unique uses, contexts and populations. Based on the distinct concerns of the test and the testing situation, researchers develop validity evidence to support the interpretations and uses of the test scores. Kane also noted that validation of inferences was not a one-time assessment, rather it involves re-evaluations depending on how researchers use the test.

Kane's (2012) argument-based approach to validation considered two broad steps. The first step regards the intended inferences and uses of a measure's assumptions associated with the inferences and uses are made explicit. The second step provides evidence to either support or refute the inferences and uses of the measure. Kane noted that the evidence that provides information about troublesome interpretations or uses is

the most beneficial type of evidence. Kane also noted that a clear argument has three purposes. The first is that the explicit statement of the inferences and uses of a test uncovers the assumptions of a test. Second, the argument specifies the type of evidence gathered. Third, the argument supplies a basis for evaluating the intended inferences and uses of the test.

Measures to Assess Quality of Life

From the International Association for the Scientific Study of Intellectual Disabilities (IASSID) consensus statement about quality of life, two measurement considerations emerged. These criteria include that measures should: (a) contain subjective and objective assessments and (b) reflect the multi-dimensional nature of the concept. In addition, when selecting a measurement tool, it is important to also consider reliability and validity evidence as well as the ability to use proxy-responses when necessary.

Several comprehensive instruments meet the quality of life measurement criteria. These include: (a) Cummins' (1997) *Comprehensive Quality of Life – Intellectual Disabilities (COMQoL-ID)*, (b) Schalock and Keith's (1993) *Quality of Life Questionnaire (QoL-Q)*, (c) Bonham et al.'s (2004) *Ask Me! Survey*™, and (d) *National Core Indicators (NCI)*, the instrument used in this study. Table 2 provides a comparison of these measures.

Table 2

Measures identified through review of Agency for Healthcare Research and Quality (2010), Cummins (1997), and Li, Tsoi, Zhang, Chen, and Wang (2012)

Measure	AHRQ	Cummins	Li et al.
Ask Me! Survey/Ask Me! Survey - 2	X		X
Comprehensive Quality of Life Scale – Intellectual Disabilities (COMQoL-ID)		X	X
Choice Questionnaire (CQ)			X
Experience of Care and Health Outcomes Survey (ECHO) - Adult Survey	X		
Importance/Satisfaction (I/S) Map	X		
National Core Indicators (NCI)	X		
Personal Outcomes Measures	X		
Personal Life Quality Protocol (PLQ)	X		
Personal Outcomes Measures	X		
Personal Outcome Scale (POS)			X
Quality Indicators for Medicaid Services for People with Developmental Delay	X		
Quality of Life Questionnaire (QoLQ)	X	X	X

National Core Indicators. A group of researchers worked collaboratively with multiple states to create the National Core Indicators (NCI) to measure satisfaction with services (Stancliffe, Lakin, Taub, Chiri, & Byun, 2009) and track program outcomes, which allow states to make comparisons (Bradley & Moseley, 2007). The NCI survey consists of four sections: (a) background, (b) direct interview of subjective items, (c) direct or proxy interview of objective questions, and (d) interviewer feedback.

Researchers assessed the reliability of the NCI through several studies that used this instrument (Bradley & Moseley, 2007). In a sample of 27 individuals, the University of Minnesota's Research and Training Center on Community Living (2006) identified test-retest reliability of the NCI at 80%. Recent estimates of internal consistency reliability were not available for all domains of the NCI, though several are noteworthy due to their sample size. Ticha et al. (2012) found in a sample of 8,892 adults with

intellectual and developmental disabilities that the choice scale had an internal consistency of .80. In another study consisting of a national sample of 11,508 individuals, the community inclusion scale obtained a coefficient alpha of .64 (NCI, 2011a).

How states use the NCI. States use findings from the NCI in a number of ways, to educate individuals with ID/DD and their families, guide state policy change, and meet federal compliance standards. In developing these reports, states use a number of scales from the NCI, but the focus for this section is on the scales previously examined in this dissertation to determine how states practically use the scales.

The California Developmental Disabilities Consumer Advisory Committee (2012) used the *community inclusion* questions and the *life decisions* items to report outcomes with the intended audience as individuals with ID/DD and their families. In this report, California reported the outcomes without interpreting the results.

Multiple states use the NCI to inform policy decisions. As an example, Kentucky's Division of Developmental and Intellectual Disabilities (2013) compared findings of the *community inclusion* scale from 2008/2009 and 2011/2012 to find that individuals with ID/DD have had less community involvement in recent years. Based on these results, Kentucky launched a new service, known as community access under the Supports for Community Living waiver. The purpose of community access is "to encourage people with disabilities to engage in community life" (Kentucky Division of Developmental and Intellectual Disabilities, 2013, p. 3).

States also use the NCI findings to measure compliance with Centers for Medicare and Medicaid Services (CMS), which provides regulatory oversight for the

Medicaid waivers that serve to support individuals with ID/DD in their community. For example, Tennessee used NCI data to assess observation of CMS requirements (Henderson & Baird, 2014). In the state report, Henderson and Baird (2014) used the findings of the *community inclusion* scale to show that individuals who received Department of Intellectual and Developmental Disabilities (DIDD) services did participate in the community, meeting the Home and Community Based Services (HCBS) requirements. In addition, Tennessee reported the findings of the *life decisions* scale to identify state trends (Henderson & Baird, 2014).

Current Measurement Recommendations to Assess Quality of Life

In recent years, several researchers have developed measurement guidelines for people with disabilities, including the use of proxies. From the applied literature, two recommendations emerged related to subjective and objective measures. The first was that proxy-responses should not be used for subjective measures of quality of life (Cummins, 1997). The second was that it is acceptable to use proxy responses on objective measures, but only in specific conditions. Schalock (2010) provided several points of guidance related to when it is appropriate to use a proxy. The first is that proxies should be well known to the individual with a disability. Second, two proxies should be interviewed, and their average scores should be used. Third, when it is necessary to use proxy-responses, researchers should highlight that the proxy has a different frame of reference than the self-respondent. Finally, statistical techniques should be used to model a separate path for proxy data to account for their response differences.

In a similar vein, Verdugo, Schalock, Keith, and Stancliffe (2005) identified several approaches to proxy use, which are similar to Schalock's (2010) recommendations. However, Verdugo et al. recommended that analysis should not include both self-report data and proxy data. These authors state that the data are confounded when analysis includes both self- and proxy-responses. Stancliffe (2000) further clarified this position, noting that the proxy-responses systematically vary in the independent variable. The next section provides a critique of the measurement recommendations associated with the use of proxies with consideration of social justice aspects.

Critical review. If a researcher followed Verdugo et al.'s (2005) recommendations, then data would exclude perspectives from either self-respondents or non-verbal individuals. For example, a study intending to represent the full spectrum of individuals, meaning those with ID/DD who are verbal and those who are non-verbal, would not use data obtained from self-respondents, because this would introduce systematic differences. Or if a researcher only accepted the perspective of verbal respondents, then the data would not represent the experiences of non-verbal participants. Then, research in this area would exclude either the self-respondents or representations of non-verbal individuals. This is inconsistent with the concept of social justice, which the National Association of Social Workers defines as "equality of opportunity; and meaningful participation" (NASW, 2008, *Social justice*). A method used to assess systematic differences is differential item functioning, which assesses consistency of responses between groups.

Previous investigations have not made attempts to analyze these systematic differences. One method that could assess the systematic operations between self- and proxy-respondents is differential item functioning. This statistic assesses whether items operate in a different way depending on who answers the items rather than a person's knowledge, skills, or ability (Osterlind & Everson, 2009).

Research Purpose

The published research to date focuses on analyses that take into account general differences between self-responses and proxy-reports when self-respondents were verbal and proxies were known to them. In addition, the previous research considered overall differences and was unable to capture item-level differences. This dissertation is a between-subjects design that compares two groups - responses of individuals who answer for themselves and proxy-respondents - to determine if there were systematic differences between these groups at both the item and scale level. Through secondary data analysis of the 2009-2010 NCI objective quality of life measures, it was possible to assess the internal structure of the three quality of life scales when considering these two groups of respondents.

The NCI is recognized as a leading survey tool, used to capture a cross-section of experiences of individuals with disabilities. A concern with the NCI is that relatively little published data exist related to the psychometric properties of the instrument. This research intended to build evidence related to the internal psychometric properties of the NCI with special consideration of self- and proxy-responses.

The rationale of this research was to examine if there were systematic differences that occur between self- and proxy-respondents and what, if any, impact these differences

had on the three scale scores. It advances research in this area, since differential item functioning allows a researcher to consider dissimilarities in item functioning between groups of individuals who are able to respond for themselves and groups in which proxies provide the responses when conditioned on the total score for the scale. In addition, this research considered the factor structure of each of the scales for these groups of respondents, as well as measurement invariance.

Chapter IV: Methods

This dissertation assessed the internal structure of the National Core Indicators (NCI) using secondary data from the 2009-2010 NCI objective quality of life scales. The NCI established three objective quality of life scales, which included: (a) *Community Inclusion,* (b) *Life Decisions,* and (c) *Everyday Choices*. This study evaluated the internal structure of these three scales when considering two types of respondents: (a) self-respondents and (b) proxy-respondents. Self-respondents were those individuals with ID/DD who answered the questions for themselves. Proxy-respondents were primarily family members or paid staff who responded on behalf of individuals with ID/DD who could not answer for themselves. This dissertation seeks to evaluate the following questions and their corresponding hypotheses.

Research Question One

A. What are the internal consistency reliabilities of the three objective quality of life measures for both self- and proxy-respondents? Do these findings provide evidence to indicate adequate score consistency for applied research ($\alpha \geq .7$) for both the self- and proxy-respondents?

H1. The internal consistency reliability for self- and proxy-respondents will be close to acceptable coefficients for applied research for the *Community Inclusion* scale.

H2. The internal consistency for the *Life Decision* scale will be higher than acceptable score consistencies for applied research for both the self- and proxy-respondents.

H3. Coefficient alpha for self- and proxy-respondents for the *Everyday Choice* scale will be higher than .7.

Justification for research question and hypotheses. In quantitative studies in an applied setting, researchers recommend that a measure have at least an internal consistency of .7, which indicates that the scale does not measure more than 30% random error (Thorndike & Thorndike-Christ, 2010). The NCI (2011a) published the internal consistency reliabilities for each of the three scales as *Community Inclusion* $\alpha = .64$, *Life Decisions* $\alpha = .80$ and *Everyday Choices* $\alpha = .76$ for the 2009-2010 dataset. Human Services Research Institute (HSRI) calculated these reliability indices based on the overall sample, which included all respondent types. This study anticipated that the internal reliability indices would approximate the overall reliability indices for the three scales with the two respondent groups.

Research Question Two

B. To what extent is there evidence of meaningful item impact between self- and proxy-respondents for each question of the three objective measures?

H1. For the items in the *Community Inclusion* scale, self- and proxy-respondents will show similar amounts of community inclusion with no large effect sizes identified.

H2. For the items in the *Life Decisions* scale, there will be no practical differences between the two groups (self- versus proxy-respondents).

H3. For the items in the *Everyday Choices* scale, the self- and proxy-respondents will not show practical significance of choices between the response groups.

Justification for research question and hypotheses. For all of the items on the three scales, the author was unable to find any published studies that compared self- and proxy-respondents on the individual items. However, previous research did find that self-respondents tended to answer that they had more objective quality of life experiences when compared with individuals for whom a proxy responded when considering several scales (Bonham et al., 2004). However, Bonham et al.'s (2004) study used the entire scale score for the study. In considering individual items, it is unknown how each group will respond. Therefore, these hypotheses are exploratory in nature and do not suggest a specific direction.

This study used a large dataset. Due to the high power associated with a large dataset, it was important to not just consider statistical significance, but practical significance, which the effect size analysis determined. In considering the effect size, practical significance was any item that not only showed statistical significance, but also had an effect size measure that had small, medium, or large practical differences between the two groups of respondents.

Research Question Three

C. When conditioning on the total score, to what extent did the item response probabilities differ based on whether individuals with ID/DD responded for themselves or a proxy responded on their behalf? Is the effect size noteworthy if the analysis uncovers differences?

H1. The *Community Inclusion* scale will show evidence of meaningful differential item functioning for the question associated with *entertainment*.

a. The item related to *entertainment* will show differential item functioning between self- and proxy-respondents when conditioned on the total score with proxy-responses indicating greater community inclusion than self-respondents. This difference will also be meaningful.

H2. Each item in the *Life Decisions* scale will show no evidence of differential item functioning between these two groups when considering practical significance.

H3. Each item in the *Everyday Choices* scale will show no evidence of statistical or practical differential item functioning between these two groups of respondents when considering practical significance.

Justification of question and hypotheses. For this research question, no previous published studies were available. In a pilot study that the author completed on another dataset, it was found that the *entertainment* item exhibited DIF for self- and proxy-reports. This research continued to explore that finding. The remaining hypotheses were exploratory in nature, since there has not been any previous research on these items.

In a similar vein to the item impact question, the DIF analysis included an interpretation of the effect size. Since the dataset was large, it is likely the statistical analyses will indicate significance, but the interpretation of the effect size provides an understanding of the magnitude of these differences.

Research Question Four

D. How well do the data for self-respondents with ID/DD and proxy-respondents fit a uni-dimensional factor model for each of the three scales?

H1. Each of the three scales will demonstrate a uni-dimensional model for both the self- and proxy-respondents.

Justification of research question and hypothesis. Since researchers developed these three scales with a theoretical premise, it is anticipated that each of these scales will be uni-dimensional in nature.

Research Question Five

E. What type of measurement invariance exists between self- and proxy-reports on each of the three objective measures?

H1. The *Community Inclusion*, *Life Decisions*, and *Everyday Choices* scales will all demonstrate metric invariance.

Justification of question and hypothesis. Schalock (2010) noted that proxy-respondents systematically respond to questions in a different way than self-respondents. This would suggest some type of measurement invariance between these two groups. Scalar measurement invariance is the most constrained model and is difficult to achieve. With configural measurement invariance, this is the least constrained model and easiest to fit. Therefore, it is anticipated that the analysis will identify what metric measurement invariance holds for each of the three scales.

Design

Human Services Research Institute (HSRI) organized the collection of these data. As states collected these data within a similar timeframe, this study represents a cross-sectional study. The design also included a comparison of two groups – self-respondents and proxy-respondents – indicative of an independent-measures or between-subjects study. Self-respondents were those individuals with ID/DD who were able to respond for

themselves. Proxy-respondents provided answers based on what they believed individuals with ID/DD would have given had they been able to respond for themselves. This study was an observational study, since participants were not randomly assigned into either of the respondent types (Zumbo, 2007).

To answer the above research questions, the author used the same procedures for each of the three objective quality of life scales (*Community Inclusion, Life Decisions,* and *Everyday Choices*). In the first study question, the researcher calculated an internal consistency reliability coefficient. The internal consistency reliability yields a measure of accuracy for each of the three scales for both self- and proxy-respondents. Internal consistency reliability captured the average of all split-halves for each of the scales. In considering the overarching research question associated with the internal structure of the scales, this was the most salient reliability index to assess, since it was related to the internal aspects of the scales (Thorndike & Thorndike-Christ, 2010).

The purpose of item impact analysis was to determine if general group distinctions exist (Ackerman, 1992). The second research question then captured whether there were differences between these two groups of respondents. The purpose of completing this analysis was to obtain information about the distribution of scores for each of the two groups. To complete the item impact analyses, this researcher completed both an independent samples *t*-test for the *Community Inclusion* scale and the chi-square statistic test for the *Life Decision* and *Everyday Choices* scales. In this analysis, the independent variable was the type of respondent (either self or proxy), and the dependent variable was the response to the item.

The third research question addressed the differential item functioning (DIF) evaluation. In this analysis, all of the items on the three scales were assessed using the Mantel-Haenszel method. In addition, a hierarchical logistic or ordinal regression model fit comparison determined the statistical significance and magnitude of DIF for each of the questions. The author used both of these methods to answer this question since they were standard DIF assessment methods (Zumbo, 1999). In these analyses, the independent variable was the type of respondent (either self or proxy), the matching variable was the total score on the scale, and the dependent variable was the response to the item.

The fourth research question considered the dimensionality of each of the three scales for both self- and proxy-respondents. The purpose of uni-dimensional analysis was to verify that the underlying factor was uni-dimensional for both respondent types on each of the three scales. Since researchers developed these scales with a theoretical perspective, the next step was to determine if one latent factor held for both self- and proxy-respondents on each of the three scales (Harrington, 2008).

The final research question captured whether a configural, metric, or scalar measurement model fits these data. In completing this analysis, this researcher developed a nested model whereby additional constraints were placed on the data. In comparing this nested model, it was possible to then determine the level of measurement invariance.

Measures

As previously described, the National Core Indicators (NCI) is one of the quality of life assessment tools available to study this population (AHRQ, 2010). The National Association of State Directors of Developmental Disabilities Services (NASDDDS) and

the Human Services Research Institute (HSRI) partnered to establish this survey (NCI, 2011b). It has been used for over 15 years in multiple states (NCI, 2013).

The NCI included a structured interview comprised of five components. The first was a pre-survey portion that included general information, intended to assist with interview scheduling and collected before the interview took place (NCI, 2011c). The second segment was background information acquired before the interview, which contained information such as health history, employment information and living arrangements. States obtained the third and fourth portions through direct interview with individuals with ID/DD. In the third segment (identified as Section I in the survey), the questions related to subjective measures of quality of life, and interviewers only accepted self-responses for this part. The fourth aspect (denoted as Section II in the survey) was an assessment of objective quality of life. In Section II, interviewers accepted both self- and proxy-responses. Interviewers accepted proxy-responses from the following objective quality of life scales: (a) *Community Inclusion*, (b) *Life Decisions*, and (c) *Everyday Choices*. NCI (2011a) reported an internal consistency reliability of .64, .80, and .76, respectively, for the scales based on the 2009-2010 sample. This research project used all three objective scales for the analyses. The fifth and final section included the interviewer feedback section, which recorded the length of time, date, and any concerns associated with the interview.

The NCI manual provided guidance about who could answer as a proxy in Section II of the survey. Interviewers accepted responses from individuals who knew the person well, such as a family member, a friend, or a staff member. In addition, the NCI

specified that interviewers could not accept responses from individuals who provided service coordination, due to potential conflicts of interest (NCI, 2011c).

The NCI also established who could conduct the interview of individuals with ID/DD. The NCI (2011c) noted that individuals who served as assessors were people who did not know the person interviewed. HSRI used a train-the-trainer method to prepare interviewers for this role (NCI, 2011c). This process involved providing instruction to a cadre of staff in each state through a webinar. Those who participated in the webinar then presented this information to other trainers in their respective states. In doing this, the survey was implemented in a standardized method across all states (NCI, 2011c).

Objective quality of life scales. The 2009-2010 Final Report for the NCI (2011c) contained the three objective quality of life scales and described the items that comprised these scales. The next sections consider the items that make up these scales, response options, and scoring procedures for each scale.

Community inclusion scale. The components that make up the *Community Inclusion* composite scale included four items: (a) "In the past month, did you (did this person) go shopping? If yes, how many times?" (National Association of State Directors of Developmental Disabilities Services [NASDDDS] and Human Services Research Institute [HSRI] 2003, p. 40); (b) "In the past month, did you (did this person) go out on errands or appointments? If yes, how many times?" (NASDDDS & HSRI, 2003, p. 40); (c) "In the past month, did you (did this person) go out for entertainment? If yes, how many times?" (NASDDS & HSRI, 2003, p. 41); and (d) "In the past month, did you (did this person) go out to a restaurant or coffee shop? If yes, how many times?" (NASDDS

& HSRI, 2003, p. 41). This paper uses *shopping*, *errands*, *entertainment*, and *eating out* to refer to each question, respectively, to avoid restating each question. The NCI also included some additional questions related to community inclusion, but researchers did not include these items in the composite score, and are therefore not considered for this dissertation.

The *Community Inclusion* scale included open-ended counts of the activity. If someone responded that he/she had not completed the activity, he/she received a zero for that item. If someone participated in an activity, he/she or the proxy provided counts to reflect the number of times that the individual participated in the activity in the last month.

To score this scale, the frequency responses were aggregated for the four questions. If an NCI respondent indicated that he/she did not know, or the response was not understandable to the interviewer, those individuals did not receive a total score for the scale. In other words, the individual or the proxy needed to supply the interviewer with a response to all four questions to have a valid *Community Inclusion* scale score. The dataset included a total score for the *Community Inclusion* scale, which this dissertation uses as the total score for the *Community Inclusion* composite scale.

In addition to the frequency of each of the four activities in the scale, the interviewer also recorded who responded to each question. The response options included: (a) individual, (b) family/friend, (c) staff, or (d) other (NASDDDS & HSRI, 2003). For the purpose of this research the three types of proxy-reports (family/friend, staff, and other) were grouped together to represent a unified proxy-response.

Life decisions scale. The *Life Decisions* scale included seven indicators. The first question was: "Who chose (or picked) the place where you live?" (NASDDDS & HSRI, 2003, p. 45). The next question addressed choice of roommates: "Did you choose (or pick) the people who you live with (or did you choose to live by yourself)?" (NASDDDS & HSRI, 2003, p. 45). Question three captured the choice of staff: "Do you choose (or pick) who helps you at home?" (NASDDDS & HSRI, 2003, p. 46). To understand choice related to day activities, the NCI asked: (a) "Who chose (or picked) the place where you work?" and (b) "Who chose (or picked) where you go during the day?" (NASDDDS & HSRI, 2003, p. 47). The last two questions captured choice of staff during the day, and included: (a) "Do you choose (or pick) who helps you at your job?" and (b) "Do you choose (or pick) who helps you during the day?" (NASDDDS & HSRI, 2003, p. 48-50). This paper uses *home*, *roommate*, and *staff at home* to refer to the first three items.

To capture the final two indicators, this research combined two questions to reflect choice of day activity and choice of staff at the day activity. The two questions that reflected choice of day activities were: (a) "Who chose (or picked) the place where you work?" (NASDDDS & HSRI, 2003, p. 47) or (b) "Who chose (or picked) where you go during the day?" (NASDDDS & HSRI, 2003, p. 48). The two items that gathered information about choice of day staff included: (a) "Do you choose (or pick) who helps you at your job?" (NASDDDS & HSRI, 2003, p. 48) or (b) "Do you choose (or pick) who helps you during the day? (NASDDDS & HSRI, 2003, p. 49). The reason that these items were combined was that typically individuals with ID/DD either participate in a job environment or attend a day training and habilitation program, which led to a more robust

analysis. The dataset contained two additional variables to reflect these combined indicators. This dissertation uses *day activity* and *day staff* to refer to these combined items.

The *Life Decisions* scale included several response options from which the respondent could answer. These response options included: (a) not applicable, (b) yes, person chose, (c) some choice, (d) someone else chose, and (e) don't know or response unclear. The response options (b) and (c) were combined to reflect some choice. In other words, responses to these questions were scored dichotomously to reflect no choice (scored as 0) or choice (scored as 1). The scale score included only those questions in which participants chose response options (b), (c), or (d). To score this scale, researchers averaged the scores across the valid responses. For example, an individual could respond with either a (b), (c), or (d) response to three of the five questions, and the scale score was calculated based on the average of the three questions to which they responded. The data set included a total score for the *Life Decisions* scale, which this researcher used for the analysis.

In addition, the interviewer also recorded who responded to the question. The respondents were the same as those outlined in the *Community Inclusion* scale (individual, family/friend, staff, and other). For the purpose of this dissertation, the three types of proxy-respondents were recorded to reflect a single proxy group.

Everyday choices. The third scale, *Everyday Choices,* comprised three questions. The first question addressed schedules and read: "Who decides your daily schedule (like when to get up, when to eat, when to go to sleep)?" (NASDDDS & HSRI, 2003, p. 46). The second question captured choice of free time: "Who decides how you spend your

free time (when you are not working, in school, or at the day program)?" (NASDDDS & HSRI, 2003, p. 46). The final question was: "Do you choose what you buy with your spending money?" (NASDDDS & HSRI, 2003, p. 46). Rather than using the entire question, this paper uses *schedule*, *free time,* and *choose to buy* to refer to these items.

As with the *Life Decisions* scale, the response options for these questions were: (a) not applicable, (b) yes, person chose, (c) some choice, (d) someone else chose, and (e) don't know or response unclear. Also like the *Life Decisions* scale, the response options that demonstrated some choice and person chose were combined, which meant that valid responses were dichotomous. A zero response reflected no choice, whereas a one demonstrated some amount of choice.

To score this question, researchers averaged the items. If someone did not respond to an item, those items were not incorporated into the total average score. The dataset also included the total scale score for *Everyday Choices*, which this research used for the analysis.

Similar to the previous two scales, the interviewer also recorded who responded to the questions, which included: individual, family/friend, staff, or other. This researcher recorded the three types of proxy-respondents to reflect one proxy-respondent type rather than each individual proxy type.

Participants

A description of the participants is organized into two broad components: (a) participant selection and (b) sample characteristics. Within each of these components, this paper describes the entire sample and the study-specific participants.

Participant selection. To be eligible to participate in the NCI survey, individuals must have been older than 18 with a diagnosis of an intellectual or developmental disability, and a recipient of at least one service other than case management (NCI, 2011a). The dataset included 11,599 respondents, representing 16 states, the District of Columbia, and Orange County California (Human Services Research Institute (HSRI), 2010). The states included in this sample were Alabama, Arkansas, Georgia, Illinois, Kentucky, Louisiana, Maine, Missouri, New Jersey, New York, North Carolina, Ohio, Oklahoma, Pennsylvania, Texas, and Wyoming. Since this sample was large, it provided an opportunity to complete statistical analyses with adequate power to detect item and scale differences. HSRI did not require a pre-screening process, so all individuals who met the initial eligibility requirements of receiving at least one service other than case management were able to participate. Each state interviewed at least 400 individuals selected from those who were eligible to complete the survey across the state.

Each state was able to identify specific procedures for sampling individuals. Several states established a stratified selection process from which states randomly selected participants. For example, the classification for several states included random sampling based on (a) provider (i.e., Arkansas and North Carolina), (b) region of the state (i.e., Georgia, Missouri, Texas), and (c) region and service (i.e., Louisiana) (HSRI, 2010). The remaining states sampled participants based on a random selection processes.

HSRI developed criteria to exclude surveys at the time of the analysis. If individuals did not provide any responses to the subjective measures in the face-to-face interview portion (identified as Section I), and a proxy-report did not provide responses for the face-to-face interview to assess the objective measures (identified as Section II),

then HSRI excluded those surveys (NCI, 2011c). For the 2009-2010 data set, HSRI reported that 98.8% of all NCI surveys collected were valid for Section II, which was the focus of this study. The final data set for the 2009-2010 NCI consisted of 11,599 participants.

Sample characteristics of the entire dataset. HSRI provided a description of the 2009-2010 dataset in the Final Report. Tables 3 to 8 outline the proportion of the total sample according to gender, race, ethnicity, level of intellectual disability, age, and language. The NCI Final Report for 2009-2010 also provided information broken into each participating state; this study did not report this level of detail, as the analysis did not include individual states.

Table 3

Proportion of the 2009-2010 Entire NCI Sample before Identifying Studied Sample of Self- and Proxy- Respondents by Gender

Male	Female	*N*
56.9%	43.1%	11,508

Note. Information in this table is adapted from "Annual Summary Report 2009-2010" by National Core Indicators (2010). Retrieved from http://www.nationalcoreindicators.org/upload/coreindicators/NCI_Annual_Summary_Report_2009-10_FINAL.pdf

Table 4

Proportion of the 2009-2010 Entire NCI Sample before Identifying Studied Sample of Self- and Proxy-Respondents by Race

N-A[a]	Asian	A-A[b]	P-I[c]	White	Other[d]	Further[e]	Unknown	*N*
.9%	1.1%	20.3%	.1%	73.4%	2.8%	.6%	.8%	11,044

Note. Information in this table is adapted from "Annual Summary Report 2009-2010" by National Core Indicators (2010). Retrieved from http://www.nationalcoreindicators.org/upload/coreindicators/NCI_Annual_Summary_Report_2009-10_FINAL.pdf

N-A[a] represents Native Americans, A-A[b] indicates African-American, P-I[c] is Pacific-Islander, Other[d] signifies Other race not listed, and Further[e] symbolizes more than two races.

Table 5

Proportion of the 2009-2010 Entire NCI Sample before Identifying Studied Sample of Self- and Proxy-Respondents by Ethnicity

Non-Hispanic	Hispanic	Unknown	*N*
92.0%	7.1%	.9%	11,433

Note. Information in this table is adapted from "Annual Summary Report 2009-2010" by National Core Indicators (2010). Retrieved from http://www.nationalcoreindicators.org/upload/coreindicators/NCI_Annual_Summary_Report_2009-10_FINAL.pdf

Table 6

Proportion of 2009-2010 Entire NCI Sample before Identifying Studied Sample of Self- and Proxy-Respondents by Level of ID

No ID label	Mild	Moderate	Severe	Profound	Unknown	*N*
7.2%	32.0%	23.9%	14.0%	19.8%	3.2%	11,321

Note. Information in this table is adapted from "Annual Summary Report 2009-2010" by National Core Indicators (2010). Retrieved from http://www.nationalcoreindicators.org/upload/coreindicators/NCI_Annual_Summary_Report_2009-10_FINAL.pdf

Table 7

2009-2010 Entire NCI Sample before Identifying Studied Sample of Self- and Proxy-Respondents by Age

Min	Max	Mean	SD	Median	*N*
18	95	43.5	14.5	44	11,559

Note. Information in this table is adapted from "Annual Summary Report 2009-2010" by National Core Indicators (2010). Retrieved from http://www.nationalcoreindicators.org/upload/coreindicators/NCI_Annual_Summary_Report_2009-10_FINAL.pdf

Table 8

Proportion of 2009-2010 Entire NCI Sample before Identifying Studied Sample of Self- and Proxy-Respondents Who Speak English

English	Other	*N*
96.4%	3.6%	*11,412*

Note. Information in this table is adapted from "Annual Summary Report 2009-2010" by National Core Indicators (2010). Retrieved from http://www.nationalcoreindicators.org/upload/coreindicators/NCI_Annual_Summary_Report_2009-10_FINAL.pdf

Study-specific participation. An important assumption with this study related to the independence of the two respondent groups, meaning that the respondents consisted of only self-respondents and proxy-respondents. To accomplish this assumption, this researcher screened participants out when respondents were not consistent within each of the three objective quality of life scales (*Community Inclusion, Life Decisions,* and *Everyday Choices*). For example, if the case involved a combination of self- and proxy-respondents, then this researcher excluded the case from the scale analysis.

For the objective items, the NCI included information about who answered the question. The various types of respondents comprised: (a) individual, (b) family/friend, (c) staff, and (d) other (NASDDDS & HSRI, 2003). To assure respondent consistency, this researcher verified that the respondent type remained consistent throughout each scale. For example, if a staff member answered the first question, then a staff member consistently responded to the remaining scale questions. When the respondent type was inconsistent across the scale, those cases were excluded from the analysis.

To meet the assumption of only one respondent type, each of the scales contained different samples. The next section clarifies the samples for each of the three objective measures, provides information specific to each scale, and includes information related to excluded cases followed by a description of the included cases for each scale.

Excluded cases from study. This study excluded cases from the three scale analyses for a number of reasons. These included: (a) lack of information about who

responded to the question, (b) missing data, and (c) alternating respondents. This screening process eliminated 3,166 cases from the *Community Inclusion* analysis. Of these, 2,393 cases were excluded due to missing data, represented in points (a) and (b) from above, and the additional cases included inconsistent respondents. The *Life Decisions* scale had 1,823 cases excluded from the analysis primarily due to points (a) and (c) previously identified. The *Everyday Choices* scale had 1,173 cases that did not meet the criteria for analysis. Tables 16 through 22 provide demographic information related to gender, race, ethnicity, level of ID, age, legal status, and home type for the excluded cases. In addition, Tables 16 through 22 include a chi-square and *t*-test analysis to determine if there were differences between participants in this study and those not included in the analysis.

Community inclusion sample. The *Community Inclusion* scale consisted of a total of 8,433 cases. Of these, 4,280 were self-respondents (50.79%) and 4,153 (49.12%) were proxy-respondents. Tables 9 through 15 incorporate demographic information, including gender, ethnicity, legal status, race, level of ID, age, and home type that comprised the study cases for this sample. The tables provide a statistical comparison between self- and proxy-respondents based on these sample characteristics.

Life decisions sample. The *Life Decisions* scale included 4,935 (50.4%) self-respondents and 4,841 (49.5%) proxy-reports. The demographic information and a statistical comparison between self- and proxy-respondents for this sample are in Tables 9 through 15.

Everyday choices sample. This scale had 5,459 self-respondents and 4,967 proxy-respondents, with a total of 10,426 cases used for this analysis. Tables 9 through

15 show the demographic information and a statistical comparison between self- and proxy-respondents.

Table 9

Proportion of the Included Cases by Gender based on Respondent Type

Scale	Respondent	Male	Female	Unknown	*N*	χ^2
Community Inclusion						
	Self	56.2	43.2	.5	4280	.738
	Proxy	57.3	42.4	.2	4153	
Life Decisions						
	Self	55.6	43.5	.9	4935	1.66
	Proxy	57.2	42.4	.4	4841	
Scale	Respondent	Male	Female	Unknown	*N*	χ^2
Everyday Choices						
	Self	55.4	43.7	1.0	5459	2.73
	Proxy	57.3	42.3	.4	4967	

*$p < .10$. **$p < .05$. ***$p < .001$.

Table 10

Proportion of Included Cases Sample by Ethnicity based on Scale and Respondent Type

Scale	Respondent	Non-Hispanic	Hispanic	Unknown	χ^2	ϕ
CI[a]						
	Self	92	6.4	1.6	29.22***	.06
	Proxy	89.1	9.3	1.6		
LD[b]						
	Self	92.1	5.7	2.2	30.18***	.06
	Proxy	89.7	8.5	1.9		
EC[c]						
	Self	91.9	5.8	2.3	35.52***	.06
	Proxy	89.2	8.7	2.1		

Note. CI[a] indicates the scale *Community Inclusion*, LD[b] indicates the *Life Decisions* scale and EC[c] represents the scale *Everyday Choices*.
*$p < .10$. **$p < .05$. ***$p < .001$.

Table 11

Proportion of Included Cases Sample by Legal Status Based on Scale and Respondent Type

Scale	Respondent	Independent[a]	Limited[b]	Full[c]	Unknown	χ^2	ϕ
CI[d]							
	Self	65.6	4.6	26.7	3.1	449.85***	.23
	Proxy	43.1	6.7	46.7	3.5		
LD[e]							
	Self	67.0	4.5	24.3	4.1	529.53***	.24
	Proxy	44.7	6.4	44.4	4.4		
EC[f]							
	Self	66.5	4.6	24.4	4.5	564.11***	.24
	Proxy	44.9	6.6	45.3	3.2		

Note. Independent[a] is independent of guardianship, Limited[b] represents limited guardianship and Full[c] indicates full guardianship, CI[d] indicates *Community Inclusion*, LD[e] indicates *Life Decisions* and EC[f] represents *Everyday Choices*.
*$p < .10$. **$p < .05$. ***$p < .001$.

Table 12

Proportion of the Included Cases by Race Based on Scale and Respondent Type

Scale	Respondent	N-A[a]	Asian	A-A[b]	P-I[c]	White	Other[d]	Further[e]	Unknown	χ^2	ϕ
CI[f]											
	Self	.8	.8	20.3	.1	70.5	2.9	.5	4.1	28.64***	.06
	Proxy	1.2	1.5	17.6	.0	69.1	2.6	.4	7.5		
LD[g]											
	Self	.8	.8	20.3	.1	70.5	2.9	.5	4.1	22.63*	.05
	Proxy	1.2	1.5	17.6	.0	69.1	2.6	.4	7.5		
EC[h]											
	Self	.8	.8	19.5	.1	71.3	2.7	.5	4.2	25.22***	.05
	Proxy	1.1	1.4	17.2	.0	69.9	2.6	.5	7.3		

Note. N-A[a] represents Native Americans, A-A[b] indicates African-American, P-I[c] is Pacific-Islander, Other[d] signifies Other race not listed, Further[e] symbolizes more than two races, CI[f] indicates the scale *Community Inclusion*, LD[h] indicates the *Life Decisions* scale and EC[g] represents the scale *Everyday Choices*.
*$p < .10$. **$p < .05$. ***$p < .001$

Table 13

Proportion of Included Cases by Level of ID Based on Scale and Respondent Type

Scale	Respondent	None[a]	Mild	Moderate	Severe	Profound	Unknown	χ^2	ϕ
CI[b]									
	Self	10.7	50.4	25.7	5.6	2.2	5.3	3135.58***	.62
	Proxy	5.9	10.4	17.2	21.2	41.6	3.7		
LD[c]									
	Self	9.0	51.1	26.1	6.0	2.1	5.7	3353.96***	.59
	Proxy	5.1	11.7	18.7	20.9	39.0	4.6		
EC[d]									
	Self	9.0	50.1	27.0	6.1	2.3	5.5	3651.67***	.59
	Proxy	5.4	11.0	18.2	21.0	39.9	4.5		

Note. None[a] indicates no ID label, CI[b] indicates the scale *Community Inclusion*, LD[c] indicates the *Life Decisions* scale and EC[d] represents the scale *Everyday Choices*.
*$p < .10$. **$p < .05$. ***$p < .001$.

Table 14

Proportion of Included Cases Sample by Age Based on Scale and Respondent Type

Scale	Respondent	Min	Max	Mean	*SD*	*N*	*t*	*d*
CI[a]								
	Self	18	95	41.95	13.981	4271	-9.61***	-.21
	Proxy	18	95	44.93	14.494	4152		
LD[b]								
	Self	18	95	42.14	14.04	4916	-9.18***	-.19
	Proxy	18	92	44.81	14.59	4837		
EC[c]								
	Self	18	95	42.11	14.12	5440	-9.66***	-.19
	Proxy	18	92	44.84	14.66	4963		

Note. CI[a] indicates the scale *Community Inclusion*, LD[b] indicates the *Life Decisions* scale, and EC[c] represents the scale *Everyday Choices*.

*$p < .10$. **$p < .05$. ***$p < .001$.

Table 15

Proportion of Included Cases Sample by Type of Home Based on Scale and Respondent Type

Scale	Respondent	Inst.[a]	Group[b]	Agency[c]	Alone[d]	Parent[e]	Foster[f]	Nurse[g]	Other	Unknown	χ^2	ϕ
CI[h]												
	Self	9.3	24.2	7.8	15.3	29.0	4.4	.6	6.0	3.4	1106.13***	.36
	Proxy	37.3	24.4	2.0	8.1	19.9	2.7	.7	3.3	1.6		
LD[i]												
	Self	8.1	24.9	7.2	15.6	29.3	3.9	.7	6.2	3.9	1206.17***	.35
	Proxy	34.1	25.5	2.0	7.3	22.0	3.0	.8	3.3	2.1		
EC[j]												
	Self	8.0	25.4	7.3	15.1	30.1	3.8	.8	5.9	3.7	1360.23***	.36
	Proxy	35.0	25.5	2.0	6.9	21.6	2.8	.8	3.3	2.0		

Note. Inst.[a] represents specialized institution for individuals with IDD, Group[b] indicates group home residence, Agency[c] is an agency run apartment or home, Alone[d] signifies independent home, Foster[f] represents foster home and Nursing[g] is a nursing home, CI[h] indicates the scale *Community Inclusion*, LD[i] indicates the Life Decisions scale and EC[j] represents the scale *Everyday Choices*.
*$p < .10$. **$p < .05$. ***$p < .001$.

Table 16

Proportion of Excluded Cases by Gender and Scale

Scale	Male	Female	Unknown	χ^{2a}
CI[b]	55.6	42.6	1.8	.11
LD[c]	56.9	41.6	1.5	.63
EC[d]	58	40.7	1.4	1.86

Note. χ^{2a} represents a comparison between the cases included in the study and those excluded, CI[b] indicates the scale *Community Inclusion*, LD[c] indicates the *Life Decisions* scale, and EC[d] represents the scale *Everyday Choices*.
*$p < .10$. **$p < .05$. ***$p < .001$.

Table 17

Proportion of Excluded Cases by Race and Scale

Scale	N-A[a]	Asian	A-A[b]	P-I[c]	White	Other[d]	Further[e]	Un-known	χ^{2f}, ϕ
CI[g]	.6	.8	20.4	0	70.0	2.5	0	5.7	11.28
LD[h]	1.0	.7	26.3	.1	62.2	3.2	0	6.6	78.10***, .08
EC[i]	.5	.9	27.7	0	62.8	2.8	0	5.2	60.99***, .07

Note. N-A[a] represents Native Americans, A-A[b] indicates African-American, P-I[c] is Pacific-Islander, Other[d] signifies other race not listed, and Further[e] symbolizes more than two races, χ^{2f} represents a comparison between the cases included in the study and those excluded, CI[g] indicates the scale *Community Inclusion*, LD[h] indicates the *Life Decisions* scale, and EC[i] represents the scale *Everyday Choices.*
*$p < .10$. **$p < .05$. ***$p < .001$.

Table 18

Proportion of Excluded Cases by Ethnicity and Scale

Scale	Non-Hispanic	Hispanic	Unknown	χ^{2a}, ϕ
CI[a]	91.0	4.8	4.2	32.09***, .05
LD[b]	89.4	6.4	4.2	7.29*, .03
EC[c]	91.5	5.1	3.4	6.72*, .02

Note. χ^{2a} represents the comparison between included cases and excluded cases, CI[b] represents the *Community Inclusion* scale, LD[c] indicates the *Life Decisions* scale, and EC[d] is the *Everyday Choices* scale.
*$p < .10$. **$p < .05$. ***$p < .001$.

Table 19

Proportion of Excluded Cases by Level of ID and Scale

Scale	None[a]	Mild	Moderate	Severe	Profound	Unk[b]	χ^{2c}, ϕ
CI[d]	3.5	32.4	28.1	14.6	13.2	8.1	211.82***, .14
LD[e]	6.6	29.0	28.2	15.2	13.7	7.3	74.43***, .08
EC[f]	4.8	28.6	27.9	17.6	11.6	9.6	117.47***, .10

Note. None[a] indicates no ID label, Unk[b] represents unknown χ^{2c} is a comparison between the included and excluded cases, CI[d] represents *Community Inclusion*, LD[e] is *Life Decisions,* and EC[f] denotes *Everyday Choices*.
*$p < .10$. **$p < .05$. ***$p < .001$.

Table 20

Proportion of the Community Inclusion Excluded Cases by Age

Scale	Min	Max	Mean	*SD*	*N*	*t* statistic
CI[a]	18	91	43.63	14.9	3137	-.73
LD[c]	18	90	43.5	14.9	1804	-.14
EC[b]	18	94	44.02	14.6	1156	-1.36

Note. CI[a] indicates the scale *Community Inclusion*, LD[b] indicates the *Life Decisions* scale, and EC[c] represents the scale *Everyday Choices*.
*$p < .10$. **$p < .05$. ***$p < .001$.

Table 21

Proportion of the Community Inclusion Excluded Cases by Legal Status

Scale	Independent[a]	Limited[b]	Full[c]	Unknown	χ^{2d}, ϕ
CI[e]	57.0	5.7	28.3	9.1	113.13***, .10
LD[f]	50.8	6.8	34.4	7.9	25.75***, .05
EC[g]	48.6	7.2	36.4	7.9	24.68***, .05

Note. Independent[a] is indepenent of guardianship, Limited[b] represents limited guardianship and Full[c] indicates full guardianship, χ^{2d} is a comparison between studied cases and excluded cases, CI[e] represents *Community Inclusion*, LD[f] is *Life Decisions*, and EC[g] indicates *everyday choice*.
*$p < .10$. **$p < .05$. ***$p < .001$.

Table 22

Proportion of the Community Inclusion Excluded Cases by Type of Home

Scale	Instit.[a]	Group[b]	Agency[c]	Alone[d]	Parent[e]	Foster[f]	Nurse[g]	Unknown	χ^{2h}, ϕ
CI[i]	10.7	30.0	5.6	9.1	32.2	3.1	1.3	7.8	45.71***, .31
LD[j]	19.7	25.9	5.1	11.0	26.6	3.5	.8	7.5	144.24***, .11
EC[k]	9.3	30.3	8.4	9.0	31.4	4.8	.9	6.0	145.84***, .11

Note. Inst.[a] represents specialized institution for individuals with IDD, Group[b] indicates group home residence, Agency[c] is an agency run apartment or home, Alone[d] signifies independent home, Foster[f] represents foster home and Nursing[g] is a nursing home, χ^{2h} indicates a comparison between studied cases and excluded cases, CI[i] represents the *Community Inclusion* scale, LD[j] indicates the *Life Decisions* scale, and EC[k] represents the *Everyday Choices* scale.
*$p < .10$. **$p < .05$. ***$p < .001$.

Procedures

This next section provides a description of the Institutional Review Board (IRB) approval process, then presents an overview of the analyses used to assess the questions identified earlier. To review, these three questions relate to: (a) exploration of the internal consistency reliability of the three scales for self- and proxy-respondents, (b) assessment of item impact for all of the items on the three scales, (c) evaluation of differential item functioning for each item on the three scales, (d) determination of uni-dimensionality for both self-respondents and proxy-respondents for the three scales, and (e) study of measurement invariance for these two populations on all three of the scales.

Institutional review board. The author submitted an IRB Determination of Human Subject Research form to inquire about the need for an IRB review. After reviewing the determination document, the University of Minnesota's IRB indicated that approval for this research was not required, since the data was stripped of all identifying information.

Data Analyses. For all three scales of the NCI, the analyses were the same for each. The next sections provide information on how this researcher examined the data to address the questions outlined in this study.

Reliability. The reliability index provided an estimate of measurement consistency within a scale. It was calculated for both the self- and proxy-respondents for all three scales. This analysis was completed using SPSS, version 20 (2011). For the *Community Inclusion* scale, SPSS used the coefficient alpha formula, whereas for the dichotomous scales (*Life Decisions and Everyday Choices)*, the SPSS default was the Kuder-Richardson 20 formula.

Item Impact. Ackerman (1992) noted that item impact provides information about general group differences. If the analysis revealed group dissimilarities, these differences were due to differences in knowledge, skills, or abilities that the measure captured (Zumbo, 1999). The item impact assessment involved both an independent samples *t*-test (Angoff, 1993; Dorans & Holland, 1993) and chi-square. This researcher used SPSS, version 20 for the item impact analyses.

The level of measurement for the *Community Inclusion* scale was continuous, an assumption of the *t*-test. Since the two groups compared were independent of each other, this researcher used an independent samples *t*-test for the items on this scale. In addition, the effect size was calculated for these items.

Researchers scored the *Life Decisions and Everyday Choices* item responses dichotomously, which required a chi-square analysis. In addition to the chi-square statistic, an effect size was completed for each of the items.

Differential Item Functioning. The differential item functioning analysis provided an indication of whether there were systematic differences in how people responded to questions based on which type of respondent answered the item. Rather than include the theoretical methods and their processes in this section, a more detailed description of differential item functioning (DIF) is provided in Appendix A. Also included in this appendix are descriptions of the statistical methods used to assess DIF.

Mantel-Haenszel analysis. This study included a DIF analysis for each item of the three scales. The Mantel-Haenszel (M-H) analysis was completed using the jMetrik (Meyer, 2014b) software for the *Community Inclusion* scale, and SPSS, version 20 for the *Life Decisions* and *Everyday Choices* scales. This researcher used the jMetrik software

for the *Community Inclusion* scale because SPSS was unable to calculate the M-H statistic with polytomous data.

The effect size was also calculated for each of the Mantel-Haenszel analyses. For the polytomous data, jMetrik reported a converted Standardized Mean Difference (SMD) effect size (Meyer, 2014a), referred to as sP-DIF. Meyer noted that the interpretation for this effect size is: "A" or negligible DIF has an sP-DIF of less than |.05|, "B" or slight/moderate DIF when sP-DIF is greater than |.05| but less than |.10|. Whereas "C" items have an sP-DIF of greater than |.10|.

For the *Life Decision* and *Everyday Choices* scales, Educational Testing Services developed criterion to interpret the ln(α) with the Mantel-Haenszel (Zwick, 2012). To interpret this, "A" items have ln(α) of less than |.43|, whereas "B" items have an ln(α) of greater than |.43|, but less than |.64| (Zwick, 2012). Those items identified as "C" items have an effect size of greater than |.64| (Zwick, 2012).

The effect size sign is also important. A negative effect size indicated that the item favored self-respondents (or the reference group), while a positive sign showed that the proxy-respondents (or the focal group) tended to answer with a higher response than the self-respondents (Osterlind & Everson, 2009).

For all of these effect size methods, "A" items did not require any additional attention and did not show practical DIF. "B" items had some DIF, but usually only required an item review to confirm that the items did not appear to be problematic. With "C" items, Osterlind and Everson (2009) recommended not using these items.

Logistic regression analysis. With the logistic regression analyses, Zumbo's (1999) hierarchical model was used for the analysis. The SPSS DIFLRT macro was

employed to complete the statistical tests for all of the items. In addition, a test of the model fit and an effect size was calculated. Jodoin and Gierl's (2001) ΔR^2 effect size was used to interpret the magnitude of the model fit. With this effect size measure, type "A" items were those with a ΔR^2 of less than .035, and type "B" items were those where the ΔR^2 was above .035 and less than .070. Finally, type "C" items were those where the ΔR^2 was above .070.

It was necessary to conduct thick matching with the *Community Inclusion* scale due to the large score ranges for each of the items. The largest response fluctuation was for the *shopping* item, which spanned from 0 to 99. The smallest range response was for the *eating out* question, which extended from 0 to 55. One concern related to these large ranges was that multiple cells would not be populated in the Mantel-Haenszel contingency table. Donoghue and Allen (1993) noted that when a cell has a zero frequency it is excluded from the analysis, which reduces statistical power. As a result, Donoghue and Allen recommended using the percent of the total sample method with short assessments. With this method, item totals (both for self- and proxy-respondents) were combined and equal groupings were formed. For the *shopping*, *errands,* and *eating out* items, the new compressed range was four points. For the *entertainment* question, the item responses were compressed to three points. The same thick matching procedures were used both for the Mantel-Haenszel and logistic regression analyses.

For the *Life Decisions* scale, the scores were compressed again using the percent of the total score method. This compressed score was used as the matching score for both the Mantel-Haenszel and the logistic regression analyses. Donoghue and Allen

(1993) indicated that this method was also acceptable for use with short scales and protected against Type I error rates.

Uni-dimensional analysis. The purpose of a uni-dimensional analysis was to verify that there was only one latent factor. Harrington (2008) noted that when researchers developed a scale it was important to assess if the scale was uni-dimensional. To assess whether the scales had one underlying factor for self- and proxy-reports, a uni-dimensional analysis was conducted on each group for the three scales. The uni-dimensional analyses included the use of Item Response Theory for Patient Reported Outcomes (Cai, Thissen, & du Toit, 2012 [IRTPRO]). All of the latent variables and intercepts were free to vary in order to assess uni-dimensionality. For the *Community Inclusion* scale, the extraction method was a graded response due to the polytomous response options for this scale. With the *Life Decisions* and *Everyday Choices* scales, the extraction method was 2 PL, since the item-level data were dichotomous.

Measurement invariance. When researchers compared groups on an outcome, they assumed that measurement invariance would hold, but in reality this assumption was infrequently evaluated (Vandenburg & Lance, 2000). The crux of measurement invariance was that regardless of group membership, all people interpreted both the question and the underlying latent trait in a similar manner (van de Schoot, Lugtig, & Hox, 2012). When considering measurement invariance, three hierarchical approaches were evaluated: (a) configural, (b) metric, and (c) scalar invariance. *Configural invariance* or weak factorial invariance tests whether groups assessed the underlying trait as uni-dimensional with the same items manifesting a single dimension in both groups. In this model, both the intercepts and factor loadings were allowed to be free. In

comparing these different models, configural invariance had the fewest restrictions. In *metric invariance*, the factor loadings were set to be equal for each item between the groups, but the intercepts were allowed to vary (van de Schoot et al., 2012). As an example, the item *shopping* was selected, and the intercept was fixed between the self- and proxy-reports, then the *errands* question was fixed between self- and proxy-respondents. This analysis matched all of the intercepts for all of the items. Van de Schoot et al. (2012) noted that the purpose of this test was to determine if people in the two groups associated the same meaning to the latent construct. Metric invariance then had more restrictions than configural invariance. In *scalar invariance* or strong factorial invariance, the factor loadings and intercept estimates were set to be equal between the groups (Vandenberg & Lance, 2000). This model had the most restrictions and was the most constrained model fit described here.

To complete the analysis of measurement invariance, this researcher first established that configural invariance was met before proceeding to metric invariance. Then, once metric invariance was met scalar, invariance was tested. The three models were then nested and their results were compared to determine the best-fitting model.

Measurement invariance analysis. To evaluate the three hierarchical models of measurement invariance, the IRTPRO (Cai, Thissen, & du Toit, 2012) software was used and the various constraints were placed on each of the models. Goodness of fit indices were calculated for each of the models with all of the scales.

Chapter V: Results

The analyses are presented based on each scale. Within this grouping, the results were conveyed in the order of the research questions for this paper. To review, the first question considered whether the internal consistency reliabilities were adequate ($\alpha \geq .70$) for both the self- and proxy-respondents on all three scales. The second question examined the degree of item impact for self- and proxy-respondents. The purpose of the item impact analysis was to explore the extent to which the various scales showed practical differences between these two groups (i.e., self- versus proxy-respondents). The third question considered the extent to which items exhibited differential item functioning (DIF) using both the Mantel-Haenszel and logistic regression analyses. The general hypothesis was that all items except the *entertainment* item in the *Community Inclusion* scale would not show a meaningful difference. The fourth question bore in mind whether the latent structures of the scales were uni-dimensional. Finally, the last question examined whether the scales showed configural, metric, or scalar measurement invariance.

Community Inclusion Scale

The specific hypotheses for this scale included that the internal consistency reliability would be acceptable ($\alpha \geq .70$) for both the self- and proxy-respondents. For the item impact, it was believed that people with ID/DD and proxy-respondents would have similar means for *Community Inclusion,* indicating that there would not be any practical differences between the two groups. For the DIF analysis, it was hypothesized that the items related to *shopping, errands,* and *eating out* would not present DIF, but that the item related to *entertainment* would show meaningful DIF. For the uni-dimensional

hypotheses, it was believed that the two samples would be uni-dimensional, and finally that the two samples would demonstrate metric measurement invariance.

Reliability. The internal consistency reliabilities for self-respondents did not show evidence of adequate reliability ($\alpha = .59$) when the entire range of response options was used, whereas the proxy-reports approached acceptability ($\alpha = .69$). This indicates that for the self-respondents, 59% of score variance was true score variance for the entire response option. For proxy-respondents, 69% of score variance was true score variance. When calculating this statistic, there were no missing values in any of the questions, since this was a scoring requirement for this scale. The use of all four items resulted in the highest coefficient alpha for both self- and proxy-respondents.

The corrected item-total correlation, which captures the item's ability to discriminate between high and low performers, was acceptable for self-respondents and proxy-respondents for the entire range of response options. For self-respondents, the corrected item-total correlation ranged from $r = .32$ to .41. For the proxy-respondents, the corrected item-total correlation ranged from $r = .40$ to .55.

A second analysis was completed once the item responses were compressed using the percent total method. In using this compressed scale, the internal consistency reliability improved for both the self and proxy-respondents. The self-respondents returned a coefficient alpha of $\alpha = .66$, with $\alpha = .76$ for the proxy-respondents. In addition, all items equally contributed to this finding, meaning that the removal of items did not improve the reliability coefficient.

Item impact. Table 23 includes the distribution and summary statistics for the item scores. Each of the items showed a floor effect where the sample distribution was

greater at the lowest scores. The distributions for the scale scores were not normal, but the independent samples *t*-test was robust to violations of a normal distribution when the sample size was large and the two groups had nearly the same numbers in each group (Huck, 2009).

In considering the item related to *shopping*, self-respondents ($M = 4$, $SD = 4.4$, $N = 4{,}280$) reported more shopping experiences in the last month than did those individuals for whom the proxy-responded on their behalf ($M = 2.9$, $SD = 3.9$, $N = 4{,}153$). This mean difference of 1.11 was significantly different: $t(8{,}431) = 12.16$, $p < .001$, 95% CI [.93, 1.29]. The effect size ($d = .27$) was small. This finding did not support the hypothesis, since self-respondents reported a modest increase in shopping experiences over proxy-respondents.

For the question related to going out for *errands*, self-respondents ($M = 2.9$, $SD = 3.6$, $N = 4{,}280$) reported more opportunities than individuals with proxy-respondents ($M = 2.2$, $SD = 3.1$, $N = 4{,}153$). The mean difference (.69) between these groups was significant: $t(8{,}431) = 9.18$, $p < .001$, 95% CI [.54, .83]. The effect size was small at $d = .20$, which indicated that 14.7% of the score distribution for these two groups did not overlap. These results did not support the hypothesis, since there was a small effect in favor of self-respondents.

The item related to *entertainment* revealed that self-respondents ($M = 2.3$, $SD = 3.0$, $N = 4{,}280$) reported slightly fewer opportunities than individuals for whom a proxy responded ($M = 2.4$, $SD = 3.2$, $N = 4{,}153$) on their behalf. The *t*-te*st* revealed that these mean differences (.11) were not statistically significant: $t(8{,}431) = -1.59$, $p = .11$, 95%

CI [-.24, .025]. The effect size was trivial (d = -.03). These results did support the hypothesis, since the findings did not provide any practical significance.

Eating out captured the number of opportunities to visit a restaurant or coffee house in the last month. Self-respondents (M = 3.7, SD = 4.3, N = 4,280) reported more experiences than people who had a proxy respond for them (M = 2.6, SD = 3.4, N = 4,153). This difference between the groups was statistically significant: t(8,431) = 12.24, p < .001, 95% CI [.87, 1.21]. The effect size was small at d = .28. The findings did not support the hypothesis, because small practical differences were identified with respect to eating out for these two groups.

In sum, three out of four items showed significant mean differences between self- and proxy-respondents for the *Community Inclusion* scale. The items related to *shopping*, *errands*, and *eating out* were meaningful, whereas the *entertainment* item did not reveal any practical differences between the groups. When considering practical significance, the three items related to shopping, errands, and eating out had small effect sizes.

Differential item functioning. For the *shopping*, *errands*, and *eating out* questions, the hypothesis was that while controlling for overall community inclusion, there would be no relationship between who responded and their experiences with *shopping*, running *errands*, or *eating out*. However, for the *entertainment* item, it was anticipated that when controlling for overall community inclusion, DIF would be present both from a statistical and practical perspective. For each item, the analysis for the Mantel-Haenszel was presented first, followed by the regression models.

Table 23

Mean Comparison of Items for the Community Inclusion Scale

Scale Variable	*N*	*M*	Median	Skew	*SD*	*df*	*t*-statistic	*p* - value	Cohen's *d*
Shopping[a]	8433					8431	12.63	.000***	.265
Self	4280	4.0	4.0	5.1	4.4				
Proxy	4153	2.9	2.0	6.1	3.9				
Errands[b]	8433					8431	9.18	.000***	.201
Self	4280	2.9	2.0	5.3	3.6				
Proxy	4153	2.2	1.0	4.1	3.1				
Entertainment[c]	8433					8358.46	-1.59	.111	-.032
Self	4280	2.3	2.0	4.7	3.0				
Proxy	4153	2.4	2.0	3.8	3.2				
Eating Out[d]						8431	12.24	.000***	.262
Self	Self	3.7	3.0	3.7	4.3				
Proxy	Proxy	2.6	2.0	3.0	3.4				

Note. Shopping[a] indicates the question "In the past month, did you (did this person) go shopping? If yes, how many times?", Errands[b] corresponds to the question "In the past month, did you (did this person) go out on errands or appointments? If yes, how many times?" Entertainment[c] represents the question "In the past month, did you (did this person) go out for entertainment? If yes, how many times?", Eating Out[d] indicates the question "In the past month, did you (did this person) go out to a restaurant or coffee shop? If yes, how many times?"

*$p < .10$. **$p < .05$. ***$p < .001$.

Shopping. Adjusting for overall community inclusion, significant differences were found between self- and proxy-respondents and their shopping experiences: $\chi^2_{M\text{-}H}(3, N = 8{,}433) = 61.34$, $p < .001$. The sP-DIF effect size was .12, 95% CI [.09, .16], which pointed to minimal practical significance for this finding. This indicated that when comparing responses based on the overall community inclusion, both groups responded in a similar way to this item.

The results for the ordinal regression are in Table 24. The analysis of any DIF revealed a $\Delta X^2(2, N = 8{,}433) = 122.29$, $p < .001$, but the effect size was trivial at $\Delta R^2 = .006$. The results comparing model fit for uniform DIF reveal $\Delta X^2(1, N = 8{,}433) = 89.53$, $p < .001$. The effect size for this was minor ($\Delta R^2 = .004$). Finally, the model fit to assess non-uniform DIF identified $\Delta X^2(1, N = 8{,}433) = 32.76$, $p < .001$, and the effect size was again negligible ($\Delta R^2 = .002$).

These findings taken together suggest that while the results were statistically significant, they do not represent practically significant results. These outcomes then support the hypothesis that there was negligible DIF evident in this item. This indicated that the item can be used as written.

Errands. While controlling for the overall score, responses were different based on who answered the question: $\chi^2_{M\text{-}H}(3, N = 8{,}433) = 23.44$, $p < .001$. The effect size for this item was small at .09 (95% CI [.05, .13]).

Table 25 reports the results of this ordinal regression. In comparing the model to assess any DIF, $\Delta X^2(2, N = 8{,}433) = 42.82$, $p < .001$. The effect size revealed $\Delta R^2 = .003$ or a trivial finding. To determine whether uniform DIF existed the examination for the model fit revealed that $\Delta X^2(1, N = 8{,}433) = 32.57$, $p < .001$ with an effect size

of $\Delta R^2 = .003$. While this result was statistically significant, the effect size indicated that this finding was not practically significant. Finally, to assess non-uniform DIF, the model fit was statistically significant: $\Delta X^2(1, N = 8{,}433) = 10.25$, $p = .001$. However, the effect size was trivial at $\Delta R^2 < .001$.

The findings for both the Mantel-Haenszel and the ordinal regression for the errands item uncovered statistical significance. However, when considering the trivial effect sizes for both analyses, these findings were not practically important. These results then supported the hypothesis that no meaningful DIF existed for this item and that it can be used as written.

Entertainment. While controlling for the overall community inclusion with the total score of this scale, responses were different based on who answered the question: $\chi^2_{M\text{-}H}(2, N = 8{,}433) = 220.60$, $p < .001$. When considering the effect size for this item, it showed a large effect (sP-DIF = -.21; 95% CI [-.24, -.18]), which indicated that this item favored proxy-respondents.

Table 26 includes the results of this ordinal regression. In comparing the model to assess any DIF: $\Delta X^2(2, N = 8{,}433) = 210.250$, $p < .001$. The effect size revealed that $\Delta R^2 = .019$, which was just below Jodoin and Gierl's (2001) effect size criteria. When determining whether uniform DIF exists, the *entertainment* ordinal regression models revealed $\Delta X^2(1, N = 8{,}433) = 210.04$, $p < .001$, with an effect size of $\Delta R^2 = .019$. While this result was statistically significant, the effect size indicated that this finding was below practical significance. Finally, when assessing non-uniform DIF, $\Delta X^2(1, N = 8{,}433) = .21$, $p = .65$, which was not statistically significant.

Table 24

Ordinal Regression Results Community Inclusion Shopping Item

	Step 1			Step 2			Step 3		
Variable	β	*S. E.*	Wald	β	*S. E.*	Wald	β	*S. E.*	Wald
Shopping = 1	3.058 [2.94, 3.17]	.061	2535.08***	3.230 [3.10, 3.36]	.064	2509.41***	3.518 [3.35, 3.68]	.084	1757.15***
Shopping = 2	5.089 [4.94, 5.24]	.078	4225.99***	5.285 [5.12, 5.44]	.082	4134.78***	5.58 [5.39, 5.78]	.099	3169.24***
Shopping = 3	7.22 [7.03, 7.42]	.099	5343.81***	7.42 [7.22, 7.62]	.102	5286.12***	7.702 [7.48, 7.93]	.115	4470.39***
Self				.426 [.338, .514]	.045	89.96***	1.03 [.81, .89]	.116	79.72***
Self * Total							-.103 [.00, -.14]	.018	33.23***
Total Scale	.821	.012	4709.95***	.814 [.79, .84]	.012	4600.19***	.865 [.84, .89]	.015	3252.55***
R^2		.321			.321			.322	
-2 Log Likelihood		809.81			809.81			777.05	

Note. R^{2*} the Nagelkerke R^2 statistic is reported; Step 1[a] is the model: logit (Item Response) = $\beta_0 + \beta_1$(Total Score); Step 2[b] represents the model in which logit (Item Response) = $\beta_0 + \beta_1$(Total Score) + β_2(Respondent Type); Step 3[c] indicates the model in which logit (Item Response) = $\beta_0 + \beta_1$(Total Score) + β_2(Respondent Type) + β_3(Total Score * Respondent Type).
*$p < .10$. **$p < .05$. ***$p < .001$.

Table 25

Ordinal Regression Results Community Inclusion Errands Item

	Step 1			Step 2			Step 3		
Variable	β	*S. E.*	Wald	β	*S. E.*	Wald	β	*S. E.*	Wald
Errands = 0	1.31 [1.21, 1.40]	.049	710.21***	1.38 [1.29, 149]	.051	732.86***	1.49 [1.37, 1.62]	.063	569.61***
Errands = 1	2.92 [2.81, 3.03]	.057	2663.70***	3.01 [2.89, 3.13]	.059	2598.46***	3.13 [2.99, 3.27]	.071	1971.52***
Errands = 2	4.92 [4.78, 5.07]	.073	4498.18***	5.01 [4.86, 5.16]	.075	4425.06***	5.13 [4.96, 5.29]	.084	3727.47***
Self				.24 [.16, .33]	.043	32.54***	.52 [.33, .65]	.090	28.56***
Self * Total							-.052 [-.083, -.02]	.016	10.39***
Total Scale	.61 [.59, .63]	.01	3861.27***	.60 [.58, .62]	.01	3729.79***	.624 [.60, .65]	.012	2547.08***
R^2		.481			.484			.484	
-2 Log Likelihood		894.18			861.61			851.36	

Note. R^{2*} the Nagelkerke R^2 statistic is reported; Step 1[a] is the model: logit (Item Response) = $\beta_0 + \beta_1$(Total Score); Step 2[b] represents the model in which logit (Item Response) = $\beta_0 + \beta_1$(Total Score) + β_2(Respondent Type); Step 3[c] indicates the model in which logit (Item Response) = $\beta_0 + \beta_1$(Total Score) + β_2(Respondent Type) + β_3(Total Score * Respondent Type).
*$p < .10$. **$p < .05$. ***$p < .001$.

Table 26

Ordinal Regression Results Community Inclusion Entertainment Item

	Step 1			Step 2			Step 3		
Variable	β	*S. E.*	Wald	β	*S. E.*	Wald	*β*	*S. E.*	Wald
Ent[a] = 0	1.35 [1.25, 1.45]	.049	746.89***	1.17 [1.07, 1.27]	.051	532.12***	1.15 [1.03, 1.28]	.063	336.11***
Ent[a] = 1	3.52 [3.39, 3.64]	.063	3073.52***	3.38 [3.25, 3.50]	.064	2766.86***	3.36 [3.22, 3.50]	.074	2065.48***
Self				-.65 [-.74, -.56]	.046	204.65***	-.66 [-.89, -.49]	.103	45.23***
Self * Total							.01 [.64, -.03]	.01	.21
Total Scale	.47 [.456, .492]	.009	2675.29***	.51 [.49, .53]	.010	2772.99***	.50 [.48, .53]	.012	1631.63***
R^2		.366			.385			.385	
-2 Log Likelihood		1081.06			871.02			870.81	

Note. R^{2*} the Nagelkerke R^2 statistic is reported; Step 1[a] is the model: logit (Item Response) = $\beta_0 + \beta_1$(Total Score); Step 2[b] represents the model in which logit (Item Response) = $\beta_0 + \beta_1$(Total Score) + β_2(Respondent Type); Step 3[c] indicates the model in which logit (Item Response) = $\beta_0 + \beta_1$(Total Score) + β_2(Respondent Type) + β_3(Total Score * Respondent Type).
Ent[a] represents Entertainment.
*$p < .10$. **$p < .05$. ***$p < .001$.

The findings for the Mantel-Haenszel and the ordinal regression revealed differences in the effect size. This item showed that when compared with others who expressed similar amounts of community inclusion, proxy-respondents were significantly more likely to respond that individuals with ID/DD whom they represent participated in more entertainment. Since the logistic regression effect size tended to underestimate magnitude (Hidalgo & López-Piña, 2004), the Mantel-Haenszel effect sizes were likely more accurate. Also, the ΔR^2 effect size was one of the largest effect sizes based on the logistic regression analyses. Taken together, these findings do support the hypothesis that there was practical significance for this item.

Eating out. While controlling for overall community inclusion, item-responses were not different based on who answered the question: $\chi^2_{\text{M-H}}(3, N = 8{,}433) = 3.60$, $p = .06$. This supported the hypothesis that there was no uniform DIF present with this item. In considering the effect size, it was sP-DIF = .02, 95% CI [-0.01, 0.06]. The magnitude associated with this question was not meaningful.

Table 27 provides the summary results of the ordinal regression analysis for this item. In comparing the model to assess any DIF, $\Delta X^2(2, N = 8{,}433) = 45.93$, $p < .001$. The effect size revealed that $\Delta R^2 = .003$ or a minimal finding. To determine whether uniform DIF existed, the regression model found that $\Delta X^2(1, N = 8{,}433) = 29.1$, $p < .001$, with an effect size of $\Delta R^2 = .002$. While this result was statistically significant, the effect size indicated that this finding was not practically significant. Finally, to assess non-uniform DIF, the model fit showed $\Delta X^2(1, N = 8{,}433) = 16.84$, $p < .001$,

which was statistically significant. However, the effect size was inconsequential at $\Delta R^2 = .001$.

The findings of the Mantel-Haenszel and logistic regression were different. However, when considering the chi-square statistic's sensitivity to sample size, it was not unusual to find significance due to the large sample size. The effect size for both analyses did not constitute practical significance. As such, the results supported the hypothesis of no practical DIF evident in this question.

In sum, several of the items on the *Community Inclusion* scale did not provide evidence of DIF, while the *entertainment* item did show significant DIF. ETS developed guidelines for the Mantel-Haenszel and provided test developers with guidance about how to treat items with DIF. For items with small or no DIF, ETS recommended no further action. For the *shopping, errands*, *and eating out* items, they were identified with type "A" or minimal DIF, and these items can be used as written. However, for items with severe DIF (noted as type "C"), ETS recommended not using the items as written due to the large amount of DIF detected in the analyses. Based on these analyses, the *entertainment* item presented with a substantial amount of DIF and should not be used as it is currently written.

Table 27

Ordinal Regression Results Community Inclusion Eating Out Item

	Step 1			Step 2			Step 3		
Variable	β	*S. E.*	Wald	β	*S. E.*	Wald	β	*S. E.*	Wald
Eating Out = 0	2.06 [1.96, 2.17]	.054	1473.02***	2.14 [2.03, 2.25]	.056	1461.10***	2.30 [2.17, 2.44]	.069	1117.19***
Eating Out = 1	4.79 [4.65, 4.95]	.075	4043.53***	4.89 [4.74, 5.05]	.078	3932.65***	5.06 [4.89, 5.24]	.089	3225.68***
Eating Out = 2	7.36 [7.16, 7.56]	.100	5368.12***	7.45 [7.25, 7.65]	.102	5313.24***	7.61 [7.39, 7.82]	.11	4804.00***
Self				.24 [.15, .32]	.044	29.22***	.62 [.416, .82]	.10	35.87***
Self * Total							-.067 [-.09, -.03]	.016	17.08
Total Scale	.81 [.79, .83]	.012	4657.06***	.80 [.78, .83]	.012	4555.92***	.84 [.81, .86]	.014	3456.41***
R^2		.614			.616			.617	
-2 Log Likelihood		904.49			875.40			858.56	

Note. R^{2*} the Nagelkerke R^2 statistic is reported; Step 1[a] is the model: logit (Item Response) = $\beta_0 + \beta_1$(Total Score); Step 2[b] represents the model in which logit (Item Response) = $\beta_0 + \beta_1$(Total Score) + β_2(Respondent Type); Step 3[c] indicates the model in which logit (Item Response) = $\beta_0 + \beta_1$(Total Score) + β_2(Respondent Type) + β_3(Total Score * Respondent Type).
*p < .10., **p <.05., ***p < .001.

Uni-dimensional analysis. For this analysis, a graded response extraction method was used, since this item is polytomous. The findings revealed that for both of the self- and proxy-reports, the scale was uni-dimensional. In considering proxy-respondents, the root mean square error of approximation (RMSEA) for both the G^2 (RMSEA = .02) and the χ^2 (RMSEA = .03) showed very good model fit with RMSEA below .05. Table 28 presents the graded model item parameter estimates for the self-respondents. The fit was also good for the self-respondents with both the G^2 and χ^2 RMSEA less than .05 (.03 for both). Table 28 includes the parameter estimates for the proxy-respondents.

Table 28

Uni-dimensional Evaluation with Item Parameter Estimates: Community Inclusion Scale for Self- and Proxy-Respondents

Item	*a*	*s.e.*	b_1	*s.e.*	b_2	*s.e.*	b_3	*s.e.*
Self-Respondents								
Shopping	2.72	.12	-.21	.02	.51	.02	1.12	.03
Errands	1.63	.06	-.86	.03	.06	.03	1.12	.04
Entertainment	1.21	.05	-.98	.05	.66	.04		
Eating Out	2.23	.09	-.68	.03	.39	.02	1.35	.04
Proxy-Respondents								
Shopping	1.57	.07	-1.13	.04	-.05	.03	1.11	.04
Errands	1.28	.06	-1.78	.07	-.59	.04	.78	.04
Entertainment	1.15	.05	-.98	.05	.74	.04		
Eating Out	1.57	.07	-1.62	.06	-.10	.03	1.24	.05

Measurement invariance. Table 29 provides the fit indices for the configural, metric, and scalar analyses for the *Community Inclusion* scale. In considering the fit between the configural and metric models, the chi-square statistic was large, χ^2 (4, N = 8,433) = 32.54, $p < .001$, which would indicate a poor fit between these models. However, in considering the sample size and that the chi-square statistic became very large as sample size increased, this finding was not unusual. In considering the BIC, the

metric model fit better ($BIC_{configural}$ = 80517.59; BIC_{metric} = 80513.96). An examination of the model fit between the metric and scalar analyses revealed that there was not a good congruence between these models: (χ^2 (11, N = 8,433) = 250.28, $p < .001$. A comparison of the BIC revealed that the metric model had the lowest value (BIC_{metric} = 80513.96, BIC_{scalar} = 80664.81). This finding uncovered that these two groups have different intercepts. In order to compare these two groups (such as in a *t*-test), it was necessary to equate the group intercepts.

Table 29

Configural, Metric, and Scalar Measurement Invariance: Community Inclusion Scale

Invariance Type	-2 log likelihood	AIC	BIC
Configural	80228.31	80292.31	80517.59
Metric	80260.85	80316.85	80513.96
Scalar	80511.13	80545.13	80664.81

In sum, the hypothesis related to internal consistency was supported for the proxy-respondents, but not for the self-respondents. There was item impact with small effect sizes for three of the questions, but the *entertainment* item did not reveal any statistical differences between these groups. The DIF analysis supported the hypothesis that only the *entertainment* question would show DIF. This scale was uni-dimensional for both self- and proxy-respondents. The measurement invariance results indicated that it was not possible to compare these two groups, since they had different intercepts.

Life Decisions Scale

To review, the hypotheses for this scale were that both self- and proxy-respondents would have acceptable internal consistency reliabilities for applied research. When assessing item impact, it was believed that the items would not show practically significant differences between the two groups, meaning that the effect size would be less

than .20. For the DIF analysis, none of the items would show significance. With respect to uni-dimensionality, it was hypothesized that the scale had one underlying latent trait for both the self- and proxy-respondent groups. Finally, it was hypothesized that a comparison of configural, metric, or scalar invariance would show that metric invariance would hold.

Reliability. The internal consistency reliabilities for self-respondents ($\alpha = .74$) and proxy-respondents ($\alpha = .73$) were acceptable. Thus, true score variance for self-respondents and proxy-respondents, was 74% and 73% respectively. The use of all the items resulted in the highest coefficient alpha for both groups of respondents.

There were a number of missing values for both the self- and proxy-reports. To score this scale, the average was calculated based on the number of responses to questions. To calculate the internal consistency reliability, the statistic used only those cases where a response was available for all of the items (listwise deletion). As such, 2,654 (53.9%) of the cases were excluded from the analysis with the self-respondents, and 3,973 (82.1%) of the cases were excluded for proxy-respondents. An analysis of the missing data in the *staff at home* item revealed that most individuals responded that this item did not apply, meaning that they did not have staff in their homes. The choice of *day activity* also had a large number of non-responses (46.6%), as did the question related to choice of *day staff* (48.5%).

When considering item discrimination for this scale, the corrected item-total correlation was moderate for both self- and proxy-respondents. For the self-respondents, the corrected item-total correlation ranged from $r = .44$ to .58. This indicated that these items discriminate well between those individuals who reported overall high life

decisions and responded that they experienced choice. Similar findings were evident for the proxy-respondents, as the corrected item-total correlation ranged from $r = .44$ to .55.

Item impact. Table 30 presents the contingency table for this scale. Since the scoring for this scale used dichotomous responses, a chi-square statistic was calculated to understand the group differences.

The item analysis related to choice of *home* revealed that the self-respondents indicated a greater proportion of choice about their living arrangements (2,825 out of 4,754 or 59.4%) than proxy-respondents (916 out of 4,432 or 20.6%). The findings for this analysis were statistically significant: χ^2 (1, $N = 9,186$) = 1427.16, $p < .001$. The effect size ($\phi = .39$) showed a moderate association in favor of self-respondents when considering choice of home. This finding did not support the hypothesis, since this item showed both statistical significance and meaningfulness.

The item analysis examined the relationship between choice of *roommates* and who responded. The self-respondents reported more choice (2,449 out of 4,731 or 51.8%) than proxy-respondents (996 out of 4,647 or 21.4%). This finding was statistically significant: χ^2 (1, $N = 9,378$) = 928.04, $p < .001$. The effect size ($\phi = .31$) demonstrated a moderate association in favor of self-respondents with respect to choice of roommate. This finding did not support the hypothesis that there would be no practical significance between these groups.

The item pertaining to choice of *staff at home* revealed that the self-respondents again indicated that they had more choice (2,030 out of 3,057 or 66.4%) than individuals for whom a proxy responded (813 out of 1,441 or 56.4%). This difference was statistically significant: χ^2 (1, $N = 4,498$) = 41.99, $p < .001$. The effect size ($\phi = .097$)

showed a weak association between choice of staff and who responded. Of note, 5,270 (54%) of the sample did not respond to this item and were not included in the analysis. This finding supported that there was statistical significance; however, when considering the effect size these differences were not meaningful. Specifically, the results supported the hypothesis, because no worthwhile differences existed between these groups.

The item reflecting choice of *day activities* showed that self-respondents reported more choice (2,742 out of 3,799 or 72.2%) than proxy-respondents (615 out of 1,421 or 43.3%) in selecting their day activities. As such, this finding was statistically significant χ^2 (1, N = 5,220) = 376.26, $p < .001$. The effect size ($\phi = .27$) showed a small relationship between choice of day program and who responded. This result did not support the hypothesis, since practical significance was found. Of note, 4,548 (46.6%) participants in this sample did not respond to this question and were excluded from analysis.

For the last question, choice of *day staff*, the findings supported the hypothesis. Self-respondents reported greater choice (2,179 out of 3,578 or 60.8%) when compared with proxy-respondents (775 out of 1,449 or 53.5%). The significance was: χ^2 (1, N = 5,027) = 23.4, $p < .001$. However, the effect size ($\phi = .068$) indicated a minor relationship, meaning that this item was not practically significant. There were 4,741 (48.5%) excluded from the analysis due to no response. These results supported the hypothesis that no meaningful differences existed between the two groups.

In review, statistically significant differences in choice were found between these two groups of respondents for the *Life Decisions* scale. For all of these items, self-respondents indicated more choice than proxy-reports. However, when considering the

practical significance of these findings, two items -- choice of *staff at home* and choice of *day staff* -- had minor effect sizes, which indicated that these differences were slight and not meaningful.

Table 30

Contingency Table for Life Decisions Items

Item		Respondent Type		Total	χ^2	*p* - value
		Self	Proxy			
Home[a]	None	1929	3516	5445		
	Some	2825	916	3741	1427.16	.000***
Total		4754	4432	9186		
Roommate[b]	None	2282	3651	5933		
	Some	2449	996	3445	928.04	.000***
Total		4731	4647	9378		
Staff at home[c]	None	1027	628	1655		
	Some	2030	813	2843	41.99	.000***
Total		3057	1441	4498		
Day Activity[d]	None	1057	806	1863		
	Some	2742	615	3357	376.26	.000***
Total		3799	1421	5220		
Day Staff[e]	None	1399	674	2073		
	Some	2179	775	2954	23.4	.000***
Total		3578	1449	5027		

Note. Home[a] relates to the question: "Who chose (or picked) the place where you live?" Roommate[b] indicates the question: "Did you chose (or pick) the people you live with (or did you choose to live by yourself)?" Staff at home[c] corresponds to the question: "Do you choose (or pick) who helps you at home?" Day Activity[d] relates to the combined questions: "Who chose (or picked) the place where you work?" and "Who chose (or picked) where you go during the day?" Day Staff[e] is the combined questions "Do you choose (or pick) who helps you at your job?" and "Do you choose (or pick) who helps you during the day?"

*$p < .10$. **$p < .05$. ***$p < .001$.

Differential item functioning. For the items related to choice of *home*, *roommates*, *staff at home*, choice of *day activity,* and choice of *day staff*, the hypotheses were the same for all of the questions. The hypothesis was that while

controlling for overall life decisions, the two groups would not have dissimilar expected scores on the items in this scale when matched with others of a similar ability.

Home. The Mantel-Haenszel statistic was reviewed first, followed by the logistic regression model fit. Conditioning on overall life decisions, significant differences were found between self-respondents and proxy-respondents and their choice of home: χ^2 (1, N = 9,186) = 50.09, $p < .001$. The effect size for this item was medium (ln(α) = -.56), which indicated practical significance. This question moderately favored self-respondents who were more likely to report choice in selecting where they lived when compared with others who had the same life decisions total score (OR_{MH} = .57).

Table 31 includes the results for the logistic regression. For the analyses of any DIF, a comparison of models revealed a model fit of $\Delta X^2(2, N = 9{,}186) = 67.04$, $p < .001$, but the effect size was negligible at $\Delta R^2 = .004$. The evaluation of the uniform DIF model showed significant findings: $\Delta X^2(1, N = 9{,}186) = 56.09$, $p < .001$. However, the effect size for this did not show practical significance at $\Delta R^2 = .003$. Finally, the assessment of non-uniform DIF indicated significance at $\Delta X^2(1, N = 9{,}186) = 10.95$, $p = .001$.

While both the Mantel-Haenszel and logistic regression analyses showed significant results, the interpretation of the effect size for these measures gave different findings. In considering effect size, Hidalgo and López-Piña (2004) completed a simulation study to compare the Mantel-Haenszel with logistic regression. These authors found that the logistic regression models tended to

underestimate the effect size, which likely occurred for this item and many other items in this analysis. When considering the Mantel-Haenszel results, this item did appear to moderately favor self-respondents. These findings did not support the hypothesis of no DIF present.

Roommate. When considering the Mantel-Haenszel statistic for this question, the findings uncovered significant results: $\chi^2(1, N = 9{,}378) = 7.95, p = .005$. However, the effect size ($\ln(\alpha) = .22$) was negligible.

Table 32 reports the summary findings of the logistic regression analysis. When considering the change in the -2 Log likelihood, the model detecting any DIF was significant at $\Delta X^2(2, N = 9{,}378) = 12.57, p < .002$. The effect size of $\Delta R^2 = .011$ did not constitute practical significance for this model. In evaluating the model for uniform DIF, the model comparison demonstrated significance: $\Delta X^2(1, N = 9{,}378) = 11.89, p = .001$. The effect size measure indicated only a trivial effect ($\Delta R^2 = .011$). The analysis of non-uniform DIF revealed no significance: $\Delta X^2(1, N = 9{,}378) = .67, p = .41$.

Table 31

Logistic Regression Results: Choice of Home Item in Life Decisions Scale

	Step 1			Step 2			Step 3		
Variable	β	*S. E.*	Wald	β	*S. E.*	Wald	β	*S. E.*	Wald
Home = 0	5.27 [5.02, 5.52]	.128	1696.02***	5.53 [5.27, 5.80]	.135	1668.64***	6.10 [5.64, 6.57]	.24	653.17***
Total Scale	8.95 [8.55, 9.35]	.21	1893.90***	8.75 [8.34, 9.15]	.21	1796.05***	9.72 [8.95, 10.49]	.39	610.65***
Self				.59 [.43, .74]	.08	55.95***	1.45 [.89, 2.02]	.28	26.10***
Interaction							-1.48 [-2.39, -.57]	.47	10.15***
R^2		.767			.770			.771	
-2 Log Likelihood		348.42			292.32			281.38	

Note. R^{2*} the Nagelkerke R^2 statistic is reported; Step 1[a] is the model: logit (Item Response) = $\beta_0 + \beta_1$(Total Score); Step 2[b] represents the model in which logit (Item Response) = $\beta_0 + \beta_1$(Total Score) + β_2(Respondent Type); Step 3[c] indicates the model in which logit (Item Response) = $\beta_0 + \beta_1$(Total Score) + β_2(Respondent Type) + β_3(Total Score * Respondent Type).
*$p < .10$. **$p < .05$. ***$p < .001$.

Table 32

Logistic Regression Results: Choice of Roommates Item in Life Decisions Scale

	Step 1			Step 2			Step 3		
Variable	β	*S. E.*	Wald	β	*S. E.*	Wald	β	*S. E.*	Wald
Roommates = 0	5.87 [5.59, 6.14]	.14	1806.11***	5.78 [5.51, 6.06]	.140	1708.82***	5.66 [5.26, 6.06]	.21	763.44***
Total Scale	9.15 [8.74, 9.56]	.21	1909.38***	9.31 [8.89, 9.73]	.22	1850.49***	9.08 [8.42, 9.75]	.34	723.69***
Self				-.274 [-.43, -.12]	.08	11.81	-.49 [.077, -1.04]	.28	3.14
Interaction							.36 [.41, -.49]	.44	.69
R^2		.761			.762			.762	
-2 Log Likelihood		374.97			363.07			362.39	

Note. R^{2*} the Nagelkerke R^2 statistic is reported; Step 1[a] is the model: logit (Item Response) = $\beta_0 + \beta_1$(Total Score); Step 2[b] represents the model in which logit (Item Response) = $\beta_0 + \beta_1$(Total Score) + β_2(Respondent Type); Step 3[c] indicates the model in which logit (Item Response) = $\beta_0 + \beta_1$(Total Score) + β_2(Respondent Type) + β_3(Total Score * Respondent Type).
*$p < .10$. **$p < .05$. ***$p < .001$.

Taken together, the Mantel-Haenszel and an examination of the logistic models showed statistical significance for the presence of uniform DIF. However, when the effect size was considered, this finding was not meaningful. This did support the proposed hypothesis.

Staff at home. The Mantel-Haenszel revealed that when compared with others who had similar levels of life decisions, the choice of day staff was different depending on who answered the question: $\chi^2(1, N = 4{,}498) = 74.71$, $p < .001$. The log-odds ratio for this question showed that proxy-respondents were 2.5 times more likely to respond that the individuals had a choice in selecting their day staff. The effect size for this was also large ($\ln(\alpha) = .92$).

The logistic regression models also indicated significance, and the effect size reflected a minor effect for the uniform model. Table 33 provides the summary statistics for this analysis. With respect to any DIF, the chi-square showed statistical significance at $\Delta X^2(2, N = 4{,}498) = 82.29$, $p < .001$ and an effect size of $\Delta R^2 = .013$. The model assessing uniform DIF uncovered a model fit of $\Delta X^2(1, N = 4{,}498) = 80.57$, $p < .001$, and an effect size of $\Delta R^2 = .013$. Finally, the model that assessed non-uniform DIF was $\Delta X^2(1, N = 4{,}498) = 1.72$, $p = .19$.

Both the Mantel-Haenszel and logistic regression models were consistent; this item showed uniform DIF, with proxy-respondents more likely to indicate choice when compared with others who had similar overall levels of life decisions. The logistic regression model showed a slight effect size, but the Mantel-Haenszel indicated a large effect size. Since the Mantel-Haenszel effect size was likely more

accurate (Hidalgo & López-Piña, 2004), this finding was not consistent with the hypothesis of no DIF found.

Day activity. This Mantel-Haenszel analysis revealed that there was uniform DIF present in this item, $\chi^2(1, N = 5{,}220) = 81.04$, $p < .001$, with a large effect also detected ($\ln(\alpha) = -.78$). This item appeared to favor self-respondents, since self-respondents were 46% more likely to respond that they had a choice about their day activity when compared with others who had similar life decision choices.

Table 34 lists comparable findings from the examination of the logistic regression analysis. When considering any DIF, $\Delta X^2(2, N = 5{,}220) = 82.56$, $p < .001$ and $\Delta R^2 = .013$. This indicated that there was a slight effect associated with any DIF present in this question. In evaluating uniform DIF, $\Delta X^2(1, N = 5{,}220) = 81.79$, $p < .001$ with an effect size change of .013. For the non-uniform DIF model fit, the chi-square revealed that $\Delta X^2(1, N = 5{,}220) = .76$, $p = .38$ was non-significant.

Both statistical methods found that uniform DIF was present in this item. The Mantel-Haenszel indicated that this effect was large, while the logistic regression model showed that this was a minor effect. Since the Mantel-Haenszel was likely more accurate (Hidalgo & López-Piña, 2004), this finding did not support the original hypothesis that there would be no DIF detected in this item.

Table 33

Logistic Regression Results: Choice of Staff at Home Item in Life Decisions Scale

	Step 1			Step 2			Step 3		
Variable	β	*S. E.*	Wald	β	*S. E.*	Wald	β	*S. E.*	Wald
Staff at home = 0	3.33 [3.09, 3.56]	.12	775.00***	2.96 [2.71, 3.20]	.13	560.44***	3.12 [2.77. 3.47]	.18	304.41***
Total Scale	7.78 [7.33, 8.22]	.23	1186.02***	8.34 [7.85, 8.82]	.25	1143.59***	8.77 [7.94, 9.58]	.42	436.64***
Self				-.96 [-1.17, -.74]	.11	74.27***	-.67 [-1.15, -.18]	.25	7.21***
Interaction							-.67 [-1.69, .34]	.52	1.69
R^2		.67			.68			.68	
-2 Log Likelihood		244.09			163.52			161.80	

Note. R^{2*} the Nagelkerke R^2 statistic is reported; Step 1[a] is the model: logit (Item Response) = $\beta_0 + \beta_1$(Total Score); Step 2[b] represents the model in which logit (Item Response) = $\beta_0 + \beta_1$(Total Score) + β_2(Respondent Type); Step 3[c] indicates the model in which logit (Item Response) = $\beta_0 + \beta_1$(Total Score) + β_2(Respondent Type) + β_3(Total Score * Respondent Type).
*$p < .10$. **$p < .05$. ***$p < .001$.

Table 34

Logistic Regression Results: Choice of Day Activity Item in Life Decisions Scale

	Step 1			Step 2			Step 3		
Variable	β	*S. E.*	Wald	β	*S. E.*	Wald	β	*S. E.*	Wald
Day Activity= 0	2.49 [2.32, 2.65]	.08	878.81***	2.94 [2.74, 3.15]	.104	803.81***	2.95 [2.65, 3.26]	.155	363.04***
Total Scale	6.46 [6.13, 6.79]	.17	1426.69***	6.33 [5.99, 6.67]	.172	1348.93***	6.36 [5.73, 6.98]	.320	393.33***
Self				.75 [.58, .93]	.089	71.98***	.770 [.41, 1.13]	.185	17.34***
Interaction							-0.04 [-.78, .71]	.380	.01
R^2		.598			.609			.609	
-2 Log Likelihood		529.21			456.84			456.83	

Note. R^{2*} the Nagelkerke R^2 statistic is reported; Step 1[a] is the model: logit (Item Response) = $\beta_0 + \beta_1$(Total Score); Step 2[b] represents the model in which logit (Item Response) = $\beta_0 + \beta_1$(Total Score) + β_2(Respondent Type); Step 3[c] indicates the model in which logit (Item Response) = $\beta_0 + \beta_1$(Total Score) + β_2(Respondent Type) + β_3(Total Score * Respondent Type).
*$p < .10$. **$p < .05$. ***$p < .001$.

Day staff. The Mantel-Haenszel revealed that there was a significant difference between who responded and their indication of choice while conditioned on the total *Life Decisions* scale. For this analysis, the Mantel-Haenszel showed that $\chi^2(1, N = 5{,}027) = 77.27, p < .001$. The log odds indicated that this item favored proxy-respondents ($OR_{MH} = 2.19$). The effect size for this item was large ($\ln(\alpha) = .79$).

For the logistic regression analysis, represented in Table 35, the model revealed that there was DIF present, but the effect size based on Jodoin and Gierl's (2001) criteria was negligible. For the model assessing any DIF, $\Delta X^2(2, N = 5{,}027) = 86.75, p < .001$ and $\Delta R^2 = .014$. The uniform DIF model showed statistical significance ($\Delta X^2(1, N = 5{,}027) = 67.08, p < .001$), but a slight effect size with $\Delta R^2 = .014$. Finally, the model for non-uniform DIF indicated no significance at $\Delta X^2(1, N = 5{,}027) = .026, p = .87$.

These two models detected uniform DIF in this item. However, when considering the effect size, the Mantel-Haenszel indicated large DIF while the logistic regression model just missed the cut-off for small magnitude. Given that the Mantel-Haenszel model had more power for detecting uniform DIF and the logistic regression model underestimates DIF, it was concluded that this item does exhibit DIF. This finding was not consistent with the research hypothesis.

In considering the DIF analyses, the general hypothesis of no DIF present was not supported in several of the items. In considering the ETS guidelines, the choice of *home* item did show DIF, however this item was classified as type "B" DIF based on the moderate effect size. ETS advised that this type of question should undergo a review, but that items of this nature could be retained in the scale. The item related to choice of *roommate* was classified as type "A", meaning that this item can be used

as written. Finally, the last three items were all identified as type "C" DIF. This indicated that these items should not be used as written and should be re-written.

Uni-dimensional analysis. The findings of this analysis revealed that the uni-dimensional model fit better for the proxy-respondents than the self-respondents. The root mean square error of approximation (RMSEA) for proxy-respondents was .05 for both the G^2 and the χ^2 statistics. This finding suggests that there was good fit for a uni-dimensional model. However, for the self-respondents, the RMSEA was .07 for the G^2 and the χ^2 statistics, which was at the upper limits of good model fit (Steiger, 2007). Table 36 lists the parameter estimates for these two groups.

Table 35

Logistic Regression Results: Choice of Day Activity Staff Item in Life Decisions Scale

	Step 1			Step 2			Step 3		
Variable	β	*S. E.*	Wald	β	*S. E.*	Wald	β	*S. E.*	Wald
Day Staff = 0	2.97 [2.77, 3.16]	.100	877.74***	2.60 [2.39, 2.81]	.106	605.08***	2.59 [2.29, 2.88]	.15	296.29***
Total Scale	6.23 [5.89, 6.57]	.173	1301.23***	6.66 [6.30, 7.03]	.186	1288.19***	6.62 [5.99, 7.25]	.321	426.52***
Self				-.84 [-1.02, -.66]	.093	81.67***	-.866 [-1.27, -.46]	.205	17.84***
Interaction							0.06 [-.71, .83]	.393	.03
R^2		.554			.568			.568	
-2 Log Likelihood		397.97			311.24			311.22	

Note. R^{2*} the Nagelkerke R^2 statistic is reported; Step 1[a] is the model: logit (Item Response) = $\beta_0 + \beta_1$(Total Score); Step 2[b] represents the model in which logit (Item Response) = $\beta_0 + \beta_1$(Total Score) + β_2(Respondent Type); Step 3[c] indicates the model in which logit (Item Response) = $\beta_0 + \beta_1$(Total Score) + β_2(Respondent Type) + β_3(Total Score * Respondent Type).
*$p < .10$. **$p < .05$. ***$p < .001$.

Table 36

Uni-dimensional Evaluation with Item Parameter Estimates: Life Decisions Scale for Self- and Proxy-Respondents

Item	*a*	*s.e.*	*c*	*s.e.*	*b*	*s.e.*
Self-Respondents						
Home	1.99	.10	.62	.05	-.31	.03
Roommates	2.40	.13	.12	.05	-.05	.02
Staff at Home	2.36	.16	1.35	.09	-.57	.03
Day Activity	1.67	.09	1.46	.06	-.87	.04
Day Staff	1.47	.08	.70	.05	-.48	.03
Proxy-Respondents						
Home	3.14	.29	-2.93	.22	.93	.03
Roommates	3.10	.28	-2.81	.21	.91	.03
Staff at Home	1.84	.17	.15	.08	.38	.05
Day Activity	1.21	.11	.03	.07	-.03	.06
Day Staff						

Measurement invariance. Table 37 supplies the configural, metric and scalar fit indices for the *Life Decisions* scale. In considering the fit between the models, metric invariance was the better fit than the other two models. The difference in fit between the configural and metric models was χ^2 (5, N = 9768) = 37.18, $p < .001$, which would suggest a poor fit, but the large sample size inflated this chi-square statistic. In considering the BIC, the metric model was the smallest (BIC_{metric} = 36874.18, $BIC_{configural}$ = 36882.93). When comparing the scalar versus the metric models, the metric model again was the best fit with the lowest BIC (BIC_{metric} = 36874.18, BIC_{scalar} = 37110.29). Therefore, the metric invariance model fit these data well. These findings suggested that any between-group differences on the *Life Decisions* scale were not uniquely related to the construct measured in this scale (Wicherts & Dolan, 2010).

Table 37

Configural, Metric, and Scalar Measurement Invariance: Life Decisions Scale

Invariance Type	-2 log likelihood	AIC	BIC
Configural	36680.82	36724.82	36882.93
Metric	36718.00	36752.00	36874.18
Scalar	37000.04	37024.04	37110.29

In summary, these findings showed that the internal consistency reliability for both groups was acceptable for applied research. The general item impact analysis revealed that the hypothesis of no meaningful differences between the groups was upheld for two questions – *staff at home* and *day staff*. The remaining items showed small to moderate effect sizes. When considering the DIF analysis, the general hypothesis of no meaningful DIF was consistent with only one item -- choice of *roommate*. The remaining items showed moderate or large DIF for these groups. The uni-dimensional analysis revealed that the proxy-respondents fit the uni-dimensional model well. However, the self-respondents were at the upper limits of an acceptable fit for the uni-dimensional model. In considering the configural, metric and scalar invariance, the metric model fit these data well.

Everyday Choices Scale

The general hypothesis for this scale was that the internal consistency reliability would be acceptable for applied research; there would not be any meaningful differences between self- and proxy-respondents with respect to item impact. In addition, there would not be any meaningful DIF present in this scale. The scale would be uni-dimensional for both the self- and proxy-respondents. Finally, the measurement invariance would show that the metric model fit the data well.

Reliability. The internal consistency reliabilities for self-respondents was lower than adequate ($\alpha = .48$), whereas the proxy-respondents was acceptable ($\alpha = .77$). This revealed that for self-respondents, 48% of score variance was true score variance; for the proxy-respondents, 77% of score variance was true score variance. For both groups of respondents, the percentage of missing cases was less than 5% (2.7% for self-respondents and 2.6% for proxy-respondents), which was not considered problematic.

In examining the influence of removing items and potential impact on the reliability coefficient, the question: "Do you choose what you buy with your spending money?" adversely impacted the reliability coefficient for the self-respondents. If this item was removed, then the reliability coefficient for the self-respondents improved to .55, but did not improve the scale for proxy-respondents ($\alpha = .75$).

When considering item discrimination or the relationship between individual items and the overall scale score, the corrected item-total correlation was low for the self-respondents and moderate for the proxy-respondents. This correlation ranged from .18 to .43 for the self-respondents, while it ranged from .54 to .65 for the proxy-respondents. As with the previous analysis, the item with the lowest item-total correlation was related to: "Do you choose what you buy with your spending money?" (NASDDDS & HSRI, 2003, p. 46) at $r = .18$ for self-respondents and .54 for proxy-respondents. This paper used *schedule*, *free time,* and *buy* to refer to these items respectively rather than writing them out.

Item impact. Table 38 presents the findings for these analyses. A chi-square statistic was used for all the analyses, since researchers scored these item responses dichotomously.

Table 38

Contingency Table for Everyday Choices Items

Item		Respondent Type		Total	χ^2	*p* - value
		Self	Proxy			
Schedule[a]	None	392	1624	2016		
	Some	5021	3318	8336	1081.39	.000***
Total		5413	4942	10,355		
Free time[b]	None	145	969	1114		
	Some	5250	3968	9218	768.99	.000***
Total		5395	4937	10,332		
Buy[c]	None	141	1247	1388		
	Some	5264	3636	8900	1155.58	.000***
Total		5405	4883	10,288		

Note. Schedule[a] represents the question: "Who decides your daily schedule (like when you get up, when to eat, when to go to sleep)?" Free time[b] indicates the question: "Who decides how you spend your free time (when you are not working, in school or at the day program)?" Buy[c] is the question associated with: "Do you choose what you buy with your spending money.

*$p < .10$. **$p < .05$. ***$p < .001$.

When considering the individual questions, individuals who responded for themselves indicated greater choice with respect to the *schedule* proportion of choice (5,021 out of 5,413 or 92.7%) as compared to proxy-respondents (3,318 out of 4,942 or 67%). When comparing the groups, it was found that the proportions were significantly different between the groups: χ^2 (1, N = 10,355) = 1081.49, $p < .001$. The effect size (ϕ =.32) showed a moderate association between choice of schedule and who responded. These findings do not support the hypothesis, since a meaningful difference was indicated for these groups.

The item reflecting choice of *free time* showed that self-respondents (5,250 out of 5,395 or 97.3%) reported more choice when compared with the proxy group (3,968 out of 4,937 or 80.4%). This finding was statistically significant: χ^2 (1, N = 10,332) = 768.99, p

< .001. The effect size ($\phi = .27$) indicated a small relationship between choice of free time and who responded. As there was small practical significance, these data did not support the hypothesis.

Finally, the item related to choice of what to *buy* indicated that self-respondents (5,264 out of 5,405 or 97.4%) also demonstrated a higher proportion of choice over individuals for whom a proxy responded (3,636 out of 4,883 or 74.4%). This finding was again statistically significant: χ^2 (1, N = 10,288) = 1155.58, $p < .001$. The effect size (ϕ = .34) showed a moderate relationship between choice of what to buy and the respondent type. Due to a moderate effect, these findings did not support the hypothesis.

All of these items on the *Everyday Choices* scale showed small to moderate practical differences between the groups. Self-respondents reported more choice on all three items of this scale. In considering how self-respondents answered these questions, there was little variability in the responses they offered. To clarify, self-respondents answered that they had choice in their *schedule* 96% of the time, choice of *free time* 97% of the time and choice of what to *buy* 97.4% of the time.

With so many individuals reporting that they had choice on all three items, there was (obviously) not a lot of variation in how they responded. This lack of variation in the self-respondents then contributed to the low internal consistency reliability for this scale.

Differential item functioning. For these three items, it was hypothesized that none of them would show any meaningful DIF, either non-uniform or uniform in nature. Each of the three item results was reviewed. The conclusions for this series of findings should be taken with caution, however, since the internal consistency reliabilities were so low for the self-respondents and there was little variability in responses.

Schedule. The Mantel-Haenszel analysis revealed that uniform DIF was detected at a small level. The statistical test returns a $\chi^2(1, N = 10{,}355) = 14.49$, $p < .001$ and a magnitude of $\ln(\alpha) = .48$. This item also appears to favor proxy-respondents ($OR_{MH} = 1.61$).

For the logistic model comparisons in Table 39, these findings revealed that there was DIF present, but the magnitude was trivial. For the model that assesses any DIF, $\Delta X^2(2, N = 10{,}355) = 54.13$, $p < .001$ and $\Delta R^2 = .004$. The uniform DIF model revealed $\Delta X^2(1, N = 10{,}355) = 12.05$, $p = .001$ with a magnitude of $\Delta R^2 = .001$. Finally, the non-uniform model uncovered statistical significance of $\Delta X^2(1, N = 10{,}355) = 42.08$, $p < .001$ and a minor effect of $\Delta R^2 = .003$.

Given these findings, the hypothesis for this item was again not supported. The Mantel-Haenszel detected DIF that was moderate in nature. The logistic regression model also found DIF, but the magnitude was trivial. Since the logistic regression model effect size tended to underestimate magnitude (Hidalgo & López-Piña, 2004), it was likely that this item DIF had small practical significance in favor of proxy-respondents.

Free time. The Mantel-Haenszel analysis revealed that there was DIF detected, $\chi^2(1, N = 10{,}355) = 29.89$, $p < .001$ and that the magnitude was large ($\ln(\alpha) = .94$). The odds ratio showed that this item tended to favor proxy-respondents ($OR_{MH} = 2.56$) when compared with others who had similar everyday choices.

Table 39

Logistic Regression Results: Choice of Schedule in Everyday Choices Scale

	Step 1			Step 2			Step 3		
Variable	β	*S. E.*	Wald	β	*S. E.*	Wald	β	*S. E.*	Wald
Schedule = 0	10.58 [9.95, 11.22]	.32	1073.99***	10.69 [10.05, 11.34]	.33	1065.84***	8.58 [7.92, 9.24]	.34	646.44***
Total Scale	15.68 [14.79, 16.56]	.45	1218.96***	15.98 [15.07, 16.91]	.47	1167.37***	12.94 [12.02, 13.86]	.47	756.62***
Self				-.30 [-.53, -.08]	.12	6.91	-6.76 [-8.61, -4.91]	.942	51.54***
Interaction							9.09 [6.45, 11.73]	1.35	45.47***
R^2		.85			.85			.86	
-2 Log Likelihood		356.57			349.58			275.64	

Note. R^{2*} the Nagelkerke R^2 statistic is reported; Step 1[a] is the model: logit (Item Response) = $\beta_0 + \beta_1$(Total Score); Step 2[b] represents the model in which logit (Item Response) = $\beta_0 + \beta_1$(Total Score) + β_2(Respondent Type); Step 3[c] indicates the model in which logit (Item Response) = $\beta_0 + \beta_1$(Total Score) + β_2(Respondent Type) + β_3(Total Score * Respondent Type).
*$p < .10$. **$p < .05$. ***$p < .001$.

The logistic regression also found DIF present in this item, but the magnitude of the statistic was trivial. Table 40 includes the model results. The model for no DIF was $\Delta X^2(2, N = 10{,}332) = 40.67$, $p < .001$ with $\Delta R^2 = .005$. The model to detect uniform DIF indicated that $\Delta X^2(1, N = 10{,}332) = 40.67$, $p < .001$, and the magnitude was small: $\Delta R^2 = .004$. Finally, the non-uniform DIF model was $\Delta X^2(1, N = 10{,}332) = 8.99$, $p = .003$. The effect for this was trivial at $\Delta R^2 = .001$.

Taken together, this item did not support the original hypothesis of no meaningful DIF present in this item. Based on the Mantel-Haenszel statistic, DIF was present and strongly favored the proxy-respondents for this item.

Buy. The Mantel-Haenszel analysis found that DIF existed and that this item favored self-respondents. The overall test statistic was $\chi^2(1, N = 10{,}288) = 54.47$, $p < .001$ with a magnitude of $\ln(\alpha) = -.88$. This would indicate that this item showed large practical significance. The log odds ($OR_{MH} = .42$) for this item exhibited that it favored self-respondents. The findings of this statistic should be taken with some caution, however, as the Breslow-Day homogeneity of the odds, was significant ($\chi^2 = 9.8$, $p = .007$). This would suggest that there was another factor impacting these results. For example, the poor reliability index and item-discrimination measures could be impacting these findings. The reliability coefficient for self-respondents was very low ($\alpha = .48$), whereas the proxy-respondents sample was better ($\alpha = .77$). This finding related to the precision of this scale, with the scale performing poorly for self-respondents. In addition, the item-discrimination index was very low ($r = .18$), for self-respondents, which could impact Type I error (Magis & DeBoeck, in press).

Table 40

Logistic Regression Results: Choice of Free time Item in Everyday Choices Scale

	Step 1			Step 2			Step 3		
Variable	β	*S. E.*	Wald	β	*S. E.*	Wald	β	*S. E.*	Wald
Free time= 0	4.06 [3.77, 4.37]	.153	705.23***	4.16 [3.85, 4.47]	.159	685.49***	3.88 [3.54, 4.22]	.175	494.03***
Total Scale	10.77 [10.13, 11.41]	.33	1091.6***	11.53 [10.79, 12.26]	.375	946.29***	10.78 [9.96, 11.59]	.418	663.39***
Self				-1.00 [-1.36, -.64]	.183	29.78***	-2.32 [-3.32, -1.31]	.512	20.49***
Interaction							2.48 [.77, 4.19]	.872	8.11***
R^2		.838			.842			.843	
-2 Log Likelihood		132.27			100.59			91.71	

Note. R^{2*} the Nagelkerke R^2 statistic is reported; Step 1[a] is the model: logit (Item Response) = $\beta_0 + \beta_1$(Total Score); Step 2[b] represents the model in which logit (Item Response) = $\beta_0 + \beta_1$(Total Score) + β_2(Respondent Type); Step 3[c] indicates the model in which logit (Item Response) = $\beta_0 + \beta_1$(Total Score) + β_2(Respondent Type) + β_3(Total Score * Respondent Type).
*$p < .10$. **$p < .05$. ***$p < .001$.

Table 41

Logistic Regression Results: Choice of What to Buy in Everyday Choices Scale

	Step 1			Step 2			Step 3		
Variable	β	*S. E.*	Wald	β	*S. E.*	Wald	β	*S. E.*	Wald
Buy = 0	3.53 [3.30, 3.77]	.119	886.15***	3.57 [3.33, 3.80]	.120	876.25***	3.49 [3.22, 3.75]	.132	695.85***
Total Scale	7.82 [7.48, 8.17]	.175	2008.72***	7.36 [7.01, 7.71]	.179	1690.64***	7.22 [6.83, 7.62]	.201	1288.72***
Self				1.14 [.91, 1.38]	.120	90.04***	.76 [1.02, 1.90]	.306	6.20***
Interaction							.58 [-.26, 1.42]	.429	1.82
R^2		.708			.719			.719	
-2 Log Likelihood		737.78			640.73			638.87	

Note. R^{2*} the Nagelkerke R^2 statistic is reported; Step 1[a] is the model: logit (Item Response) = $\beta_0 + \beta_1$(Total Score); Step 2[b] represents the model in which logit (Item Response) = $\beta_0 + \beta_1$(Total Score) + β_2(Respondent Type); Step 3[c] indicates the model in which logit (Item Response) = $\beta_0 + \beta_1$(Total Score) + β_2(Respondent Type) + β_3(Total Score * Respondent Type).

*$p < .10$. **$p < .05$. ***$p < .001$.

The logistic regression for this item revealed statistical significance, but again no practical significance. The no DIF model showed $\Delta X^2(2, N = 10{,}288) = 98.92$, $p < .001$ and an effect size of $\Delta R^2 = .008$. The uniform DIF model revealed $\Delta X^2(2, N = 10{,}288) = 97.06$, $p < .001$, $\Delta R^2 = .008$. Table 41 includes the summary statistics for this item.

The findings of this scale showed that two items, choice of *schedule* and *free time*, favored the proxy-respondents, but the choice of what to *buy* favored the self-respondents. The Mantel-Haenszel detected large effect sizes for these items, whereas the logistic regression detected only minor magnitude.

In considering the DIF analyses for these items, the hypothesis of no DIF evident was not supported for all three items. The ETS guidelines suggest that the item related to *schedule* was a type "B" classification and should be reviewed, but could be retained as written. However, for the choice of *free time* and choice of what to *buy*, both items revealed significant DIF and met the type "C" ETS criterion. As such, it is difficult to defend using these items as they are currently written and should be eliminated or undergo new item writing.

Uni-dimensional analysis. For both groups, the uni-dimensional model fit was very good and Table 42 lists the item parameter estimates for both groups. For the self-respondents, the RMSEA for both the G^2 and χ^2 was .00, which represented a perfect fit. This finding was very unusual given this large sample size and was likely due to a lack of variability in this scale. The findings for the proxy-respondents showed that the uni-dimensional model also fit the data well, with the RMSEA for both the G^2 and χ^2 at .01.

Table 42

Uni-dimensional Evaluation with Item Parameter Estimates: Everyday Choices Scale for Self- and Proxy-Respondents

Item	*a*	*s.e.*	*c*	*s.e.*	*b*	*s.e.*
Self-Respondents						
Schedule	2.82	.24	4.87	.30	-1.73	.05
Free time	6.08	.27	12.19	.51	-2.01	.04
Buy	1.27	.14	4.34	.17	-3.42	.27
Proxy-Respondents						
Schedule	3.70	.22	1.82	.11	-.49	.02
Free time	6.93	.67	6.11	.60	-.88	.02
Buy	2.10	.16	1.78	.07	-.85	.03

Measurement invariance. In examining the entire sample together, the metric model showed the best fit for these data, and Table 43 presents the summary statistics. The fit between the configural and metric models showed χ^2 (3, N = 10,426) = 3.32, p =.34, which indicated a good fit even with this large sample size. The BIC shows that the metric model fits better ($BIC_{configural}$ = 18600.30, BIC_{metric} = 18575.86). In considering whether scalar or metric fits best, the comparison showed that χ^2 (3, N = 10,426) = 167.44, p < .001. The BIC results also supported this finding: BIC_{metric} = 18575.86, BIC_{scalar} = 18715.55. Therefore, the hypothesis was not upheld, as the scalar invariance model did not fit these data.

Table 43

Configural, Metric, and Scalar Measurement Invariance: Everyday Choices Scale

Invariance Type	-2 log likelihood	AIC	BIC
Configural	18470.77	18498.77	18600.30
Metric	18474.09	18496.09	18575.86
Scalar	18641.53	18657.53	18715.55

In sum, many of the findings related to the *Everyday Choices* scale were not supported with this data. For example, the internal consistency reliability was very low for self-respondents. This low reliability coefficient was likely related to the limited variability in the scores for self-respondents. In addition, the internal consistency reliability improved for self-respondents when one of the three items was removed from the scale. Next, the item impact results indicated that all of the items showed small to moderate differences between these groups. The DIF analyses revealed that all of the items showed large DIF. Two of the items favored proxy-respondents, while the choice of what to *buy* favored self-respondents. With respect to the uni-dimensional nature of this scale, both groups showed that this model fit well, but for the self-respondents it was a perfect fit. Finally, it was determined that it was not possible to compare scores between these two groups, since the analysis confirmed metric measurement invariance.

Chapter IV: Discussion

This research examined the internal structure of three objective quality of life measures from the National Core Indicators (NCI). The NCI has emerged as a leading measure to assess quality of life for people with intellectual and developmental disabilities (ID/DD) (Moseley, 2011). Policy makers use the data and reports from the NCI to: (a) make policy decisions, (b) direct program planning, and (c) manage quality (Moseley, 2011). However, researchers have not completed extensive evaluations related to the validity of this measure with respect to self and proxy-respondents. Since policy makers use this tool to make key decisions about programs for people with ID/DD, it is important that the inferences of the instrument are valid. In addition, this dissertation explores the use of self- and proxy-responses to understand if these groups respond to questions in a systematically different way.

This is the first study to use differential item functioning to identify how well items are operating for self- and proxy-respondents in the field of ID/DD. In effect, the purpose of this dissertation is to determine if the type of respondent (i.e., self versus proxy) could confound the results of the objective quality of life measures for the NCI. This study also examines measurement invariance to determine if it is possible to make comparisons between self- and proxy-respondents.

The underlying theme of these analyses is fairness in testing. Huggins (2013) notes that differential item functioning evaluates bias at the item-level, whereas measurement invariance assesses equitability at the test level. It is important to establish both item- and test-level equivalence when there are different groups of test takers.

Two broad sections organize the following discussion. The first portion presents the findings associated with each of the research questions and situates these results in the current literature both on the NCI in general, and with respect to the use of self- and proxy-respondents. The second portion considers the implications of the research findings and how this study is useful to policy makers and social scientists examining quality of life for people with ID/DD.

Situating Findings within the Literature

Reliability. The first research question addresses the internal consistency reliability for all three scales when considering both self- and proxy-respondents. The NCI (2011c) reported the coefficient alpha for the three scales as $\alpha = .64$ for the *Community Inclusion* scale, $\alpha = .80$ for the *Life Decisions* scale, and $\alpha = .76$ for the *Everyday Choices* scale. NCI researchers calculated these reliability indices based on the entire sample. This study, however, considered the internal consistency reliability for self- and proxy-respondents. These analyses yielded different findings than what the NCI reported on the *Community Inclusion* and *Everyday Choices* scales. Specifically, the proxy-respondents had higher internal consistency reliabilities for the *Community Inclusion* scale ($\alpha = .69$) and for the *Everyday Choices* scale ($\alpha = .77$). For this sample of self-respondents, the internal consistency reliability for the *Community Inclusion* scale was $\alpha = .59$, and for the *Everyday Choices* scale it was $\alpha = .48$. Nimon, Zientek, and Henson (2012) note that the reliability coefficient provides information about measurement error. Thompson, Johnstone, and Thurlow (2002) indicate that an assessment for people with disabilities should include a tolerance for error. It is possible that these lower reliability coefficients for self-respondents are then a reflection of the

additional error inherent in assessing individuals with ID/DD (Thompson, Johnstone & Thurlow, 2002), which in turn impacts the ability to detect differences for this group. When reliability is low, statistical findings become attenuated (Thorndike & Thorndike-Christ, 2010). This error then becomes embedded in the data and reduces the precision and power in a statistical analysis.

Each of the scales has a different justification for the low internal consistency reliabilities for these self-respondents. For the *Community Inclusion* scale, the internal consistency reliability improves when considering the compressed scale scores. Coefficient alpha increases for both self- and proxy-respondents to .66 and .76, respectively. This finding indicates that the use of the entire range of scores reduced the reliability coefficient with this sample, whereas the use of the compressed scale scores led to a higher coefficient alpha for this sample.

The *Everyday Choices* scale, however, has very low internal consistency reliability, likely due to the lack of variability in how self-respondents answered these questions. In examining the score distribution for self-respondents, 92% of self-respondents answer that they have choice in their *schedule*, 97% indicate that they have choice in their *free time,* and 97.4% have choice in what they *buy*. In other words, nearly all of the self-respondents select the same response to these questions, which results in a lower coefficient alpha (Thorndike & Thorndike-Christ, 2010).

In considering the literature associated with response concerns for people with ID/DD, a number of other possible reasons account for why self-respondents answer that they experienced more choice related to these items. For example, in the literature related to individuals with ID/DD, researchers identify several areas of concern when

gathering information from people with ID/DD that are related to acquiescence and recency.

The first, acquiescence, is when individuals reply 'yes' to any question regardless of content (Heal & Sigelman, 1995; Perry & Felce, 2002). It is possible that acquiescence impacted the findings for this scale, since most responded that they had choice. Another possible response bias is when people choose the last option rather than answering questions with a variety of all possible response choices (Heal & Sigelman, 1995; Perry & Felce, 2002). In the NCI, the last response option for all three questions is that individuals have choice.

Item impact. The next research question considers the differences in how self- and proxy-respondents answered each item on the three scales. However, the findings related to the measurement invariance assessment are important in understanding the item differences detected between self- and proxy-reports for the NCI. The measurement invariance findings indicate that it is not possible to compare self- and proxy-respondents without first equating scores. Since metric invariance holds for each of these three scales, it means that the intercepts are different for each respondent type. These differences between self- and proxy-respondents could then be an artifact of measurement and not necessarily true differences between these groups.

Multiple items on these scales show differences between self- and proxy-respondents, with self-respondents indicating greater community inclusion and choice than individuals for whom a proxy responded. While this research considers item-level comparisons, the overall findings are similar to Bonham et al.'s (2004) findings with the Ask Me! project. In this evaluation of several scales on the Ask Me! measure, Bonham et

al. (2004) found that self-respondents reported greater choice and social inclusion than proxy-respondents.

Differential item functioning. The third research question evaluates whether items functioned differently depending on who answers the item while conditioning or matching on the total score of the scale. Based on a previous pilot study that the author completed, the *entertainment* item in the *Community Inclusion* scale was believed to function differently. No other DIF studies had been identified with the items on the *Life Decisions* and *Everyday Choices* scales. The hypotheses related to these items were then exploratory in nature.

Community inclusion. In considering the *Community Inclusion* scale, the findings confirmed the hypotheses for this scale. Meaningful DIF with proxy-respondents answering higher than self-respondents is only present in the *entertainment* item when considering the Mantel-Haenszel findings. For the logistic regression analyses, none of the items show practical significance, but the *entertainment* item did have one of the largest effect sizes calculated. From a methodological perspective, Hidalgo and López-Piña (2004) conducted a study to compare the effect size detection for the Mantel-Haenszel and logistic regression. Hidalgo and López-Piña found that the logistic regression consistently returned low effect sizes. In comparing the findings for the Mantel-Haenszel and logistic regression, it is believed that the Mantel-Haenszel effect size is more accurate in detecting moderate to severe DIF.

ETS, a national testing organization, provides classification and recommendations when DIF is present. When considering the DIF analysis for the *entertainment* item, the effect size is large and classified as a type "C" item (Zwick, 2012). ETS recommends

that type "C" items should be re-written. One possible reason that this item may exhibit DIF is because it is more difficult for self-respondents to answer (Osterlind & Everson, 2009).

In considering how these findings could impact the reliability coefficient, it is possible that the DIF evident in the *entertainment* item could impact the precision of this scale. Thorndike and Thorndike-Christ (2010) note that when assessments are difficult for one group of people the reliability coefficient becomes reduced due to guessing.

Life decisions. In examining the DIF analysis for the *Life Decisions* scale, the findings do not support the general hypothesis of no DIF present for several of the items. For the question related to choice of *home*, self-respondents tend to reply that they have more choice than proxy-respondents when conditioned on the overall *Life Decisions* scale. In considering the ETS guidelines, this item is classified as type "B" DIF (Zwick, 2012). ETS recommends that these items undergo a review, but researchers may continue to use these items in the scale.

The item related to choice of *roommate* revealed a negligible amount of DIF in this sample. In considering the ETS criteria, this item is identified as a type "A" item since the effect size is so small. ETS indicates that type "A" items do not require any additional type of review, and Zwick (2012) notes that researchers may continue to use these items as written.

The analysis of the choice of *staff at home* item reveals considerable DIF that favors proxy-respondents. This finding means that when comparing self- and proxy-respondents at similar levels of life decisions, the proxy-respondents tend to answer that the people whom they represented have more choice than self-respondents. According to

the ETS criteria, this item falls into the type "C" category and should not be used as written (Zwick, 2012).

The analysis for the choice of *day activity* item also exhibits significant DIF. This question appears to favor self-respondents, meaning that when comparing self- and proxy-respondents who had similar levels of overall life decisions, the self-respondents tend to reply that they have more choice related to their day activity than proxy-respondents. Based on the ETS classification, this item is identified as type "C." Again, ETS recommends not using type "C" questions in their current form (Zwick, 2012). Similar to the other recommendations, it may be necessary to re-write this question.

The final item, choice of *day staff*, also demonstrates considerable DIF, with proxy-respondents identifying more choice when matched with self-respondents at the same level of overall life decisions. This item introduces a considerable amount of construct-irrelevant variance into the item. According to ETS criteria, this item is a type "C" item, which indicates researchers should not use the item as written (Zwick, 2012).

Everyday choices. In considering the DIF analyses for each of the items in this scale, the results should be interpreted with caution. When considering DIF, there are a number of factors that could impact the results. An important consideration is the matching criterion. In this case, the matching criterion is the total scale score. From the previous analysis, we know that this scale has questionable reliability and limited variability of scores for the self-reports. In addition, most of the self-respondents rated their choice on all of these items as high, meaning that most people indicated that they had choice in their everyday decisions. This finding indicates that the self-respondent

sample is very homogeneous and does not provide a very full range of responses for this scale.

When considering the DIF analysis, the findings reveal that DIF is present in these items. The item related to choice of *schedule* reveals that proxy-respondents tend to report more choice when matched with self-respondents who have similar levels of everyday choice. In considering the ETS recommendations, the effect size for this item places it within the type "B" category. This finding indicates that this item should undergo review, but could be maintained in the scale (Zwick, 2012).

For the item related to choice of *free time*, this analysis finds large DIF, which favors proxy-respondents. In applying the ETS criteria, this item is a type "C." As such, this item should also undergo re-writing (Zwick, 2012).

The final item, choice of what to *buy*, again shows significant DIF, with the self-respondents more likely to reply that they have greater choice. This item also meets the ETS type "C" classification and should be changed rather than used in the present form (Zwick, 2012).

The method of evaluating items through a DIF analysis is unique to this dissertation. Previous research used matched pairs of verbal individuals with ID/DD with proxies who knew them well. In other words, previous research conditions the analyses based on the relationship between the self- and proxy-respondent. A limitation with this type of research is that it requires a verbal individual with ID/DD, and it is not possible to include information about people with ID/DD who are not verbal. In using the DIF analysis, the comparisons between self- and proxy-respondents are based on the total

scores for each of the three scales. In doing this, it is possible to include representations of people who are unable to respond for themselves.

Uni-dimensional analyses. For all three scales, the uni-dimensional analyses of both self- and proxy-respondents indicate that these measures capture a single underlying trait for both groups. This single latent construct is important for a number of reasons. For example, an assumption of the internal consistency reliability is that scales measure a single latent trait (Thorndike & Thorndike-Christ, 2010). In addition, this finding provides additional validity evidence related to each of these scales. Researchers created these scales with the intention that they would each assess one underlying latent trait. These results then support that the scales measure one underlying trait for each of the respondent types.

Also, in considering the literature related to quality of life, researchers note that the concept of quality of life has multiple dimensions (Cummings, 1991; Cummins, 2005; Rapley, 2003; Schalock, 2000; Schalock & Felce, 2004). The uni-dimensional analyses findings confirm that the NCI has three one-dimensional scales that relate to quality of life.

Measurement invariance. For all three scales, metric invariance fits these data well. As such, the factor loadings are consistent for both self- and proxy-respondents, but the intercepts are different between these two groups. This invariance in the intercept means that these two groups have different latent means (Hortensius, 2012). Most researchers do not often assess this finding, but it is a necessary assumption of statistical analyses that compare two groups, such as with a *t*-test or *ANOVA*. If there are non-invariant intercepts, Wicherts and Dolan (2010) state that any differences in means

between groups could be due to measurement inconsistencies rather than true differences. These findings support Verdugo, Schalock, Keith, and Stancliffe's (2005) concerns about the valid use of comparing self- and proxy-responses.

There are several possible reasons why metric invariance occurred. The first is that several items on the three scales have DIF for this sample. Wicherts and Dolan (2010) note that non-invariant intercepts could be due to DIF in items. In writing items that are fair to both self- and proxy-respondents, this reduces the risk for a test or scale to function differently depending on who takes it. Wicherts and Dolan (2010) identify a second possible reason, which is that the two groups of respondents have differences that are not due to the between-group differences measured in the scale. In considering this possibility, the self- and proxy-respondents did have very different group characteristics. Race, ethnicity, level of ID, home type, and legal status are all statistically different for self- and proxy-respondents (see Tables 10 to 16), but only level of ID, home type, and legal status show practical significance. These group characteristics could then contribute to the intercept non-invariance in these two groups.

Since these scales do not provide evidence of scalar measurement invariance, it is difficult to interpret between-group differences on these scales (Millsap & Kwok, 2004). Psychometrists note that there are three options available when non-invariance occurs. The first is to use only those items shown to be invariant between groups (Millsap & Kwok, 2004). However, there are several problems with this option; for example, test length and content could change depending on the groups being compared (Millsap & Kwok, 2004). The second option is to discontinue use of the scale (Millsap & Kwok,

2004), but this is problematic from a practical sense, since a lot of time and money are invested in the use of the NCI.

The final option is to estimate the magnitude of invariance and account for it statistically (Millsap & Kwok, 2004; Wicherts & Dolan, 2010). In this process, the intercepts are transformed to be equal between these two groups for the items where the intercepts are not equivalent. Schmitt and Kuljanin (2008) describe a method of assessing partial invariance where intercepts with the largest differences between groups are allowed to be free. A researcher would continue to complete this until the model fit is stable. When the researcher identifies the partial invariance model, it is then possible to compare mean differences between groups.

Schalock (2010) indicates that proxy-respondents answer questions in a systematically different way than self-respondents. The findings of the measurement invariance evaluation for the three NCI measures support Schalock's view. In addition, these findings shed some additional light on Schalock's recommendations related to the use of proxies. First, Schalock indicates that when researchers use proxies, it is important to consider that proxies have different perspectives than self-respondents. The present findings also support this, because the determination of metric invariance shows that these two groups of respondents have different intercepts when the latent trait takes on a value of zero.

In addition, Schalock (2010) also recommended that researchers use statistical techniques to account for the different perspectives that proxy-respondents experience. Schalock's recommendations are consistent with Wicherts and Dolan's (2010) as well as Millsap and Kwok's (2004) guidance associated with metric invariance.

Implications of findings

This section first considers specific changes related to item and scale refinements for the NCI. Second, it regards fairness and ties these ideas more directly to social work and the principle of social justice. Third, this paper identifies policy considerations and fourth, study limitations. Finally, it explores future directions for research.

Item modifications. The differential item functioning analyses suggest that there are several items that should be re-written as they have an ETS classification of type "C." Type "C" items have very large effect sizes associated with the DIF analysis. These items include the *entertainment* item in the *Community Inclusion* scale, and the choices of *staff at home*, *day activity,* and *day staff* in the *Life Decisions* scale. In addition, the *free time* and choice of what to *buy* items in the *Everyday Choices* scale should be revised.

Since significant DIF is evident in several items, it is important to consider why this occurred. As a case in point, consider the *entertainment* item in the *Community Inclusion* scale. One possible explanation for this DIF result and the fact that the item favors proxy-respondents could be due to the use of the word "entertainment," as it may be too complex for individuals with ID/DD to understand. In developing the Ask Me! survey, Bonham et al. (2004) note that word choices should be limited to one or two syllables to aid in understanding for people with ID/DD. In addition, other researchers in the ID/DD field agree that questions and response options should be clear (Verdugo et al., 2005) and that questions should be conclusive (Heal & Sigelman, 1995; Perry & Felce, 2002). Haladyna, Downing, and Rodriguez (2002) also identify that the language in an assessment should be simple and understandable to the people who take it. As such, it seems plausible that the use of a four-syllable word (i.e., "entertainment") could add

some unintended difficulty to this item for the self-respondents, which could then lead to significant DIF.

To correct this, researchers could potentially re-write this item using simpler words to reflect entertainment; for example, "How often do you go out to do fun things (such as go to a movie, watch sports or play sports)?" Researchers could then field-test the item with both self- and proxy-respondents to collect item-response data with the revised question. With the revised item data, another DIF analysis could determine if the word changes resulted in reduced DIF for this item.

It is often hard to identify the actual reason for DIF. In considering the other type "C" items, plausible reasons for DIF are not as readily evident. For example, the item related to staff at home is more difficult for self-respondents. Frequently, individuals who receive support in their homes have more than one staff. As such, it may be more difficult for individuals with ID/DD to consider the varying degrees of choice they have in the staff who help. This item could be confusing for self-respondents (Verdugo et al., 2005).

One method for uncovering confusion in items is to use a cognitive interview process to understand how people (both self- and proxy-respondents) process questions and arrive at their answers. In this procedure, researchers ask respondents to talk through their problem-solving processes as they answer items. The researchers record the verbalizations and analyze them for themes. Johnstone, Bottsford-Miller, and Thompson (2006) note that this cognitive interview or think-aloud procedure is an effective way to understand how people process questions. In completing this type of qualitative analysis, it is possible to detect items that have ambiguous constructs, unclear language and

imprecise directions (Johnstone et al., 2006). Identifying and correcting these problematic elements can help reduce the construct-irrelevant variance detected with the DIF analysis.

From another perspective, social justice principles in social work would also be consistent with the notion that it is important to include high-quality items that are accessible to the people for whom the assessment is intended. When items are understandable and accessible to people with ID/DD, they have an opportunity to fully participate. A relatively new method in test development is a universal assessment design, intended to increase accessibility for students with disabilities (Thompson, Johnstone, & Thurlow, 2002). Researchers could consider some of the universal design elements when re-writing items for the NCI, which could increase accessibility and participation for people with ID/DD. The purpose of the universal test design method is to allow people with disabilities every opportunity to demonstrate what they know and believe. Thompson, Johnstone, and Thurlow (2002) note that assessments should be flexible, simple, and require little physical effort.

The universal test design elements are comparable to several recommendations espoused by researchers in the ID/DD field. For example, Verdugo et al. (2005) and Bonham et al. (2004) recommend the use of simple words that contain one to two syllables. In addition, Perry and Felce (2002) indicate that picture responses may be a flexible way to capture the experiences of people with ID/DD. In using these universal design elements, it may be possible to increase self-responses to the quality of life questions and not rely as much on proxies, which multiple researchers support (Schalock, 2010; Stancliffe, 2000; Verdugo et al., 2005).

Scale recommendations. This research establishes two general scale recommendations. First, the present study found that applying the percent total method to the individual scores for the items on the *Community Inclusion* measure serves to increase the internal consistency reliability for both the self- and proxy-respondents. Improving internal consistency would add precision and accuracy to the scale.

Second, this research suggests that HSRI should revise the *Everyday Choices* scale. This scale is extremely short and has limited variability for self-respondents, which has a couple of implications. In increasing the length of the scale through the use of high-quality items, it would be possible to add items that could capture a fuller breadth of the construct of everyday choices, which could also increase reliability of the scale. From a measurement perspective, this lack of variability negatively impacts the internal consistency reliability for this scale and makes DIF interpretation difficult. Objective quality of life measures are frequently used to make program decisions (Verdugo et al., 2005). From a program evaluation perspective, the use of this scale would make it difficult to show any improvement in outcomes, given that most of the self-respondents replied that they had choice.

As previously discussed, it is possible that response biases associated with acquiescence and recency impacted the results for this scale. Researchers note that pre-screening individuals with ID/DD is one way to reduce these response-bias options. The NCI does not require any type of pre-screening to determine if individuals tend to respond to questions with a response bias. Other measures, such as the Comprehensive Quality of Life Scale – Intellectual Disabilities (COMQoL-ID, Cummins, 1993) include pre-screening methods to identify individuals who demonstrate response biases. It is

possible that by including a pre-screening process, the NCI could identify those individuals who tend to respond in a biased way. However, screening individuals to participate in a quality of life study does not seem consistent with social justice principles.

Social Justice

Huggins (2013) indicates that both differential item functioning and measurement invariance relate to issues of fairness in testing. Fairness and these methods of evaluation are tied to the value of social justice in social work practice, which notes that individuals should have "meaningful participation in decision making" (NASW, 2008, preamble). As noted earlier based on the present research, it is possible that the use of the word "entertainment" presents more difficulty for people with ID/DD to answer. In considering this, self-respondents are then unable to sincerely provide their input about this aspect of their community inclusion experience. The same point is true for the other items that display DIF.

Other quality of life measures exist, and to date researchers have not completed a DIF analysis or measurement invariance analyses with these other tools. Completing this type of analyses, could identify other items and scales that function differently depending on who answers the questions. This research would extend the field's understanding of fairness and social justice for individuals who participate in these other measures.

Policy Considerations

One of the most controversial aspects of modern measurement theory relates to the consequences of test scores (Messick, 1980). Individuals, states, and the federal government use the three NCI scales as well as other evidence to inform policy decisions

and determine funding, and it is crucial to accurately capture what a measure intends to assess. For example, Kentucky developed a new service to enhance community inclusion (Kentucky Division of Developmental and Intellectual Disabilities, 2013), and based their reasoning in part on a comparison of 2 years of the NCI's *Community Inclusion* scale. In developing this new service, it is important that the scale scores do not also capture another construct-irrelevant component, which could then impact scale scores and result in unintended consequences for individuals with ID/DD (Nichols & Williams, 2009). From these results, we also know that individuals for whom proxies respond on their behalf tend to report fewer instances of community inclusion. Analyses of this service suggest that it may be necessary to provide additional support to these individuals.

The present analyses of the NCI instrument indicate that it is not possible to compare self-and proxy-responses without equating the scores. Researchers frequently compare states using the NCI scale scores and apply regression weights to scales when making these comparisons (NCI, 2011a). One approach that would provide greater precision over regression weights is to equate the scales using measurement invariance principles (Wicherts & Dolan, 2010). In equating the scores, policy and decision makers could be more confident in their decisions based on these scales.

Study Limitations

Two primary limitations are evident in this study. The first is the large sample size, which resulted in more complex interpretations of the findings. The second is the matching scales for the DIF analyses. The next section further explores each of these limitations.

Sample size. This study used a large dataset, which afforded a number of advantages from a research perspective. These benefits include: (a) more accuracy, (b) increased power and (c) reduced error (Gravetter & Wallnau, 1996). However, when using large datasets, small differences can be detected as statistically significant, even though they may have little practical significance. This study also relied on the use of the chi-square statistic, which is well known to inflate when sample sizes are large. As such, it is important to use effect sizes to interpret the meaningfulness of the statistically significant findings.

For most of the analyses presented in this dissertation, the effect size interpretations are established and shown to be effective. However, the DIF effect size criteria for logistic regression is one area where researchers have called for additional study (Hidalgo & López-Piña, 2004; Teresi, 2006). Hidalgo and López-Piña (2004) found that the effect size method used with logistic regression tended to under-identify DIF as compared with the Mantel-Haenszel effect size measures. The small logistic regression effect size was also problematic in this dissertation. The results of the logistic regression analyses in the present study found that all of the items detected statistically significant DIF, but the effect size measure did not detect any practical significance for any of the items. These findings are consistent with Hidalgo and López-Piña's findings. These researchers completed a simulation study where some items had DIF and others did not. The logistic regression effect size measure only distinguished large DIF in 5% of items in the large 75-item measure.

The interpretation of the measurement invariance analysis became more difficult with the large sample size. In this analysis, a chi-square statistic is generated to compare

the fit between the various measurement invariance models. The goodness of fit indices are then used to determine fit. One potential solution for this issue could involve drawing a smaller random sample from the larger national sample.

In exploring the possibility that a smaller sample could have resulted in different conclusions, this researcher randomly drew a smaller sample from the national dataset and re-analyzed measurement invariance using the *Community Inclusion* scale and the *Life Decisions* scale. The smaller sample consisted of 15% of the self-respondents and 15% of the proxy-respondents for both the *Community Inclusion* and the *Life Decisions* measures. This process was not carried out for the *Everyday Choices* scale, since the reduced variability made the chi-square statistic small in the national sample.

The conclusions from the smaller sample are the same as the larger sample. However, with a smaller sample, the findings are much more explicit, since the chi-square statistic generated through the model comparison is much smaller. For example, in comparing the configural and metric model fit for the *Community Inclusion* scale, the metric model fit these data well, χ^2 (4, N = 1,282) = 3.8, p = .43, whereas the scalar model did not fit the smaller sample well, χ^2 (11, N = 1,282) = 56.03, p < .001. In considering the *Life Decisions* scale, the smaller dataset randomly drawn from the national sample also indicates a smaller chi-square statistic, which leads to easier interpretation of these findings. Using a smaller dataset, the model fit is also good for the metric invariance model, with χ^2 (5, N = 1,488) = 8.69, p = .122. The scalar model did not fit these data as well and revealed that χ^2 (5, N = 1,488) = 59.18, p < .001. Thus, even though these findings are the same as the larger dataset, it is easier to understand the model fit based on a smaller sample.

Matching criteria. The reliability and validity of the matching criteria, which was the total score for each of the three scales, was problematic for the DIF analysis. Psychometric researchers emphasize that the matching criteria used in the DIF analysis are important (Osterlind & Everson, 2009). One general way to increase the psychometric properties of a measure is to include many high-quality items that capture the full breadth of the construct. All three of these measures are short, which may mean that the scales did not include the entire construct. By including additional items to capture the entire domain, it is possible to increase the internal consistency reliability coefficient for all of the scales (Thorndike & Thorndike-Christ, 2010) and yield a more valid measure.

Future Research

In addition to re-writing the items noted above, there are a number of future directions for this line of study. In the NCI measure, additional analysis could be completed to examine DIF based on different group comparisons. For example, DIF analyses based on race, gender, or age could provide additional validity evidence for the objective measures, but could also be extended to examine DIF for the subjective quality of life measures found on the NCI. For the subjective quality of life scale, only self-respondents are accepted, so it would not be possible to compare self- and proxy-respondents, but it would be possible to compare other groupings to see if any of the items are functioning differently for one group over another.

Another possible area of research is to apply the DIF and measurement invariance analyses to other measures researchers use to the assess quality of life in individuals with ID/DD. For example, the Ask ME! survey (Bonham et al., 2004) includes several scales

that capture objective quality of life. The DIF analysis and measurement invariance analyses could be applied to this measure to determine if similar findings were evident. In addition, Cummins' (1997) Comprehensive Quality of Life – Intellectual Disabilities (COMQol-ID) includes objective quality of life measures, to which researchers could apply the DIF and measurement invariance methods. Finally, Schalock and Keith (1993) include objective measures in their Quality of Life Questionnaire (QOL-Q). In applying the DIF and measurement invariance analyses, it may be possible to learn more about how items can function differently depending on who answers them. This knowledge then adds to the validity of evidence related to each of these measures.

From a methodological perspective, other fields such as medicine and psychology, adopted DIF analyses as a way to understand how items function. However, social work researchers have completed minimal research using these methods, but the application to cultural diversity is evident. Azocar, Areán, Miranda, and Muñoz (2001) assessed a Spanish translation of the Beck Depression Inventory for bias using DIF methods. These researchers found that Spanish-speaking individuals tended to endorse items related to tearfulness and punishment when compared with English-speaking individuals who had similar levels of depression. Social workers often engage in practice with minority groups, and Azocar et al.'s (2001) findings could have practical implications for direct practice with clients who are Spanish-speaking and have depression.

In conclusion, this dissertation appraises the internal structure of the objective quality of life measures of the National Core Indicators. A central finding is that from a measurement perspective, we cannot directly compare mean differences between self-

and proxy-respondents. This research generated several recommendations for ways to improve the objective quality of life measures through amending several items. Revising the items would help to improve the accuracy and fairness with which we measure outcomes for people with disabilities. It is possible that improved precision could then positively impact the lives of people with ID/DD, since the outcomes of these measures could measure the quality of their lives in a more accurate and accessible way.

References

Abery, B. H. (1994). A conceptual framework for enhancing self-determination. In M.F. Hayden & B. H. Abery (Eds.), *Challenges for a service system in transition: Ensuring quality community experiences for persons with developmental disabilities* (pp. 345-380). Baltimore: Brookes.

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29(1),* 67-91.

Administration on Intellectual and Developmental Disabilities (2000). *Developmental disabilities assistance and bill of rights act*. Retrieved from www.acl.gov/Programs/AIDD/DDA_BoR_ACT_2000/p2_tt_subtitleA.aspxs

Agency for Healthcare Research and Quality (2010). Environmental scan of measures for medicaid title XIX home and community-based services: Final report. (AHRQ Publication No. 10-0042-EF, June 2010). Retrieved from http://www.ahrq.gov/research ltc/hcbsreport/

American Association on Intellectual and Developmental Disabilities (2013). Definition of intellectual disabilities. Retrieved from http://www.aaID.org/ content_100.cfm?navID=21

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing (US) (1999). *Standards for*

educational and psychological testing. Washington, DC: American Educational Research Association.

Angoff, W. H. (1993). *Perspectives on differential item functioning methodology*. Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc

Azocar, F., Arean, P., Miranda, J., & Muñoz, R. F. (2001). Differential item functioning in a Spanish translation of the Beck depression inventory. *Journal of Clinical Psychology*, *57*(3), 355-365.

Bobko, P. (2001). *Correlation and regression: Principles and applications for industrial/organizational psychology and management* (2nd ed.). London: Sage.

Bonham, G. S., Basehart, S., Schalock, R. L., Boswell Marchand, C., Kirchner, N., & Rumenap, J. M. (2004). Consumer-based quality of life assessment: The Maryland ask me! Project. *Mental Retardation, 42(5)*, 338-355.

Braddock, D. L., & Parish, S. L. (2001). An institutional history of disability. In Gary L. Albrecht, Katherine D. Seelman and Michael Bury (Eds.) *Handbook of Disability Studies*. Thousand Oaks, CA: Sage.

Bradley, V. J., & Moseley, C. (2007). National core indicators: Ten years of collaborative performance measurement. *Intellectual and Developmental Disabilities, 45(5)*, 354-358.

Brockley, J. (2004). Rearing the child who never grew: Ideologies of parenting and intellectual disability in American history. *Mental retardation in America: A historical reader*, 130-164.

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

Burnett, P. C. (1989). Assessing satisfaction in people with an intellectual disability. Living in community-based residential facilities. *Australian Disability Review 1(1)*, 14-19.

Cai, L., Thissen, D., & du Toit, S. (2012). Item Response Theory for Patient Reported Outcomes (IRTPRO) for windows, version 2.1 [software]. Skokie, IL: Scientific Software International.

California Developmental Disabilities Consumer Advisory Committee (2012). What We Have Learned from the National Core Indicators Adult Consumer Survey: NCI Results from People across California in 2010: User Friendly Version 2012. Retrieved from http://www.dds.ca.gov/QA/docs/what_we_learned.pdf

Campbell, A., & Converse, P. E. (1972). Social change and human change. In A. Campbell and P.E. Converse (Eds.). *The Human Meaning of Change.* New York, NY: Russell Sage Foundation.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment.* Beverly Hills, CA: Sage.

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement. 68 (3)*, 397-412.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*(1), 98.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando, Fl: Holt, Rinehart and Winston.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological bulletin*, *52*(4), 281.

Cummins, R. A. (1993). *Comprehensive quality of life scale* (4th edition). Melbourne: Deakin University.

Cummins, R. A. (2005). Moving from the quality of life construct to a theory. *Journal of Intellectual Disability Research, 49(10),* 699-706.

Cummins, R. A. (1997). Self-rated quality of life scales for people with an intellectual disability: A review. *Journal of Applied Research in Intellectual Disabilities 10(3)*, 199-216.

Cummins, R. A. (1991). The comprehensive quality of life scale – Intellectual disability: An instrument under development. *Australian and New Zealand Journal of Developmental Disabilities*, *17(2)*, 259-264.

Cummins, R. A., McCabe, M. P., Romeo, Y., Reid, S., & Waters, L. (1997). An initial evaluation of the comprehensive quality of life scale – Intellectual disability. *International Journal of Disability, Development and Education, 44(1),* 7-19.

Cusick, C. P., Brooks, C. A., & Whiteneck, G. G. (2001). The use of proxies in community integration research. *Archives of Physical Medicine Rehabilitation, 82*, 1018-1024.

Developmental Disability Act of 2000, 42 USC 15001 Sec. 102 (8).

Diener, E., & Suh, E. (1997). Quality of life: Economic, social and subjective indicators. *Social Indicators Research, 40(1/2)*, 189-216.

Doll, E. A. (1937). The institution as a foster parent. In J. Blacher & B. L. Baker (Eds.) *The Best of AAMR: Families and Mental Retardation: A Collection of Notable*

Journal Articles Across the 20th Century. J. Blacher & B. L. Baker, eds. Printed in 2002, Washington, DC: American Association on Mental Retardation (pp.15-18). Retrieved from http://books.google.com.ezp2.lib.umn.edu/books

Donoghue, J. R., & Allen, N. L. (1993). Thin versus thick matching in the Mantel-Haenszel procedure for detecting DIF. *Journal of Educational and Behavioral Statistics*, *18*(2), 131-154.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Lawrence Erlbaum Associates.

Downing, S. M. (2003). Validity: on the meaningful interpretation of assessment data. *Medical education*, *37*(9), 830-837.

Felce, D., & Perry, J. (1995). Quality of life: Its definition and measurement. *Research in Developmental Disabilities, 16(1)*, 51-74.

Fidalgo, A. M., & Madeira, J. M. (2008). Generalized Mantel-Haenszel methods for differential item functioning detection. *Educational and Psychological Measurement*, *68*(6), 940-958.

Fleishman, J. A., Spector, W. D., & Altman, B. M. (2002). Impact of differential item functioning on age and gender differences in functional disability. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *57*(5), S275-S284.

Fujiura, G. T. (1994). Research perspectives and the community living experience. In Mary F. Hayden & Brian H. Abery (Eds.) *Challenges for a Service System in*

Transition: Ensuring Quality Community Experiences for Persons with Developmental Disabilities. Baltimore: Paul H. Brookes Publishing.

Geiser, S., & Studley, W. R. (2002). UC and the SAT: Predictive validity and differential impact of the SAT I and SAT II at the University of California. *Educational Assessment*, *8*(1), 1-26.

Gorin, J. S. (2007). Reconsidering issues in validity theory. *Educational Researcher*, *36*(8), 456-462.

Gravatter, F. J., & Wallnau, L. B. (1996). Statistics for the Behavioral Sciences (4th Ed.). Pacific Grove: Brooks/Cole Publishing Company.

Guion, R. M. (1997). *Assessment, measurement, and prediction for personnel decisions*. Mahway, NJ: Lawrence Erlbaum Associates, Inc.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, *23*(1), 17-27.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied measurement in education*, *15*(3), 309-333.

Harrington, D. (2008). *Confirmatory factor analysis*. New York, NY: Oxford University Press.

Heal, L. W., & Sigelman, C. K. (1995). Response bias in interview of individuals with limited mental ability. *Journal of Intellectual Disability Research, 39(4),* 331-340.

Henderson, L. M., & Baird, J. V. (2014, September 24). Using Tennessee's NCI Performance Indicators: Evidence for New HCBS Requirements and Revised

HCBS Assurances in Tennessee's Waiver Planning and Revision Process: Tennessee Data from 2013-2014. Retrieved from http://vkc.mc.vanderbilt.edu/assets/files/resources/NCI%20HCBS%20crosswalk%20Final%20Report%209-24-14.pdf

Hidalgo, M. D., & López-Piña, J. A. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement*, *64*(6), 903-915.

Holland, P. W., & Thayer, D. T. (1998). Differential item performance and the Mantel-Haenszel procedure. *Test Validity*, 129-145.

Hortensius, L. (2012). Project for Introduction to Multivariate Statistics: Measurement Invariance. Retrieved from http://www.tc.umn.edu/~horte005/docs/MeasurementInvariance.pdf

Huck, S. W. (2009). *Statistical misconceptions*. New York, NY: Routledge.

Huggins, A. C. (2013). The Effect of Differential Item Functioning in Anchor Items on Population Invariance of Equating. *Educational and Psychological Measurement*, 0013164413506222.

Human Service Research Institute (2010). 2009-2010 Final Report. Retrieved from: http://www.nationalcoreindicators.org/upload/core-indicators/NCI_CS_09-10_FINAL_Report_2.pdf

Jaeger, P. T., & Bowman, C. A. (2005). *Understanding disability: Inclusion, access, diversity and civil rights.* Westport, CT: Praeger Publishing.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, *14*(4), 329-349.

Johnstone, C. J., Bottsford-Miller, N. A., & Thompson, S. J. (2006). Using the Think Aloud Method (Cognitive Labs) to Evaluate Test Design for Students with Disabilities and English Language Learners. Technical Report 44. *National Center on Educational Outcomes, University of Minnesota*.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological bulletin*, *112*(3), 527.

Kane, M. T. (2006). Content-related validity evidence in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 131-153). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational measurement*, *38*(4), 319-342.

Kane, M. T. (2012). Validating score interpretations and uses. *Language Testing*, *29*(1), 3-17.

Kentucky Division of Developmental and Intellectual Disabilities (2013, July 19). *Kentucky National Core Indicators Quality Improvement Committee: 2013 Recommendations Report*. Retrieved from http://www.belongingky.org/wpcontent/uploads/2013/09/Recommendations-Report_2013final.pdf

Ladd-Taylor, M. (2011). Eugenics and social welfare in new deal Minnesota. In Paul A. Lombardo (Ed.), *Century of Eugenics in America* (p.117-140). Bloomington, IN: Indiana University Press

Land, K. C. (1983). Social indicators. *Annual Review of Sociology, 9*, 1-26.

Landesman, C. (1986). Quality of life and personal life satisfaction: Definition and measurement issues. *Mental Retardation, 24(3)*, 141-143.

Lakin, K. C., Hayden, M. F., & Abery, B. H. (1994). An overview of the community living concept. In M. F. Hayden & B. H. Abery (Eds.) *Challenges for a service system in transition.* Baltimore, MA: Paul H. Brooks Publishing co.

Larson, S. A., Lakin, K. C., Anderson, L., Kwak, N., Lee, J. H., & Anderson, D. (2001). Prevalence of mental retardation and developmental disabilities: Estimates from the 1994/1995 National Health Interview Survey Disabilities supplements. *American Journal on Mental Retardation, 106(3)*, 231-252.

Li, C., Tsoi, E.W.S., Zhang, A.L., Chen, S. & Wang, C.K.J. (2012). Psychometric properties of self-reported quality of life measures for people with intellectual disabilities: A systematic review. *Journal of Developmental and Physical Disabilities.* DOI 10.1007/s10882-012-9297-x

Long, J. S. (1983). *Confirmatory Factor Analysis.* Newbury Park, CA: Sage.

Lyons, G. (2010). Quality of life for persons with intellectual disabilities: A review of the literature. In R. Kober (Ed.), *Enhancing the quality of life of people with intellectual disabilities,* Social Indicators Research Series 41, doi: 10.1007/978-90-481-9650-0_6

Magis, D., & De Boeck, P. (in press). Type I error inflation in DIF identification with Mantel-Haenszel: An explanation and a solution. *Educational and Psychological Measurement*.

McVilly, K. R., Burton-Smith, R., & Davidson, J. A. (2000). Concurrence between subject and proxy ratings of quality of life for people with and without intellectual disabilities. *Journal of Intellectual and Developmental Disability, 25(1)*, 19-39.

Messick, S. (1989a). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18(2)*, 5-11.

Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research*, *45*(1-3), 35-44.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*, 1012-1027.

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

Messick, S. (1989b). Validity of test interpretation and use. Princeton, NJ: Educational Testing Services. Retrieved from http://files.eric.ed.gov/fulltext/ED395031.pdf

Meyer, J. P. (2014a). *Applied Measurement with JMetrik*. New York, NY: Routledge, Taylor & Francis Group.

Meyer, J. P. (2014b). jMetrik for Max OSX, version 3.1 [software]. Retrieved from http://www.jmetrik.com/

Millsap, R. E. (2012). *Statistical approaches to measurement invariance*. New York, NY: Routledge.

Millsap, R. E., & Kwok, O. M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological methods*, *9*(1), 93.

Moseley, C. (2011, September 30). Using National Core Indicator Data to Improve Service Quality and System Performance. Retrieved from http://www.nasuad.

org/documentation/hcbs2011/Presentations/TPlenaryNASDDDS.pdf

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement*, *18*(4), 315-328.

National Association of Social Workers (2008). Code of ethics of the national association of social workers. Retrieved from http://www.socialworkers.org/pubs/code/code.asp.

National Association of State Directors of Developmental Disabilities Services and Human Services Research Institute (2003). National core indicators: Adult consumer survey 2008-2009

National Core Indicators (2011a). Annual Summary Report 2009-2010. Retrieved from http://www.nationalcoreindicators.org/upload/core-indicators/NCI_Annual_Summary_Report_2009-10_FINAL.pdf

National Core Indicators (2011b). Welcome. Retrieved from http://www2.hsri.org/nci/

National Core Indicators (2011c). Consumer Outcomes: Phase XII Final Report 2009-2010 Data Set.

National Core Indicators (2013). About the National Core Indicators. Retrieved from http://www.nationalcoreindicators.org/

Nichols, P.D., & Williams, N. (2009). Consequences of test score use as validity evidence: Roles and responsibilities. *National Council on Educational Measurement, 28(1), 3-9.*

Nimon, K., Zientek, L. R., & Henson, R. K. (2012). The assumption of a reliable instrument and other pitfalls to avoid when considering the reliability of data. *Frontiers in psychology*, *3*.

Noll, S. (1995). *Feeble-minded in our midst: Institutions for the mentally retarded in the South, 1900-1940*. UNC Press Books.

Novak Amado, A., Stancliffe, R. J., McCarron, M., & McCallion, P. (in press). Social inclusion and community participation of individuals with intellectual/developmental disabilities. *American Association on Intellectual Disabilities.*

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory.* New York: McGraw-Hill, inc.

Nussbaum, M., & Sen, A. (1993). *The quality of life*. Oxford University Press.

Osburn, H. G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological Methods,* 5, 343-355.

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning* (2nd ed.). Thousand Oaks, CA: Sage.

Perry, J., & Felce, D. (2002). Subjective and objective quality of life assessment: Responsiveness, response bias, and resident:proxy concordance. *Mental Retardation, 40(6)*, 445-456.

Pfeiffer, D. (1993). Overview of the disability movement: History, legislative record, and political implications. *Policy Studies Journal 21(4),* 724-734.

Popham, W. J. (1997). Consequential validity: Right Concern-Wrong Concept. *Educational measurement: Issues and practice*, *16*(2), 9-13.

Rapley, M. (2003). Quality of life research: A life of quality-just what does QOL mean? doi: 10.4135/9781849209748, p. 1-41.

Rapley, M., Ridgway, J., & Beyer, S. (1998). Staff:staff and staff:client reliability of the schalock and keith (1993) quality of life questionnaire. *Journal of Intellectual Disability Research, 42,* 37-42.

Reamer, F. G. (2013). *Social work values and ethics*. Columbia University Press.

Rogers, W. T., & Bateson, D. J. (1991). The influence of test-wiseness on performance of high school seniors on school leaving examinations. *Applied Measurement in Education*, *4*(2), 159-183.

Schalock, R. L. (2000). Three decades of quality of life. In M.L. Wehmeyer and J.R. Patton (Eds.). *Mental Retardation in the 21st Century.* Austin, TX: PRO-ED.

Schalock, R. L. (2004). The concept of quality of life: What we know and do not know. *Journal of Intellectual Disability Research, 48(3)*, 203-216.

Schalock, R. L. (2010). The measurement and use of quality of life-related personal outcomes. In R. Kober (Ed.), *Enhancing the quality of life of people with intellectual disabilities: From theory to practice* (pp. 3-16). New York, NY: Springer Science.

Schalock, R. L., Brown, I., Brown, R., Cummins, R. A., Felce, D., Matikka, L., Keith, K. D., & Parmenter, T. (2002). Conceptualization, measurement, and application of quality of life for persons with intellectual disabilities: Report of an international panel of experts. *Mental Retardation, 40, (6),* 457-470.

Schalock, R. L., & Felce, D. (2004). Quality of life and subjective well-being: Conceptual and measurement issues. In E. Emerson, C. Hatton, T. Thompson &

T. Parmenter (Eds.), *International handbook of applied research in developmental disabilities*. West Sussex, England: John Wiley and Sons Ltd

Schalock, R. L., & Keith, K. D. (1993). Quality of Life Questionnaire. IDS publishing.

Schalock, R. L., Luckasson, R. A., Shogren, K. A., Bradley, V., Buntinx, W. H. E. et al. (2007). The renaming of *mental retardation*: Understanding the change to the term *intellectual disability*. *Intellectual and Developmental Disabilities, 45(2)*, 116-124.

Schalock, R. L., & Verdugo, M. A. (2002). *Handbook on quality of life for human service practitioners.* Washington, DC: American Association on Mental Retardation.

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review*, *18*(4), 210-222.

Scotch, R. K. (1989). Politics and policy in the history of the disability rights movement. *The Milbank Quarterly*, 380-400.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference* (2nd ed.). New York, NY: Wadsworth Cengage learning.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, *16*(2), 5-24.

Sigelman, C. K., Budd, E. C., Spanhel, C. L., & Schoenrock, C. J. (1981). Asking questions of retarded persons: A comparison of yes-no and either-or formats. *Applied Research in Mental Retardation*, 347-357.

SPSS (2011). SPSS for windows, version 20. Chicago, IL: IBM Corp.

Stancliffe, R. J. (1995). Assessing opportunities for choice making: A comparison of self-and staff reports. *American Journal on Mental Retardation, 99,* 418-429.

Stancliffe, R. J. (2000). Proxy respondents and quality of life. *Evaluation and Program Planning, 23*, 89-93.

Stancliffe, R. J. (1999). Proxy respondents and the reliability of the quality of life questionnaire empowerment factor. *Journal of Intellectual Disability Research, 43(3),* 185-193.

Stancliffe, R. J., Lakin, K. C., Taub, S., Chiri, G., & Byun, S. (2009). Satisfaction and sense of well being among Medicaid ICF/MR and HCBS recipients in six states. *Intellectual and Developmental Disabilities, 47(2),* 63-83.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational measurement*, *27*(4), 361-370.

Tabachnick, B. G., & Fidell, L. S. (2001). Using multivariate statistics (5th ed.). Boston: Pearson

Teresi, J. A. (2006). Different approaches to differential item functioning in health applications: Advantages, disadvantages and some neglected topics. *Medical care*, *44*(11), S152-S170.

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved from http://education.umn.edu/NCEO/OnlinePubs/Synthesis44.html

Thorndike, R. M., & Thorndike-Christ, T. (2010). *Measurement and evaluation in psychology and education.* (8th ed.). Boston, MA: Pearson.

Ticha, R., Lakin, C., Larson, S. A., Stancliffe, R. J., Taub, S., Engler, J., Bershadshy, J., & Moseley, C. (2012). Correlates of everyday choice and support-related choice for 8,892 randomly sampled adults with intellectual and developmental disabilities in 29 states. *Intellectual and Developmental Disabilities, 50(6)*, 486-504.

University of Minnesota Research and Training Center on Community Living (2006). *Medicaid home and community-based services for persons with intellectual and developmental disabilities: Background and findings from consUM-RTCer interviews and the medicaid statistical information systems.* Retrieved from: http:// rtc.UM-RTCn.edu/search/searchresults1.asp

U.S. Government (2013). Americans with disabilities act of 1990, as Amended. Retrieved from http://www.ada.gov/pubs/adastatute08.pdf.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational research methods*, *3*(1), 4-70.

van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, *9*(4), 486-492.

Verdugo, M. A., Schalock, R. L., Keith, K. D., & Stancliffe, R. J. (2005). Quality of life and its measurement: important principles and guidelines. *Journal of intellectual disability research*, *49*(10), 707-717.

Wehmeyer, M. L., & Abery, B. H. (in press). Self-determination and choice. *American Association on Intellectual Disabilities.*

White, G. W., Simpson, J. L., Gonda, C., Ravesloot, C., & Coble, Z. (2010). Moving from independence to interdependence: A conceptual model for better understanding community participation of centers for independent living consumers. *Journal of Disability Policy Studies*, *20*(4), 233-240.

Wicherts, J. M., & Dolan, C. V. (2010). Measurement invariance in confirmatory factor analysis: An illustration using IQ test performance of minorities. *Educational Measurement: Issues and Practice*, *29*(3), 39-47.

Wolfensberger, W. (1972). The principle of normalization in human services.

World Health Organization (1948). *Constitution*. Retrieved from http://www.who.int/about/definition/en/print.html

World Health Organization Quality of Life Group (WHOQOL group) (1995). The world health organization quality of life assessment (WHOQOL): Position paper from the world health organization. *Social Science Medicine, 41(10)*, 1403-1409.

Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning (DIF). *Ottawa: National Defense Headquarters*.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language assessment quarterly*, *4*(2), 223-233.

Zwick, R. (2012, May). A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements and criterion refinement. Retrieved from http://www.ets.org/Media/Research/pdf/RR-12-08.pdf

Appendix A: Differential Item Functioning: Terms and Methods

Osterlind and Everson (2009) defined differential item functioning as the process of statistically examining items to determine if questions systematically function in a different way based on who answers the question rather than their knowledge, skills, or ability to answer an item. The purpose of this analysis was to determine if items consistently exhibit different probabilities of answering correctly based on group membership when conditioned on the total score of the scale (Angoff, 1993). Since it was related to the systematic function of items or measures, it can be leveraged when developing validity evidence for a measure (Angoff, 1993; Osterlind & Everson, 2009). Specifically, it provided evidence related to the internal structure of a measure (Cole, 1981).

Researchers (Dorans & Holland, 1993; Osterlind & Everson, 2009) highlight that DIF analysis was distinct from group analysis or item impact. DIF analysis was conditioned on the latent trait or ability, whereas item impact did not include this additional specification. This matching allows researchers to identify unexplained variation in responses between groups (Dorans & Holland, 1993) to identify construct irrelevant differences in groups (AERA, APA, & NCME, 1999). In using a DIF analysis versus an item impact analysis, Dorans and Holland (1993) stressed that it was possible to "compare the comparable at each ability level" (p. 38). Holland and Thayer (1988) defined comparability as the points where individuals may diverge, which then translate to performance on a measure. A statistically significant item impact finding did not necessarily constitute item bias but did provide evidence that there were overall differences in performance between the groups (Angoff, 1993).

Terms Unique to DIF analysis

The differential item functioning literature contained a number of frequently used terms that have specific meanings. The following paragraphs clarify these terms.

The first distinction was between the reference group and focal group. The *reference group* was the dominant, majority group. In this sample, the reference group contained the individuals who responded for themselves. The *focal group* was the studied group and in this study consisted of proxy-respondents. Osterlind and Everson (2009) indicated that the focal group was somehow conceived of as being hindered in their ability to respond to an item. In the present study, it was anticipated that proxies, who were not with individuals at all times, could be disadvantaged when responding to the objective quality of life questions.

The second key term related to DIF was the *conditioning or matching criterion*. This was the underlying skill, ability, or knowledge measured. The reference and focal groups were matched based on this conditioning criterion. It was possible to use either an internal or external measure with which to match, but the internal matching was more common (Osterlind & Everson, 2009). An internally matched measure was when a total test score that also contains the studied item was used as the matching criteria. An externally matched measure occurred when researchers used a different measure to provide the matching criteria that was separate from the items being assessed for DIF. This study used internally matched criteria with the total score on the scale functioning as the matching criteria.

Another matching consideration was thin versus thick matching. Researchers used thin matching with all possible scores (Meyer, 2014a). A potential problem was that

with tests that were polytomous or had a large range of total scores, some test assumptions for both the Mantel-Haenszel and logistic regression may not be met. For example, with logistic regression, there should be no more than 20% of cells with zero counts. Thick matching was when the researcher compressed the score bands or responses. Donoghue and Allen (1993) noted that thick matching had two primary advantages, (a) stability and (b) power, with different matching approaches controlling Type I error better than others. In the first point, the authors indicated that thick matching helped to increase the number of instances within the full chi-square contingency table. Donoghue and Allen's second point was that when it was not possible to match instances within a cell, the observations within the cell were excluded from the analysis. Through increasing the data found for each cell, it was more likely that this information could then be assessed in the statistic, thereby increasing power. Donoghue and Allen also found that the percent total method was an acceptable thick matching method. This was used both for the *Community Inclusion* and the *Life Decision* scales.

A third term included the type of DIF. Researchers described this as either uniform or non-uniform/crossing varieties. In *uniform* DIF, one group consistently performed lower than the other group when conditioned on the matching criteria (Osterlind & Everson, 2009). Findings graphically represented that the trace curves of one group were constantly lower than the other group and did not intersect or cross. Osterlind and Everson (2009) explained, "Uniform DIF is present when the probabilities of success on the flagged test item for one group are consistently higher than the probabilities of success for the focal group over all trait levels" (p. 11). In *non-uniform* DIF, one group initially performed lower than the other group and then, at higher levels

of ability, the previously underperforming group switched position and became the better performing group (Osterlind & Everson, 2009). Graphically, this occurs when the trace curve of one group begins lower than the other group, then intersects and surpasses the other group's trace curve.

Statistical Methods to Assess DIF

Millsap (2012) noted that there are multiple methods to assess DIF. These methods included the use of either Item Response Theory (IRT) models or non-IRT procedures (Millsap, 2012). To conduct an IRT-based analysis, researchers needed large sample sizes, which can be difficult to obtain in applied settings (Millsap, 2012; Narayanan & Swaminathan, 1994). Researchers used three predominant non-IRT-based procedures to detect DIF, which included standardization, Mantel-Haenszel, (Dorans & Holland, 1993), and logistic regression (Swaminathan & Rogers, 1990). This paper used both the Mantel-Haenszel and logistic regression procedures to assess DIF.

Mantel-Haenszel statistic. Researchers frequently used the Mantel-Haenszel method. A non-parametric analysis, the Mantel-Haenszel (M-H) procedure can be calculated with dichotomous items (i.e. correct/incorrect responses) or polytomous items (i.e., more than one response) (Fidalgo & Madeira, 2008; Osterlind & Everson, 2009). A chi-square statistic was the basis for the M-H procedure and summarized as 2 x C where 2 represents the focal and reference groups (2 groups total), and C represented the number of response scores or column groups. In dichotomous items, the statistic summarized as 2 x 2, which indicated that there were two values for the columns (i.e., correct and incorrect) (Dorans & Holland, 1993; Fidalgo & Madeira, 2008; Osterlind & Everson, 2009). However, with polytomous scored items the number of column groups

was presented as the C value. As an example of a polytomous item, if a researcher used a 4-point rating scale, then the number of response options was 4. The summary for this item was 2 x 4.

Since the chi-square test was the basis for M-H Statistic, the assumptions corresponding to the chi-square test for independence should also be met when conducting an analysis using M-H to prevent Type I errors. The two assumptions for this test were (a) that the observations are independent and (b) that at least five observations are available in each cell (Gravetter & Wallnau, 1996). The first assumption indicated that each person contributed only one data point in the analysis. The second assumption was associated with the sample size. If a sample size was small, then the test became too sensitive and overly large chi-square values were obtained (Gravetter & Wallnau, 1996).

The function of the M-H was to test for an interaction between the rows and columns while conditioned on the matching criterion (Osterlind & Everson, 2009). Another way to restate this interaction was that it assessed a group by item interaction (Angoff, 1993). The null hypothesis was then that there was no interaction between the row and column. Dorans and Holland (1993) considered the null hypothesis as the odds of getting an item correct at each level of matching for both the focal and reference group. The alternative hypothesis was that the odds of getting an item incorrect at each level of matching were different for the focal and reference groups (Dorans & Holland, 1993).

There were a number of features that made the M-H statistic popular when detecting DIF. The first was that a DIF analysis did not require large sample sizes like that of other methods for detecting DIF (Fidalgo & Madeira, 2008; Osterlind & Everson,

2009). The M-H statistic had good power for detecting uniform DIF (Dorans & Holland, 1993; Fidalgo & Madeira, 2008; Osterlind & Everson, 2009). Finally, the statistic was fairly easy to understand as compared with some of the more complex models, such as structural equation models that were available. However, a considerable disadvantage of the M-H statistic was that it did not detect non-uniform DIF well (Fidalgo & Madeira, 2008; Osterlind & Everson, 2009).

Logistic regression analysis. Logistic regression was another method used to detect DIF. This method was useful when assessing either dichotomous or continuously scored data. When considering dichotomously or polytomous scored dependent variables, logistic regression functioned in a similar way as multiple regression (Tabachnick & Fidell, 2001). Unlike multiple regression, logistical regression does not have assumptions associated with the distribution, such as a normal distribution and equal variances in the groups (Tabachnick & Fidell, 2001) since the model is non-linear. One disadvantage was that sample size could impact the findings with small sample sizes potentially leading to model over-fitting (Tabachnick & Fidell, 2001).

In this model, the consistency of a response by group membership was conditioned on a variable, usually the total score (Zumbo, 1999). The statistical model included three variables entered in the following order: (a) conditioning or matching variable, (b) the group membership variable, and (c) the interaction between the conditioning variable and the group membership variable (Millsap, 2012). The following equation depicted the model that examined non-uniform DIF:

$$Y' = \beta_0 + \beta_1 \text{Total} + \beta_2 \text{Group} + \beta_3 (\text{Total x Group}) \quad (1)$$

The next logistic model did not include the interaction term:

$$Y' = \beta_0 + \beta_1 Total + \beta_2 Group \quad (2)$$

Finally, the last model only included the conditioning variable:

$$Y' = \beta_0 + \beta_1 Total \quad (3)$$

As in the M-H statistic, the conditioning or matching variable was the total score on a measure. The group variable was a dummy coded variable to reflect membership in the group of interest in this study, either self- or proxy-respondents.

In sum, the crux of the DIF analysis was a consideration of the model fit between the various models (Tabachnick & Fidell, 2001). The -2 log likelihood for equation 3 was subtracted from equation 1 to evaluate if there was any DIF. The effect size was calculated by subtracting the Nagelkerke R-squared value of equation 3 from equation 1. To consider non-uniform or crossing DIF, a comparison of the change in -2 log likelihood for equations 3 and 2 occurred. Similarly, the Nagelkerke R-squared values for equations 3 and 2 were subtracted. To assess uniform DIF, the -2 log likelihood for equations 2 and 1 were subtracted and the Nagelkerke R-squared was subtracted to arrive at the effect size.

To contrast the Mantel-Haenszel statistic with the logistic regression model, a number of positives and negatives were evident for each method. Swaminathan and Rogers (1990) concluded that the logistical regression model and M-H detect uniform DIF equally well. The M-H statistic had considerable power for detecting uniform DIF, particularly when sample sizes were small (Osterlind & Everson, 2009). However, logistical regression did a better job of detecting nonuniform DIF (Swaminathan & Rogers, 1990).

{Replace with the body of your thesis. Do not delete the final two paragraph returns at the end of the document in the process of pasting in the body of your thesis, as this will change the page numbering. If you do paste in the body of your thesis, you may take a second, separate step to delete the extra paragraph returns if necessary, but, if you do so, be sure to double-check the page numbering afterwards to be sure it is still correct.}